VERSION 3 OF THE RETRIEVING EFFECTIVELY FROM MEMORY MODEL: EXTENSIONS AND LIMITATIONS

Tyler M. Ensor

A Dissertation submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy

Department of Psychology

Memorial University of Newfoundland

May 2019

St. John's, Newfoundland and Labrador

Abstract

Shiffrin and Steyvers (1997) presented a computational model of human memory they called Retrieving Effectively from Memory (REM). In their original report of the model, they described several REM variants. To date, the majority of papers using REM have employed the simplest version of the model, known as REM.1. Although it does not matter for most applications, REM.1 makes an important, simplifying assumption: namely, that item strengthening always accumulates in a single mnemonic trace. In other words, there is one trace for each study item, regardless of the time for which items were presented or the number of times items were repeated. In the case of spaced repetitions, this assumption of automatic, single-trace storage is untenable. Here, I use a version of REM Shiffrin and Steyvers termed REM.3. In REM.3, spaced repetitions are stored in one trace if subjects identify the repetition as previously studied, and in separate traces otherwise. I show that this model can account for two findings previously believed to be inconsistent with the REM framework: the spacing effect (Delaney, Verkoeijen, & Spirgel, 2010) and positive list-strength effects observed with spaced strengthening (Ratcliff, Clark, & Shiffrin, 1990, Experiment 5) and the strong-interference paradigm (Norman, 1999). I then test a novel prediction from REM.3 concerning the list-strength effect. Three experiments were unable to find support for this prediction. Instead, I present an explanation of the list-strength effect based on strategy disruption.

Acknowledgments

My graduate work was supported by two postgraduate scholarships awarded to me by the Natural Sciences and Engineering Research Council of Canada (NSERC). I was also supported by an NSERC Discovery Grant awarded to Aimée Surprenant. Finally, the experiments reported in this dissertation were supported by Aimée Surprenant's Discovery Grant.

I want to thank my doctoral supervisors, Aimée Surprenant and Ian Neath, for their mentorship and guidance throughout my graduate career. They provided me with invaluable advice on academia, research projects, and the job market, and gave me the flexibility to pursue the topics and projects that interested me. They also exhibited incredible patience regarding my hyphenation of compound modifiers and excessive use of Latin phrases in my writing.

My dissertation committee was composed of Aimée Surprenant, Ian Neath, and Jon Fawcett. I am indebted to all three of them, who read and provided constructive feedback on both my research proposal and dissertation.

I also want to thank the researchers with whom I have collaborated during my graduate career, both on projects reported in this dissertation and others. I present their names in alphabetical order: Tyler Bancroft, Tamra Bireta, Chrissy Chubala, Weyam Fahmy, Dominic Guitard, Bill Hockley, Jeff Jones, Maria Learning, Ian Neath, Philip Servos, Brent Snook, Aimée Surprenant, Pelin Tan, and Hannah Willoughby.

I am indebted to Christina Thorpe for her invaluable teaching mentorship. My work as a teaching assistant in her Research and Writing in Psychology course shaped my pedagogical approach and allowed me to hone and refine my teaching style.

I am grateful to Ken Malmberg, who shared some of his REM.1 code with me.

On a personal note, I want to thank my parents, Stan Ensor and Marisa Ensor, for their support and encouragement. I want to thank my sister, Tiffany Ensor, for her seemingly endless wit, which never failed to make me laugh. Finally, I want to thank Abby Ensor, whose remarkable longevity is inspiring.

Table of Contents

Chapter 1: Introduction

The overarching goal of the present work is to test the viability of a computational model of human memory called Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997). REM's applicability is widespread, with REM successfully accounting for a variety of human-memory phenomena (see, e.g., Criss, 2006; Criss, Malmberg, & Shiffrin, 2011; Lehman & Malmberg, 2013; Malmberg & Shiffrin, 2005; Malmberg, Zeelenberg, & Shiffrin, 2004; Osth, Fox, McKague, Heathcote, & Dennis, 2018; Schooler, Shiffrin, & Raaijmakers, 2001). Yet, several findings in the literature are inconsistent with the REM framework, including the spacing effect (Delaney, Verkoeijen, & Spirgel, 2010) and some positive list-strength effects (Norman, 1999, 2002).

When Shiffrin and Steyvers (1997) first proposed REM, they presented several versions varying in complexity. Importantly, a version they called REM.1, which is the simplest version of REM, is most often utilized, and it is with REM.1 that the above-mentioned phenomena are inconsistent. My first test of REM, therefore, assessed whether a more realistic version of the model—termed REM.3 by Shiffrin and Steyvers—can account for the phenomena (see Chapter 4). My simulations demonstrate that, although these phenomena were previously thought to be inconsistent with REM, a more realistic version is capable of predicting the results.

Shiffrin and Steyvers' (1997) treatment of REM.3 was quite brief, and thus the simulations presented in Chapter 4 represent the first extensive test of REM.3's viability. Importantly, REM.3 makes novel predictions concerning a phenomenon known as the list-strength effect (described in-depth in Chapter 3). As a second test of REM.3's viability, then, I conducted three experiments to explore these novel predictions.

The present dissertation has two major parts, which I will term the simulation test of REM.3 and the empirical test of REM.3. However, some background is necessary before delving into the REM tests. In the present chapter, I begin by briefly reviewing several accounts of retrieval failure: decay, interference, inhibition, and repression. This is followed by an in-depth examination of two related interference phenomena: the list-length effect and the list-strength effect (Chapters 2 and 3, respectively). Chapter 4 presents the simulation test of REM.3, and Chapter 5 presents the empirical test of REM.3. Finally, in Chapter 6, I interpret the simulation and experimental results with respect to REM.3 and other theoretical accounts of the list-strength effect.

1.1. Accounts of Retrieval Failure

For the most part, everyday remembering is accomplished quickly and seemingly without effort. At any given moment, we have access to thousands of words and their meanings; the names of numerous acquaintances, friends, and relatives; a lifetime of first-person experiences; and so on. Of course, to use one of these memories, it must be retrieved. Memory is cue dependent (J. E. Eich, 1980; James, 1890/1983; Surprenant & Neath, 2009, Chapt. 3; Tulving, 1974): A memory is accessible only when an appropriate cue is utilized to access it. For example, one might remember an acquaintance's name when presented with their face—here, the face is the cue and the name is the target. Therefore, without a cue, there is no retrieval.

Despite the frequency with which memory retrieval succeeds, it also can fail. Memory failure can take several forms: Sometimes, despite a strong phenomenological feeling of knowing the sought-after memory exists, it cannot be accessed; this is termed the tip-of-the-tongue phenomenon (Brown & McNeill, 1966). Retrieval can also fail in

the absence of a tip-of-the-tongue phenomenon, such as when a subject in a memory test cannot remember all of the words from the study list, even though they were all encoded. Other times, people incorrectly remember events that did not occur, such as when post-event information modifies the original memory (i.e., the misinformation effect; Loftus & Palmer, 1974); when subjects recall an item related to, but not present on, the study list, as in the Deese-Roediger-McDermott paradigm (Deese, 1959; Roediger & McDermott, 1995); or when subjects make a false alarm on a recognition test. What is responsible for retrieval failure?

1.1.1. Decay theory

One of the earliest explanations of retrieval failure was decay theory, according to which memories become inaccessible as the time since they were encoded or last accessed increases (Thorndike, 1913; see also J. Brown, 1958; Ebbinghaus, 1885/1913; Peterson & Peterson, 1959; Reitman, 1974). When a target is retrieved with a given cue, it is assumed that the associative strength between the cue and target increases. So, under the decay framework, cue-target associative strengths steadily weaken with time unless they are periodically refreshed through retrieval. In decay theory, then, forgetting occurs when the associative strength between a cue and target weakens so much that retrieval is no longer possible, and permanent forgetting occurs when the associative strength between any possible cue and the target decays completely.

Decay theory was challenged by a number of findings demonstrating that the passage of time alone was not sufficient to degrade memory. McGeoch (1932a) pointed out that the passage of time is confounded with intervening experiences—that is, as the time since a memory was last accessed increases, the number of events added to memory

increases. As such, time-based forgetting could instead be due to retroactive interference. For example, Jenkins and Dallenbach (1924) tested subjects' memory for a list of nonsense syllables following a retention interval of 1, 2, 4, or 8 h. One group of subjects was awake during the retention interval, whereas the other group was asleep. Subjects in the awake condition produced forgetting functions consistent with decay theory: As the length of the retention interval increased, memory decreased. Crucially, however, retention interval did not affect memory in the asleep condition. Clearly, this is inconsistent with temporal decay, but consistent with McGeoch's suggestion: In the awake condition, subjects were continuously adding new mnemonic traces to memory; conversely, subjects in the asleep condition were not. By unconfounding the passage of time from intervening experiences, then, Jenkins and Dallenbach demonstrated that, on its own, time-based decay is insufficient to explain retrieval failure.

1.1.2. Interference theory

Based on Jenkins and Dallenbach's (1924) results and similar findings (Cheng, 1929; McGeoch, 1929, 1931, 1932b, 1933a, 1933b, 1936; McGeoch & McDonald, 1931; McGeoch & McGeoch, 1937), McGeoch (1932a, 1942) proposed interference theory as an alternative to decay theory. According to interference theory, retrieval failure occurs when access to the target is impeded by other memories associated to the cue. In one case, a retrieval cue subsumes so many memories that, although it is associated to the target, it cannot be found among all of the other memories.[1] An illustrative example is provided by

1. In most models, it is assumed that, given infinite sampling attempts, every memory subsumed by a cue will eventually be retrieved (see, e.g., Raaijmakers & Shiffrin, 1980).

Wickens, Born, and Allen (1963), who conducted a study using the Brown-Peterson paradigm (J. Brown, 1958; Peterson & Peterson, 1959). On a Brown-Peterson trial, subjects study a list of three items that they must remember in order. A distractor task (e.g., counting backward by threes from a three-digit number) is interpolated between study and test. In Wickens et al.'s study, stimuli for three quarters of the trials were consonant trigrams (e.g., RQX, PFH, GZM, etc.); on every fourth trial, however, subjects memorized three-digit number triplets instead. Performance declined progressively as subjects moved through the consonant-trigram trials, but performance rebounded when the to-be-remembered stimuli were changed to three-digit numbers. After three consecutive consonant-trigram trials, the "consonant" cue subsumed nine total items, thereby reducing its utility. In contrast, when the category was changed to three-digit number triplets, the new "number" cue subsumed only the three items from the current trial, thereby enhancing its efficacy. This type of interference is called cue overload or the fan effect (Badham, Poirier, Gandhi, Hadjivassiliou, & Maylor, 2016; Surprenant & Neath, 2009, Chapt. 5; M. J. Watkins & Watkins, 1976; O. C. Watkins & Watkins, 1975).

A second way in which other memories can interfere with a target occurs when competing memories are more strongly associated to the cue than the target. For example, consider a task in which subjects are asked to name as many birds as possible. Here, the "bird" cue is used to probe memory. High-typicality exemplars, such as "robin" and "pigeon", will be easier to access than low-typicality exemplars, such as "ostrich" or

---

However, in practice, retrieval attempts are terminated because the subject gives up or the experimenter has imposed a time limit on recall.

"vulture" (Bousfield & Barclay, 1950; Bousfield, Cohen, & Silva, 1956; Bousfield & Sedgewick, 1944; Bousfield, Sedgewick, & Cohen, 1954; Bousfield, Whitmarsh, & Esterson, 1958). This is because high-typicality exemplars are more strongly associated to the "bird" cue than low-typicality exemplars, and thus the high-typicality exemplars impede access to the low-typicality exemplars. This type of interference is called occlusion or blocking. Analysis of subjects' output protocols in exemplar-naming tasks shows that high-typicality exemplars are mentioned more often across subjects than low-typicality exemplars, and that high-typicality exemplars appear earlier in subjects' output sequences than low-typicality exemplars, consistent with occlusion (Bousfield & Barclay, 1950; Bousfield & Sedgewick, 1944; Bousfield et al., 1954, 1956, 1958; W. Brown, 1915).

More traditional memory tests also produce occlusion. Wike and Wike (1970) asked subjects to study a list of 20 unrelated words for a free-recall test. Importantly, although 15 of the words were presented once, the remaining five words were presented three times. As one would expect, memory for the repeated words was better than memory for the once-presented words. Critically, and consistent with occlusion, repeated words were significantly more likely than nonrepeated words to be output early in the test phase. Wixted, Ghadisha, and Vera (1997) showed that this effect extends to response times: After studying a list of once-presented and repeated words, subjects output repeated items earlier and faster (as measured by time from the start of the test phase to output) than once-presented words.

One of McGeoch's (1932a, 1942) critical insights was that context—what he termed "stimulating conditions"—contributed to encoding and retrieval. Although

scholars disagree on the precise definition of context, the term is generally thought to refer to factors peripheral to the to-be-encoded stimulus, but which nevertheless become incorporated into its mnemonic trace. Context can include internal factors specific to the individual subject, such as pharmacological state (Stillman, Weingartner, Wyatt, Gillin, & Eich, 1974), emotional state (E. Eich, 1995), or posture (Rand & Wapner, 1967), and external factors specific to the to-be-remembered stimuli or learning environment, such as the room in which encoding takes place (Smith, Glenberg, & Bjork, 1979); the font in which to-be-remembered words are displayed (Reder, Donavos, & Erickson, 2002); the font colour, background colour, and screen location of to-be-remembered words (Murnane & Phelps, 1993); background pictures on which to-be-remembered words are displayed (Hockley, 2008); ambient odour present during encoding (Isarida et al., 2014); and background music accompanying presentation of the to-be-learned stimuli (S. M. Smith, 1985). Because context is part of both the mnemonic trace and the retrieval cue, changing the context between encoding and retrieval impairs memory. In this way, McGeoch's insight foreshadowed Mandler's (1980) butcher-on-the-bus phenomenon: If we see our butcher on the bus, although we may experience a sense of familiarity, we may be unable to retrieve the butcher's name or even why the butcher seems familiar. This is because the cue—the butcher's face—lacks the context of the butcher shop. Although empirical support for the role of context in memory was inconsistent when McGeoch developed interference theory (see, e.g., Abernethy, 1940; Burri, 1931; Dulsky, 1935; Farnsworth, 1934; Pan, 1926; Pessin, 1932; Reed, 1931), considerable evidence has since amassed supporting McGeoch's contention that reinstating the encoding context benefits retrieval (Bilodeau & Schlosberg, 1951; Godden & Baddeley, 1975; Hockley,

Bancroft, & Bryant, 2012; Isarida, Isarida, Kubota, Higuma, & Matsuda, 2018; S. M. Smith, 1979; Smith et al., 1978; Smith & Manzano, 2010; for reviews, see Baddeley, 1982; Bjork & Richardson-Klavehn, 1989; Capaldi & Neath, 1995; Klein, Shiffrin, & Criss, 2007; S. M. Smith, 1988, 1994; Smith & Vela, 2001).

To summarize interference theory, then, memory failure results from four—not mutually exclusive but, in some cases, independently sufficient—factors: (1) use of a cue that is not sufficiently associated to the target, (2) cue overload, (3) occlusion, and (4) a mismatch between the encoding and retrieval contexts. Today, interference theory is the dominant account of retrieval failure in long-term memory, with numerous quantitative and qualitative theories appealing to mechanisms of interference to explain memory phenomena (e.g., Malmberg & Shiffrin, 2005; Postman & Underwood, 1973; Raaijmakers & Jakab, 2013; Raaijmakers & Shiffrin, 1980, 1981; Reder et al., 2000; Rundus, 1973; Shiffrin, 1970; Underwood, 1957; Verde, 2012, 2013; for one exception, see Hardt, Nader, & Nadel, 2013).[2]

1.1.3. Inhibition and repression

Although I am primarily concerned with interference theory in the present paper, two other accounts of memory failure bear mentioning. Anderson (2003) has argued that, in addition to interference, memory failure can also be caused by inhibition (see also

2. It should be noted that the dominance of interference theory does not extend to short-term and working memory, where some scholars continue to regard decay as a viable account of retrieval failure (e.g., Baddeley, 2012; Burgess & Hitch, 1999; Cowan, 1999; Page & Norris, 1998).

Norman, Newman, & Detre, 2007; Weiner, 1968). According to this account, an inhibitory mechanism is deployed (consciously or unconsciously) to lower the associative strength between the cue and occluding competitors. For example, if given the cue "fruit" and the target stem "a____", the correct response—"apple"—can be retrieved more efficiently by lowering the associative strengths between "fruit" and strong competitors such as "banana" and "orange". This has been empirically tested using the retrieval-practice paradigm (Anderson, Bjork, & Bjork, 1994), in which practicing retrieval of some items depresses memory for related items, a pattern known as retrieval-induced forgetting. I discuss retrieval-induced forgetting in more detail in Chapter 3. For now, note that whether retrieval-induced forgetting actually provides evidence for an inhibitory mechanism is a point of contention among memory theorists (Anderson & Spellman, 1995; Camp, Pecher, & Schmidt, 2007; Hulbert, Shivde, & Anderson, 2012; Verde, 2012).

An account of retrieval failure related to inhibition appeals to Freudian repression (Freud, 1895/1964) (Anderson & Levy, 2002; Anderson & Green, 2001; Erdelyi, 2006). The question of whether inhibition and repression occur in memory is currently being debated (Anderson, 2007; Anderson & Hanslmayr, 2014; Anderson & Levy, 2007, 2009; Bäuml, Pastötter, & Hanslmayr, 2010; Bjork, 2007; Hu, Bergström, Gagnepain, & Anderson, 2017; Jonker, Seli, & MacLeod, 2013, 2015; Kihlstrom, 2002; Levy & Anderson, 2011; MacLeod, 2007a, 2007b; MacLeod, Dodd, Sheard, Wilson, & Bibi, 2003; Raaijmakers & Jakab, 2013a, 2013b; Storm, 2011; Verde, 2012, 2013); importantly, because most (if not all) inhibition and repression theorists accept interference as a partial cause of forgetting, this debate is not pertinent to the present

discussion. Nevertheless, inhibitory accounts of the list-strength effect (discussed later) exist (Bäuml, 1997), so inhibition theory will be revisited at that time.

1.2. List-Length and List-Strength Effects

In the following two chapters, I turn to two related interference phenomena: the list-length effect (e.g., Strong, 1912) and the list-strength effect (e.g., Ratcliff, Clark, & Shiffrin, 1990). Studies exploring the list-length effect are concerned with how memory is affected when the length of a study list is increased. Ceteris paribus, increasing the number of to-be-remembered stimuli decreases the proportion of stimuli remembered (Ebbinghaus, 1885/1913; Strong, 1912; Underwood, 1978). It is worth noting, however, that although the proportion of items remembered decreases, the total number of items remembered tends to be higher on long lists than short lists (Beaman, 2006). A phenomenon related to the list-length effect is the list-strength effect. Rather than increasing the number of to-be-remembered stimuli, experiments on the list-strength effect increase the strength of a subset of the to-be-remembered stimuli; performance is then compared to control lists of all strengthened or unstrengthened items (Ratcliff et al., 1990) (sometimes, one of the control lists is omitted; e.g., Hirshman, 1995; Tulving & Hastie, 1972). Experimenters accomplish strengthening by presenting strong items more often than weak items, providing more study time for strong items than weak items, or inducing elaborative processing of strong items and shallow processing of weak items (for a comparison of strengthening methods in the list-strength effect, see Malmberg & Shiffrin, 2005). A list-strength effect is said to occur if strengthening a subset of items yields a mnemonic benefit for the strengthened items and a mnemonic cost for the unstrengthened items, relative to the control lists. Using the interference terms described

above, list-length manipulations can be thought of as cue-overload manipulations, and list-strength manipulations can be thought of as occlusion manipulations.

Prima facie, one might suspect that the list-length and list-strength effects would be highly correlated, with paradigms producing list-length effects yielding list-strength effects, and with paradigms failing to produce list-length effects failing to produce list-strength effects. Indeed, this was the prediction made by a class of models known as global-matching models, examples of which include Search of Associative Memory (SAM; Gillund & Shiffrin, 1984; Mensink & Raaijmakers, 1988, 1989; Raaijmakers & Shiffrin, 1980, 1981), MINERVA 2 (Hintzman, 1984, 1986, 1987, 1988; Hintzman & Hartry, 1990), Composite Holographic Associative Recall Model (CHARM; J. M. Eich, 1982, 1985; Metcalfe, 1990, 1993), Theory of Distributed Associative Memory (TODAM; Murdock, 1982, 1983, 1989, 1992, 1997), and the Matrix Model (Humphreys, Bain, & Pike, 1989; Pike, 1984) (for reviews, see Clark & Gronlund, 1996; Humphreys, Pike, Bain, & Tehan, 1989). Algorithmically, early instantiations of global-matching models treated item repetitions (or increased study time of some items relative to others) as equivalent to adding a new mnemonic trace to memory. Consequently, strengthening some items was the same as adding new items to memory, and thus list-strength manipulations were de facto list-length manipulations. However, empirical tests of the co-occurrence prediction produced cases in which a list-length effect occurred without a list-strength effect (Ratcliff et al., 1990) or a list-strength effect occurred without a list-length effect (Buratto & Lamberts, 2008). Shiffrin, Ratcliff, and Clark (1990) showed that, without modification, no global-matching model could predict the presence of a list-length effect and the absence of a list-strength effect in recognition (see also Clark &

Gronlund, 1996; Criss & Howard, 2015; Criss & Koop, 2015; Murdock & Kahana, 1993a; Shiffrin & Raaijmakers, 1992).

For present purposes, I am more concerned with the list-strength effect than the list-length effect. However, a review of both literatures is warranted for two reasons: First, the list-strength effect arose from work on the list-length effect (see Tulving & Hastie, 1972). Second, accounts of the list-length effect—in particular, Dennis and Humphreys' (2001) Bind Cue Decide Model of Episodic Memory (BCDMEM) and Murdock and Kahana's (1993a, 1993b) continuous-memory version of TODAM—provide accounts of the list-strength effect that are more easily understood through the lens of the list-length effect.

Chapter 2: The List-Length Effect

Strong (1912) conducted one of the earliest experiments on the list-length effect. Subjects in his experiments studied full-page newspaper advertisements; at test, they were asked to differentiate between studied and unstudied advertisements in an old/new recognition test. Strong used study lists of 5, 10, 25, 50, 100, and 150 advertisements. Although he did not report inferential statistics, his descriptive statistics show that accuracy decreased as list length increased.

Since Strong (1912), the list-length effect has been documented in numerous memory paradigms, including learning to criterion (Ceraso, 1967; Lyon, 1914a, 1914b, 1914c; Robinson & Darrow, 1924; Robinson & Heron, 1922), free recall (Gillund & Shiffrin, 1981; Grenfell-Essam, Ward, & Tan, 2013; Howell, 1973; Roberts, 1972; Shiffrin, 1973; Tabachnik & Brotsky, 1976; Ward, 2002), serial recall (Grenfell-Essam, Ward, & Tan, 2017; Haberlandt, Thomas, Lawrence, & Krohn, 2005; Penney, 1985; Spurgeon, Ward, Matthews, & Farrell, 2015; Ward, Tan, & Grenfell-Essam, 2010), serial recognition (Scarmeas et al., 2004), cued recall (Aue, Criss, & Fischetti, 2012; Calfee & Atkinson, 1965; Ceraso, Bader, & Silverstein, 1970; Davis, 1966; Nobel & Shiffrin, 2001; Samuels, 1970; Schumsky, Grasha, Trinder, & Richman, 1969), item recognition (Anderson & Revelle, 1994; Atkinson & Joula, 1973, 1974; Beth, Budson, Waring, & Ally, 2009; Bowles & Glanzer, 1983; Bowyer, Humphreys, & Revelle, 1983; Brandt, 2007; Brandt, Zaiser, & Schnuerch, 2019; Cary & Reder, 2003; Erber, 1974; Gillund & Shiffrin, 1984; Gronlund & Elam, 1994; Nobel & Shiffrin, 2001; Ohrt & Gronlund, 1999; Ratcliff & Murdock, 1976; Schulman, 1974; Toga, 1975; Underwood, 1978; Willis & Underwood, 1983; Wolk, Gold, Signoff, & Budson, 2009; Yonelinas, 1994), associative

recognition (Criss & Shiffrin, 2004b; Kinnell & Dennis, 2012), and source recognition (Glanzer, Hilford, & Kim, 2004). The overwhelming number of demonstrations of the list-length effect may give the impression that its existence is unquestionable. Indeed, Glanzer et al. identified list-length manipulations as "a variable that reliably affects item-recognition memory" (p. 1184), on par with well-established recognition phenomena including the repetition effect (e.g., Hilford, Glanzer, & Kim, 1997), levels-of-processing effect (e.g., Glanzer, Kim, Hilford, & Adams, 1999), word-frequency effect (e.g., Allen & Garton, 1968), and the decline in recognition accuracy with aging (e.g., Gordon & Clark, 1974). However, overlooked methodological issues with the standard experimental approach have led some researchers to question whether the list-length effect is a real phenomenon (Ceraso, 1967; Dennis & Humphreys, 2001; Henmon, 1917; Hipple, 1972; Murdock & Kahana, 1993b). The existence of the list-length effect has proven particularly contentious in recognition (Annis, Lenes, Westfall, Criss, & Malmberg, 2015; Criss & Shiffrin, 2004a; Dennis & Humphreys, 2001; Murdock & Kahana, 1993b; Shiffrin, Ratcliff, Murnane, & Nobel, 1993). Dennis and Humphreys noted four confounds that could be artificially producing a list-length effect: differences in the study-test lag between short and long lists, higher likelihood of attentional lapse for long lists than short lists, rehearsal borrowing, and more effective context reinstatement for short compared to long lists. I discuss these confounds next.

2.1. Confounds in the List-Length Paradigm

2.1.1. Study-test lag

The study-test lag is a critical factor to consider in list-length experiments. If the retention interval is the same for short and long lists, then, on average, less time elapses

between studying and testing of a short-list item compared to a long-list item. It is conceivable, then, that the memory advantage observed on short lists compared to long lists is a function of study-test lag rather than list length.

The solution to the study-test lag confound is not as simple as equating the study-test lag between the start of the study and test phases. For example, if the long list takes twice as long to present as the short list, then items studied in the second half of the long list will still have a shorter study-test lag than any of the items on the short list. Two paradigms have been developed to address this confound: the retroactive design and the proactive design. In the retroactive design, the time between the start of the study and test phases is equated between short and long lists, but only items from the start of the long list are tested. So, if the short list has n_{short} items and the long list has n_{long} items (where $n_{long} > n_{short}$), all items from the short list are tested, but only long-list items studied in positions $1-n_{short}$ are tested. Conversely, in the proactive design, the lag between the end of the study list and start of the test list is equated between conditions, and only the last n_{short} items from the long list are tested.

In recognition, studies using the retroactive design have generally found null or small list-length effects (Bowles & Glanzer, 1983; Buratto & Lamberts, 2008; Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Kinnell & Dennis, 2011; Murnane & Shiffrin, 1991a; Schulman, 1974; Willis & Underwood, 1983). Note, however, that this is not universal: Some experiments employing the retroactive design have still found robust list-length effects (Cary & Reder, 2003; Gronlund & Elam, 1994). Results from experiments using the proactive design are also mixed: Although Dennis and Humphreys (2001), Kinnell and Dennis (2011), and Anderson and Revelle (1994) reported null or

small list-length effects with the proactive design, Glanzer et al. (2004), Ohrt and Gronlund (1999), and Underwood (1978) reported standard list-length effects.

2.1.2. Attentional lapse

The second confound noted by Dennis and Humphreys (2001) has to do with differences in attentional demands required between short and long lists. Clearly, long lists require subjects to maintain attention for a longer time than short lists. According to the attentional-lapse account of the list-length effect, short lists are remembered better than long lists because subjects are more attentive during the short lists. Therefore, this account attributes the list-length effect to processes at encoding rather than retrieval. To the best of my knowledge, the attentional-lapse account was first posited by Henmon (1917; but see also Underwood, 1978). Attentional lapse is particularly problematic in the proactive design, since this design compares performance on end-of-list long-list items to all of the items from the short list. If long-list subjects' attention has waned by the end of the study list, then the proactive-design list-length effect may stem from poorer encoding of critical items in the long list compared to the short list, rather than proactive interference.

Despite its long history (Henmon, 1917; Underwood, 1978) and its intuitive appeal, the attentional-lapse hypothesis has little, if any, direct support in the literature. Two assumptions underpin the attentional-lapse hypothesis: (1) subjects' attention to studying stimuli decreases as the length of the study phase increases, and (2) decreasing attention results in poorer memory. In recent years, mind wandering has emerged has an active research area in cognitive science (for reviews, see Donhoff & Fox, 2015; Kane & McVay, 2012; Mooneyham & Schooler, 2013; Randall, Oswald, & Beier, 2014; Schooler

et al., 2014; Seli, Risko, Smilek, & Schacter, 2016; Smallwood & Schooler, 2015). In mind-wandering research, experimenters ask subjects to self-report mind wandering while they perform an attention task. Typically, this is accomplished through the use of thought probes—that is, periodic inquiries about current mind wandering during the experimental session (e.g., "Are you mind wandering now?") (Seli, Jonker, Cheyne, Cortes, & Smilek, 2015; Seli et al., 2014; Smallwood, Beach, Schooler, & Handy, 2008). Memory experiments using this paradigm have demonstrated that memory decreases as mind wandering increases (Maillet & Rajah, 2013; Thomson, Smilek, & Besner, 2014; for a review on the detrimental effects of mind wandering in educational settings, see Szpunar, 2017), thereby supporting the second assumption of the attentional-lapse account.

To the best of my knowledge, no experiments have compared subjects' propensity for mind wandering on short and long lists. As a result, direct support for the first assumption of the attentional-lapse account is lacking. It should be noted that behaviourally testing this assumption is quite complicated. Imagine, for example, that thought probes are issued every fifth item. In that case, long-list subjects would receive a greater number of thought probes than short-list subjects. However, further spacing thought probes for long-list subjects introduces a different confound. Other indices of mind wandering would therefore be more appropriate in a list-length experiment. Observable indices of mind wandering include amount of fidgeting (Carriere, Seli, & Smilek, 2013; Farley, Risko, & Kingstone, 2013; Seli et al., 2014) and pupil dilation (Franklin, Broadway, Mrazek, Smallwood, & Schooler, 2013); in addition, there are mind-wandering signatures from electroencephalography (Jin, Borst, & van Vugt, 2019;

van Son et al., 2019) and functional magnetic resonance imaging (Christoff, Gordon, Smallwood, Smith, & Schooler, 2009).

2.1.3. Rehearsal borrowing

Dennis and Humphreys (2001) proposed rehearsal borrowing as another potential confound in research on the list-length effect. When the retroactive or proactive designs are used, all short-list items are tested, but only a subset of items from the long list are tested. Rehearsal borrowing occurs when subjects rehearse earlier-presented items during the presentation of other items. For short lists, this is a zero-sum game: Studying Item *A*-1 during presentation of Item *A* will increase memory for Item *A*-1 at the expense of Item *A*. Since both Items *A*-1 and *A* are tested, this is not problematic for interpreting the results. However, in the long list, subjects—being unaware that some items will not be tested—may inadvertently focus their effort on untested items at the expense of tested items. A related concern arises in the retroactive design, in which short-list subjects experience a longer study-test lag than long-list subjects. Motivated subjects may spend the retention interval rehearsing study items, thus enhancing their memorability.

2.1.4. Context reinstatement

Dennis and Humphreys' (2001) final confound concerns context reinstatement. As described in Chapter 1, McGeoch (1931a) proposed that the mnemonic trace and retrieval cue always contain contextual information. Memories associated to the same or similar contexts as that contained in the retrieval cue are easier to access than memories associated to less similar contexts. Importantly, even within a given study episode, context is not stable; rather, it is assumed to fluctuate with the passage of time (Estes, 1955a, 1955b; Malmberg & Shiffrin, 2005; Mensink & Raaijmakers, 1988, 1989). Ceteris

paribus, two adjacent study items will be associated to more similar contexts than two widely separated study items, and start-of-test context will be more similar to end-of-study context than middle- or start-of-study context. This has important implications for the retroactive design, wherein long-list subjects are tested on items from the start of the study phase. Following the retention interval, subjects may attempt to reinstate study context. Because end-of-study context will be easier to reinstate than start- or middle-of-study context, this is the context most likely to be mentally reinstated. Critically, short-list subjects are tested on end-of-study items but long-list subjects are not; consequently, list-length effects may reflect an intuitively reasonable strategy (reinstating study context) that happens to penalize long-list subjects relative to short-list subjects.

2.2. Testing the Confounds

After describing these potential confounds, Dennis and colleagues (Dennis & Chapman, 2010; Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, 2011; Maguire, Humphreys, Dennis, & Lee, 2010) conducted a number of list-length experiments to assess their merit. An illustrative example is provided by Experiment 1 of Dennis and Humphreys' (2001) original paper. In their experiment, subjects studied 24 and 72 words in the short- and long-list conditions, respectively. Subjects were not forewarned of the memory test, a methodological precaution that Dennis and Humphreys reasoned would minimize the probability of rehearsal borrowing. To maintain subjects' attention across the study phase, Dennis and Humphreys asked subjects to make a pleasantness judgment for each item. They also reasoned that focusing subjects' attention on each item as it was presented would further decrease the probability of rehearsal borrowing. In an attempt to attenuate the effects of context fluctuation, long-list subjects

took part in a puzzle task between each third of the study list. Although this methodological precaution may not entirely eliminate context fluctuation, it should at least increase the uniformity of context throughout the study phase (e.g., items in Positions 24, 48, and 72 were all followed by the same puzzle task). Dennis and Humphreys also increased the difficulty of context reinstatement by including an 8 min retention interval. If context fluctuates across time, then lengthening the time between the study and test phases should make context reinstatement far more difficult than in experiments using more typical retention-interval durations (e.g., 2 min). Finally, note that there were two short-list conditions: one in which subjects began the session by studying 24 words, and one in which participants started with a puzzle task. For the first group, the study phase took place at the same time as the first 24 items of the long list; for the second group, the study phase took place at the same time as the final 24 items in the long list. Long-list subjects were tested on items from both the first and final third of the study list. This allowed Dennis and Humphreys to assess the list-length effect in the retroactive design by comparing start-of-session short-list subjects' performance to discrimination of first-third targets from distractors in the long-list condition, and to assess the list-length effect in the proactive design by comparing delayed-study short-list subjects' performance to discrimination of end-of-study targets from distractors. Critically, and consistent with the argument that list-length effects are methodological artifacts, results did not produce a list-length effect.

2.3. Theoretical Interpretation

Dennis and colleagues (Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, 2011, 2012) interpreted the null list-length effect as support for BCDMEM

(see Dennis & Humphreys, 2001). According to BCDMEM, identifying a recognition probe as "old" or "new" involves determining whether the study context is one of the contexts in which the probe has been previously encountered. In BCDMEM, then, interference is caused entirely by previous contexts with which items have been associated, rather than other items from the same study episode. As a result, BCDMEM does not predict that lengthening a study list will affect recognition.

BCDMEM's assumption that interitem interference does not occur in recognition represents an important departure from traditional recognition models. In the global-matching models and their descendants (e.g., Clark & Gronlund, 1996; Criss & Shiffrin, 2004a; Gillund & Shiffrin, 1984; Hintzman, 1987, 1988; Shiffrin & Steyvers, 1997), interference in recognition arises from two sources: item noise (i.e., other items on the study list) and context noise (i.e., as in BCDMEM). Dennis and Humphreys' (2001) finding of a null list-length effect in recognition, then, is consistent with context-noise models and inconsistent with context-plus-item-noise models.

Two simplifying assumptions common to most global-matching models are partially responsible for the prediction of the list-length effect in recognition. First, pre-experimental memories are assumed not to interfere with information learned during a study phase. Second, previous lists from the experimental session are assumed not to interfere with subsequent lists. So, according to most instantiations of global-matching models, if an experimental session involves multiple study-test cycles, neither items studied on earlier lists, nor memories from prior to the experimental session, affect performance on the test list. Indeed, the search set (i.e., the mnemonic traces searched during the test phase) are assumed to consist entirely of mnemonic traces generated

during the study phase. Prima facie, these assumptions seem unsound, inasmuch as they render subjects' minds de facto tabula rasae. Moreover, researchers have demonstrated that interference on recognition tests can be caused by previous study lists (Criss et al., 2011) and experimental instructions (Curtis et al., 2016). However, in defence of the tabula-rasa view, because retrieval cues contain current context, it is reasonable to assume that, by and large, pre-experimental memories are excluded from the search set and that previous study lists exert much less interference than the present list. It is also important to point out that proponents of global-matching models would not seriously argue for the tabula-rasa view; rather, it is implemented in computational models to make the simulations faster, and because the inclusion of pre-experimental memories would not typically affect the qualitative pattern of the models' predictions.

As an alternative to the tabula-rasa view, Murdock and Kahana (1993a) presented what they termed the continuous-memory version of TODAM. In this version, mnemonic traces from earlier lists and prior to the experiment are present during the recognition test. This is implemented by initializing simulated subjects' memories with pre-experimental noise (i.e., a computational representation of pre-experimental memories). Shiffrin et al. (1993) demonstrated that this causes TODAM to no longer predict a list-length effect, since increases in list length produce almost no additional noise when compared to the lifetime of experiences already present in the search set. Shiffrin et al. (1993) considered this to be a limitation (see also the discussion in Shiffrin et al., 1990), but Murdock and Kahana (1993b)—using some of the arguments later leveraged by Dennis and Humphreys (2001)—suggested that the list-length effect had yet to be satisfactorily demonstrated in recognition. It is worth noting, however, that Ohrt and Gronlund (1999) found a list-

length effect in recognition using an experimental design informed by Murdock and Kahana's (1993b) discussion.

2.4. Summarizing the List-Length Effect

Despite the evidence provided by Dennis and colleagues (Dennis & Chapman, 2010; Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, 2011; Maguire et al., 2010), debate still exists concerning whether the list-length effect is a real recognition phenomenon. Although Dennis and colleagues have produced a number of null findings, it remains the case that other researchers have found list-length effects with some or all of Dennis and colleagues' controls implemented (e.g., Cary & Reder, 2003; Criss & Shiffrin, 2004a; Underwood, 1978). Annis et al. (2015) have called into question the Bayesian analyses Dennis et al. (2008) presented to support their claim that their results were consistent with context-noise models (like BCDMEM) and inconsistent with item-noise models (e.g., REM). Instead, Annis et al. showed that Dennis et al.'s results were consistent with a small, but real, list-length effect. Additional data is probably necessary to settle this debate. At present, my position is that the list-length effect exists, but that it is small when Dennis and Humphreys' (2001) methodological precautions are taken.

Chapter 3: The List-Strength Effect

Research on the list-strength effect emerged from research on the list-length effect. In list-length experiments, researchers are interested in how adding items to a list affects memory; in list-strength experiments, researchers are interested in how repeating some items affects memory. Tulving and Hastie (1972) reported the first experiment on the list-strength effect. In Experiment 1 of their paper, they manipulated both list length and list strength. Free recall was tested across three list types: a short pure-weak list, on which 10 items were presented once; a long pure-weak list, on which 15 items were presented once; and a mixed list, on which five items were presented once and five items were presented twice. As one would expect, memory was better for the short pure-weak list than the long pure-weak list, reflecting a list-length effect. Also unsurprising is that memory was better for the twice-presented items than the once-presented items. Interestingly, though, memory was worse for the once-presented items on the mixed list than the items on the short pure-weak list, demonstrating that the presence of strengthened items on the mixed list produced a mnemonic cost for the unstrengthened items. Subsequent work showed that the strengthened items also incur a mnemonic benefit from the unstrengthened items—that is, free recall is better for repeated items on a mixed list than a list on which all items are repeated (Malmberg & Shiffrin, 2005). Although the term did not exist when Tulving and Hastie conducted their experiments, Ratcliff et al. (1990) and Shiffrin et al. (1990) later termed this pattern the list-strength effect.

In Chapter 1, I described two interference phenomena: cue overload and occlusion. The list-strength effect can be thought of in terms of occlusion: On the mixed

list, the strengthened items "stand out" from the unstrengthened items, thereby enhancing their memorability. At the same time, the unstrengthened items are occluded by the strengthened items, thereby diminishing their memorability.

Several free- and cued-recall experiments followed Tulving and Hastie's (1972) initial demonstration of the list-strength effect (Fritzen, 1975; Hastie, 1975; Mueller & Brown, 1977; Robbins, Bray, & Irvin, 1974). However, it was not until a series of experiments by Ratcliff et al. (1990) that the list-strength effect was investigated in recognition. Ratcliff et al. conducted seven recognition experiments on the list-strength effect. Interestingly, the results from free recall did not extend to recognition: Although strong items were better recognized than weak items, the effect did not differ as a function of list type. Indeed, several of their experiments found a trend in the opposite direction, with strong-item performance better on the pure list and weak-item performance better on the mixed list. These results were replicated by several subsequent papers (Hirshman, 1995; Murnane & Shiffrin, 1991a; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Yonelinas, Hockley, & Murdock, 1992). This discrepancy between free recall and recognition came as a surprise: Shiffrin et al. (1990) showed that, without modification, none of the global-matching models could predict the absence of a list-strength effect in recognition (see also Clark & Gronlund, 1996).

Ratcliff et al.'s (1990) demonstration of a null list-strength effect in recognition produced a number of theoretical explanations. I describe these later in the present chapter. First, however, I describe the three paradigms used to study the list-strength effect: the A+2B, mixed-pure, and strong-interference paradigms.

3.1. Paradigms Used for the List-Strength Effect

3.1.1. The *A*+2*B* paradigm

Early studies on the list-strength effect used what Tulving and Hastie (1972) termed the *A*+2*B* paradigm. The *A*+2*B* paradigm divides targets into *A* and *B* items. On the *A*+*B* list, both *A* and *B* items are presented once; on the *A*+2*B* list, *A* items are presented once and *B* items are presented twice. A list-strength effect is present if memory is better for the *A* items on the *A*+*B* list relative to *A* items on the *A*+2*B* list. Studies using the *A*+2*B* paradigm have largely tested free recall, and thus they have typically yielded a list-strength effect (Fritzen, 1975; Hastie, 1975; Mueller & Brown, 1977; Sahakyan, Abushanab, Smith, & Gray, 2014; Tulving & Hastie, 1972).

3.1.2. The mixed-pure paradigm

The mixed-pure paradigm, introduced by Ratcliff et al. (1990), can be thought of as an extension of the *A*+2*B* paradigm. Both the *A*+*B* and *A*+2*B* lists are retained, although their names are changed to the pure-weak and mixed lists, respectively. There is also a mixed-strong list, on which all items are strengthened (in terms of the *A*+2*B* paradigm, the pure-strong list could be thought of as a 2*A*+2*B* list). Therefore, experiments using the mixed-pure paradigm have a 2 (strength: weak vs. strong) × 2 (list type: pure vs. mixed) design. Unlike experiments using the *A*+*2B* paradigm, which, to the best of my knowledge, have exclusively strengthened *B* items through multiple, spaced repetitions, experiments using the mixed-pure paradigm have explored various strengthening techniques, including multiple presentations of strong items, longer study time for strong items, and more elaborative processing of strong items (see Malmberg & Shiffrin, 2005).

Although the addition of the pure-strong list is typically considered to be the only difference between the *A*+2*B* and mixed-pure paradigms, one other difference bears

mentioning. In the *A*+2*B* paradigm, the critical comparison is between *A* items from the *A*+*B* list and *A* items from the *A*+2*B* list. In other words, performance on only half of the items from the *A*+*B* list is considered when evaluating the presence of a list-strength effect. For example, if a subject in Tulving and Hastie's (1972) experiment recalled one *A* item and four *B* items, their overall performance on the *A*+*B* list would be 5/10 = .5, but their performance for comparison to the *A*+2*B* list would be 1/5 = .2. In contrast, in the mixed-pure paradigm, performance on all pure-weak items is compared to performance on weak items from the mixed list, and performance on all pure-strong items is compared to performance on all strong items from the mixed list.

Somewhat confusingly, two empirical patterns resulting from the mixed-pure paradigm have been termed a list-strength effect. In the first, a list-strength effect is said to occur when performance is better for mixed-strong items than pure-strong items, and performance is better for pure-weak items than mixed-weak items. I will call this the crossover list-strength effect because it arises when there is a crossover interaction between strength and list type.

The second empirical pattern defines the list-strength effect in terms of what Ratcliff et al. (1990) and Shiffrin et al. (1990) called the ratio of ratios (R_r). The R_r is the ratio of strong-to-weak performance ratios in the mixed and pure lists. Let m_{MW}, m_{MS}, m_{PW}, and m_{PS} denote memory performance for mixed-weak, mixed-strong, pure-weak, and pure-strong items, respectively. Then:

$$R_r = (m_{MS}/m_{MW})/(m_{PS}/m_{PW})$$

A positive list-strength effect exists if $R_r > 1$, a null list-strength effect exists if $R_r = 1$, and a negative list-strength effect exists if $R_r < 1$. Notice that, if there is a crossover

list-strength effect, the R_r is necessarily positive, but that the R_r can still be positive in the absence of a crossover list-strength effect. As an example, consider a free-recall experiment on which the mean proportion of mixed-strong, mixed-weak, pure-strong, and pure-weak items recalled is .6, .3, .6, and .4, respectively. Here, the R_r is (.6/.4)/(.6/.4) = 1.333. So, these means represent a positive list-strength effect, as indexed by the R_r. However, they do not represent a crossover list-strength effect, since there is not a crossover interaction between strength and list type.

3.1.3. The strong-interference paradigm

Norman (1999, 2002) introduced the strong-interference paradigm as an alternative to the mixed-pure paradigm. In the strong-interference paradigm, study lists are composed of tested and untested items, called targets and interference items, respectively. There are two list types: the weak-interference list and the strong-interference list. The targets are presented once on each list. However, interference items are presented multiple times on the strong-interference list and once on the weak-interference list. At test, only targets are tested, and a list-strength effect occurs if performance is better on the weak-interference list than the strong-interference list.

I will have more to say about the strong-interference paradigm in a later section. For now, it bears mentioning that, unlike the mixed-pure paradigm, the strong-interference paradigm typically yields a list-strength effect in recognition (Diana & Reder, 2005; Norman, 1999, 2002; Norman, Tepe, Nyhus, & Curran, 2008; Osth, Dennis, & Kinnell, 2014; but see Osth, Fox, et al., 2018).

3.2. Theoretical Accounts of the List-Strength Effect

3.2.1. Differentiation

Experts are able to capitalize on information that laypeople cannot, thereby giving them an advantage in distinguishing among similar stimuli. For example, violins and violas look very similar and, to a layperson, are likely indistinguishable without research and/or assistance from an expert. Yet, experts can easily categorize them as different instruments. Unlike laypeople, experts know that violas have slightly longer strings than violins, that violas have a lower note range than violins, and that violins and violas have distinct timbres. Indeed, some experts will be able to listen to a piece of music and easily identify its likely composer and composition era. The critical difference between laypeople and experts is that, unlike laypeople, experts are able to draw fine distinctions between and within categories. This ability—called differentiation—has a long history in psychological research (Gibson, 1940; Gibson & Gibson, 1955; Goldstone, 1998; Nosofsky, 1987; Saltz, 1963), and has been applied to a number of domains of expertise, including reading sheet music (Waters & Underwood, 1998), predicting the movement of billiards balls (Crespi, Robino, Silva, & de'Sperati, 2012), fingerprint identification (Searston & Tangen, 2017), identifying the best move in a chess game (Sheridan & Reingold, 2014), discriminating among wines (Foroni et al., 2017; Parr, Heatherbell, & White, 2002), and identifying clinically relevant abnormalities in chest X-rays (Myles-Worsley, Johnston, & Simons, 1988).

Does differentiation play a role in memory? One phenomenon that implicates differentiation is the so-called paradox of interference (see Smith, Adams, & Schorr, 1978). Recall that, as the number of memories subsumed by a retrieval cue increases, the retrieval cue's efficacy diminishes, a phenomenon called cue overload (Surprenant & Neath, 2009, Chapt. 5; O. C. Watkins & Watkins, 1975). Notice that cue overload implies

that, as one learns more about a topic, one's ability to retrieve facts about that topic will become increasingly difficult. For example, a strict interpretation of cue overload would predict that students who have just completed an introductory neuroscience course will be faster at naming the parts of a neuron than their professor, for whom the cues "neuroscience" or "neuron" subsume far more information.

Differentiation provides a solution to the paradox of interference: As experience with a topic increases, more specific cues can be utilized to retrieve desired information. For example, in retrieving information about cardinals, laypeople will probably have to use the cue "cardinal", but experts will be able to use cues corresponding to different cardinal species (e.g., "vermilion cardinal", "northern cardinal", "red-crested cardinal", etc.). Notice, too, that encoding also benefits from differentiation: For laypeople, a picture of a vermilion cardinal is likely to be given a verbal label such as "cardinal" or "red bird", but experts will easily encode the picture with a label such as "female vermilion cardinal" or "unusually large vermilion cardinal". If a picture of a northern cardinal is then shown, laypeople will probably end up with the same label used for the vermilion cardinal, but experts will generate a mnemonic trace corresponding to a separate cue. So, experts' ability to draw finer distinctions within and between categories allows them to increase the specificity of both their encoded mnemonic traces and their retrieval cues.

One memory phenomenon for which differentiation has been used as an explanation is the strength-based mirror effect. In recognition, a mirror effect is said to occur when an increase in the hit rate is accompanied by a decrease in the false-alarm rate (Glanzer & Adams, 1985, 1990). Mirror effects have generally been divided into stimulus-based mirror effects and strength-based mirror effects. In the stimulus-based

mirror effect, targets and distractors are evenly divided between two types of stimuli, and one of the stimulus types produces a higher hit rate and lower false-alarm rate, relative to the other stimulus type. Manipulations that produce a stimulus-based mirror effect include low- versus high-frequency words (Reder et al., 2000), concrete versus abstract words (Hockley, 1994), and pictures versus words (Snodgrass, Wasser, Finkelstein, & Goldberg, 1974). In contrast, the strength-based mirror effect tests subjects' recognition of two list types: a weakly encoded list and a strongly encoded list. These lists are analogous to the pure-weak and pure-strong lists in the mixed-pure paradigm. Unsurprisingly, hit rates are higher for strongly encoded lists than weakly encoded lists; more surprising is that the false-alarm rate is lower on the strongly encoded list than the weakly encoded list (Bruno, Higham, & Perfect, 2009; Criss, 2006, 2009, 2010; Criss, Aue, & Kılıç, 2014; Criss & McClelland, 2006; Criss, Wheeler, & McClelland, 2013; Hockley & Niewiadomski, 2007; Kılıç & Öztekin, 2014; Koop, Criss, & Pardini, 2019; Starns, Ratcliff, & White, 2012; Starns, White, & Ratcliff, 2010, 2012; Stretch & Wixted, 1998; Wixted & Stretch, 2000).

Stretch and Wixted (1998) appeal to criterion placement to explain the strength-based mirror effect. In item recognition, although measures of discrimination (e.g., d') are most often the dependent variable of interest, some experiments investigate measures of criterion placement (e.g., β). Subjects' criterion placement reflects the amount of evidence they need to call an item "old"; if a probe's activation level (match, familiarity, etc.) exceeds the criterion, an "old" response is made, and if it does not, a "new" response is made. According to the criterion-placement account of the strength-based mirror effect, subjects adopt a stricter criterion (i.e., a more conservative criterion) on a strong test list

than a weak test list. The targets from the strong list are sufficiently well encoded that they exceed this higher criterion, but the higher criterion results in fewer distractors being called "old" than on the weak list since, presumably, the average activation of distractors is the same between lists.

Differentiation can explain the strength-based mirror effect without appealing to criterion placement (Criss & McClelland, 2006; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Most memory models assume that studying an item produces an error-prone, mnemonic trace of the original stimulus. Models also generally assume that the fidelity of the mnemonic trace increases as encoding time increases. A mnemonic trace can be said to be differentiated when it is better encoded than other, related mnemonic traces. Trace differentiation has two important consequences: First, differentiated traces are more distinctive than related items, thus decreasing the potential for them to produce interitem interference. Second, differentiated traces are less confusable with other items, thus decreasing the probability that a distractor on a recognition test will be called "old". The differentiation account, then, proposes that the strength-based mirror effect is a necessary outcome of better encoding: The hit rate is higher for strong items than weak items because their mnemonic traces are more accurate representations of the studied items, and the false-alarm rate is lower on strong lists than weak lists because the greater distinctiveness of strong mnemonic traces decreases the probability of erroneous distractor matches.

What about the list-strength effect? Although it is often said that the list-strength effect tends to be null in recognition, it actually tends to be slightly negative (i.e., $R_r < 1$; see Ratcliff et al., 1990, 1992, 1994). When a negative list-strength effect occurs,

performance is better on mixed-weak items than pure-weak items, but worse on mixed-strong items than pure-strong items. Under a differentiation framework, weak items produce more interitem interference than strong items, and thus the negative list-strength effect has a straightforward explanation: Mixed-weak items are protected to a greater extent from interitem interference than pure-weak items, since half of their competitors are differentiated. Similarly, pure-strong items are better protected from interitem interference than mixed-strong items, since all of their competitors are differentiated.[3]

When Ratcliff et al. (1990) first documented the null or negative list-strength effect in recognition, Shiffrin et al. (1990) demonstrated that no extant model could simultaneously predict a positive list-length effect and negative list-strength effect in recognition. Shiffrin et al. therefore added a differentiation assumption to SAM, which allowed it to predict the requisite pattern. In earlier instantiations of SAM (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1980, 1981), strengthened items were stored in multiple traces—for example, if weak items were presented for 1 s and strong items were presented for 5 s, weak items would have one mnemonic trace and strong items would have five mnemonic traces. In the differentiation version of SAM, however, strengthened items were stored in a single trace, which resulted in SAM predicting a negative list-strength effect (see also Shiffrin & Raaijmakers, 1992). Although SAM is no longer considered a viable model of recognition memory, differentiation is still the mechanism by which the null or negative list-strength effect occurs according to REM (Shiffrin &

3. Note that the level of differentiation is always relative to a baseline, and that differentiation is best conceived as a continuum rather than a dichotomy.

Steyvers, 1997) and the Subjective Likelihood Model (SLIM; McClelland & Chappell, 1998) (see also Criss & McClelland, 2006; Criss & Howard, 2015; Criss & Koop, 2015; Kılıç, Criss, Malmberg, & Shiffrin, 2017).

Importantly, if differentiation is responsible for the null or negative list-strength effect, then disrupting differentiation should produce a list-strength effect in recognition. Murnane and Shiffrin (1991b) tested this prediction by embedding to-be-remembered words in sentences during the study phase. In Experiment 1, strong items were repeated in new sentences; in Experiment 2, strong items were repeated in the same sentence. Murnane and Shiffrin reasoned that repeating words in new sentences would result in separate mnemonic traces, but that repeating words in the same sentence would result in a single, differentiated trace. Their predictions were supported, with a list-strength effect observed in Experiment 1 but not in Experiment 2 (for similar results, see Sahakyan & Malmberg, 2018).

3.2.2. Criterion placement

As discussed above, some scholars have attributed the strength-based mirror effect to a criterion shift, with subjects adopting a more conservative criterion on the strong list than the weak list (Stretch & Wixted, 1998). The criterion shift is assumed to be because subjects recognize the greater strength of studied items on the strong list, and thus require a recognition probe to have a greater degree of familiarity to call it “old” than they do following study of the weak list. In the mixed-pure paradigm, then, one might suspect that subjects would vary their criterion from probe to probe on the mixed list according to whether they believe a given probe to be a mixed-strong or mixed-weak item. However, two factors complicate such a strategy: First, because test lists are not blocked by item

strength, it is unclear how subjects would know when to use a conservative criterion and when to use a more liberal criterion. Second, previous work has demonstrated that subjects are reluctant to modify their criterion within a test list (Verde & Rotello, 2007).

So, if subjects' criterion placement is a function of list strength, and if within-list criterion changes are unlikely, then subjects' criterion on a mixed-strength list should be more conservative than their criterion for the pure-weak list, but more liberal than their criterion for the pure-strong list. This is because the average strength of targets on the mixed list will be greater than the average strength of the targets on the pure-weak list and less than the average strength of targets on the pure-strong list. Let N denote the number of targets per list, W denote the presentation duration of weak items, and S denote the presentation duration of strong items. Then, the average target strength, T, for the pure-weak (T_{PW}), mixed (T_M), and pure-strong (T_{PS}) lists is:

$$T_{PW} = WN/N = W$$

$$T_M = (W(N/2)+S(N/2))/N = (W+S)/2$$

$$T_{PS} = SN/N = S$$

Therefore, if criterion placement is indeed a product of study-list strength, subjects' criterion should become increasingly conservative from the pure-weak list to the mixed list to the pure-strong list. Hirshman (1995) argued that the different criteria used across list types in the mixed-pure paradigm is responsible for the absence of the list-strength effect in recognition.

Why does Hirshman's (1995) criterion-placement account predict a null or negative list-strength effect? First, notice that, although the criterion is assumed to change across lists, the actual familiarity of the distractors does not change. In other words, if the

average familiarity of distractors on the pure-weak list is F_d, then the average familiarity of distractors on the mixed and pure-strong lists should also be F_d. This is because distractors across the three lists are presumably drawn from the same population of potential distractors. For example, Ratcliff et al. (1990) randomly sampled stimuli for each subject from the Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982). Therefore, because of the increasingly conservative criterion, the false-alarm rate will be highest on the pure-weak list, lowest on the pure-strong list, and intermediate on the mixed list.

What about target familiarity? Because they will be better encoded, the average familiarity of strong targets (F_{TS}) will exceed the average familiarity of weak targets (F_{TW}), but F_{TS} and F_{TW} will not differ between pure and mixed lists. However, although mixed- and pure-weak targets will be equally well encoded, the more liberal criterion on the pure-weak list will result in a higher hit rate for pure-weak targets than mixed-weak targets. Similarly, although mixed- and pure-strong targets will be equally well encoded, the more liberal criterion on the mixed list will result in a higher hit rate for mixed-strong targets than pure-strong targets.

It can be difficult to grasp why the criterion-placement account predicts a null list-strength effect. If one considers only the hit rates, then the pattern is consistent with a positive list-strength effect: the hit rate is higher for pure-weak items than mixed-weak items, because F_{TW} is the same across list types but subjects adopt a more conservative criterion on the mixed list than the pure-weak list. At the same time, the hit rate is higher for mixed-strong items than pure-strong items, again because F_{TS} is the same between lists but subjects adopt a more conservative criterion on the pure-strong list than the

mixed list. However, the pattern of hit rates is offset by the pattern of false-alarm rates: Recall that F_D is the same across lists. This means that, because the criterion becomes increasingly conservative as list strength increases, the false-alarm rate is higher on the pure-weak list than the mixed list, and the false-alarm rate is higher on the mixed list than the pure-strong list. The differences in false-alarm rates offset the hit-rate differences, producing roughly equivalent discrimination on weak and strong items across pure and mixed lists.

I find the differentiation account more persuasive than the criterion-placement account for two reasons. First, the criterion-placement account assumes a distribution of hit and false-alarm rates that is difficult to assess based on the published record. Although Hirshman (1995) found hit and false-alarm rates consistent with the criterion-placement account, most other list-strength experiments restrict their analyses to discrimination measures like *d′*. As such, it is unclear whether null list-strength effects are always due to this pattern of hit and false-alarm rates. Second, the differentiation account has generated novel predictions that have been tested and verified (Murnane & Shiffrin, 1991b; Sahakyan & Malmberg, 2018). In contrast, it is unclear what a priori predictions are made by the criterion-placement account, and it is unclear how the criterion-placement account could explain exceptions to the null list-strength effect (e.g., Sahakyan & Malmberg, 2018).

3.2.3. The continuous-memory version of TODAM

As described in the previous chapter, Murdock and Kahana (1993a, 1993b) introduced the continuous-memory version of TODAM. For simplicity, most memory models assume that memory is empty when a study list is first presented. Instead, in the

continuous-memory version of TODAM, the memory matrix consists of all pre-experimental memories when the experiment begins. As subjects move through study-test cycles, memories from previous cycles remain in memory.

Murdock and Kahana (1993a) showed that, with the addition of the continuous-memory assumption, TODAM was able to predict a null list-strength effect in recognition. This is because it renders within-list interference nearly negligible when weighed against a subject's lifetime of memories. As Shiffrin et al. (1993) pointed out, the continuous-memory version of TODAM also predicts a null list-length effect in recognition, although, given recent findings (Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, 2011), this may not be as problematic as once thought.

Even setting aside the controversy regarding the list-length effect in recognition, the continuous-memory version of TODAM does not provide an entirely satisfying solution to the list-strength effect. No scholar would seriously suggest that a subject's memory is empty at the beginning of an experimental session; nor would they suggest that the act of moving from one study-test cycle to the next reinitializes memory. However, we know that context plays a substantial role in limiting the search set. It is perfectly reasonable, then, to assume that pre-experimental memories and mnemonic traces from previous study-test cycles exert minimal interference on the current list, simply because those memories will be associated to different contexts, and thus will be largely excluded from the search set.

3.2.4. BCDMEM

As reviewed in the previous chapter, Dennis and Humphreys' (2001) BCDMEM assumes that recognition decisions are unaffected by interitem interference. Instead, false

alarms occur when subjects erroneously believe they last saw the distractor during the study phase, and misses occur when subjects erroneously believe that the target was last encountered in a context other than the study phase. Therefore, because BCDMEM does not include a mechanism for interitem interference, it predicts a null list-strength effect. Note, however, that BCDMEM can only account for a null list-strength effect, and not the slightly negative list-strength effects that have been observed (Ratcliff et al., 1990).

3.2.5. Insufficient strengthening and dual-process accounts

Norman (1999, 2002) proposed two explanations for negative or null list-strength effects: insufficient strengthening of strong items and a failure to separate recollection from familiarity. According to the insufficient-strengthening hypothesis, extant studies on the list-strength effect had failed to strengthen strong items sufficiently to interfere with weak items. Because strong items are tested in the mixed-pure paradigm, experimenters have been careful to keep strong-item performance below ceiling. Norman (1999, 2002) argued that strengthening strong items to ceiling would produce a list-strength effect, at least in weak-item performance. Norman (1999, 2002) therefore developed the strong-interference paradigm as an alternative to the mixed-pure paradigm. I described the strong-interference paradigm in Section 3.1.3 but, very briefly, this paradigm eliminates the pure-strong list and tests only weak items.

Norman (1999, 2002) also proposed a dual-process interpretation of the list-strength effect. According to dual-process theorists (Jacoby, 1991; Mandler, 1980, 2008; Norman & O'Reilly, 2003; Reder et al., 2000; Tulving, 1985; Yonelinas, 2002), recognition decisions are based on information from two independent sources: familiarity and recollection (for an argument that familiarity and recollection are correlated, rather

than independent, see Moran & Goshen-Gottstein, 2015). Familiarity is a phenomenological sense that an item was studied in the absence of a concrete, episodic memory of studying the probe. In contrast, recollection involves an episodic memory of studying the probe. Although dual-process theorists assume that familiarity and recollection both contribute to recognition decisions, only recollection contributes to free-recall performance (but see Hamilton & Rajaram, 2003). Since free recall produces a positive list-strength effect but recognition does not, Norman (1999, 2002) hypothesized that the list-strength effect was a purely recollection-based phenomenon, and that the combination of familiarity and recollection in recognition was obscuring the list-strength effect.

Norman (1999, 2002) tested the insufficient-strengthening and dual-process accounts of the list-strength effect by coupling the strong-interference paradigm and remember/know paradigm (see Tulving, 1985). In the remember/know paradigm, when subjects identify a probe as "old", they are asked to indicate whether their response was based on "remembering" or "knowing" (recollection and familiarity, respectively). Consistent with the insufficient-strengthening account, Norman (1999, 2002) demonstrated that strengthening interference items to ceiling produced a list-strength effect in the strong-interference paradigm (i.e., better recognition of weak items from the weak-interference list than weak items from the strong-interference list). Consistent with the dual-process account, the list-strength effect occurred in "remember" but not "know" responses. The recollection-based list-strength effect has since been replicated by Diana and Reder (2005) and Norman et al. (2008) (but see Osth, Fox, et al., 2018, for a failed replication).

Two additional findings provide converging support for the view that the list-strength effect is a recollection-based phenomenon. In the switched-plurality paradigm, subjects study a list of singular and plural nouns, and are then tested with targets (i.e., items from the study list), switched-plurality distractors (i.e., plural versions of singular nouns or singular versions of plural nouns), and unrelated distractors (see Hintzman, Curran, & Oppy, 1992). Buratto and Lamberts (2008) found a list-strength effect in recognition using the switched-plurality paradigm. In another study, Verde and Rotello (2004) found a list-strength effect in associative recognition. Importantly, distinguishing between targets and switched-plurality distractors or between intact and rearranged pairs are assumed to rely more on recollection than standard old/new recognition.

3.2.6. Inhibition and output interference

My discussion of the theoretical accounts of the list-strength effect concludes with two accounts that have only been applied in free and cued recall. In particular, I begin by describing the inhibition account of the list-strength effect (Anderson, 2003; Bäuml, 1997), and finish by arguing that output interference provides a more compelling account of the data than inhibition.

The inhibition account relies heavily on similarities between the mixed-pure paradigm and the retrieval-practice paradigm, introduced by Anderson et al. (1994) to study retrieval-induced forgetting. The retrieval-practice paradigm consists of three distinct phases: study, retrieval practice, and test. During the study phase, subjects study a list of words from several distinct categories (e.g., fruits, animals, instruments, etc.). During the retrieval-practice phase, subjects practice retrieving half of the exemplars from half of the categories. This is accomplished by presenting subjects with a category-

exemplar stem (e.g., “instrument”-“p____” for “piano”). In the retrieval-practice literature, the convention is to refer to practiced categories as Rp categories and nonpracticed categories as Nrp categories. Practiced exemplars from Rp categories are denoted Rp+, and nonpracticed exemplars from Rp categories are denoted Rp−. At test, all exemplars from all categories are tested. Unsurprisingly, performance is best for Rp+ exemplars. More importantly, memory is better for Nrp exemplars than Rp− exemplars—in other words, the Rp− exemplars incur a mnemonic cost from the presence of Rp+ exemplars (Anderson et al., 1994; Bäuml, Zellner, & Vilimek, 2005; Cinel, Cortis Mack, & Ward, 2018; Ciranni & Shimamura, 1999; Gómez-Ariza, Pelegrina, Lechuga, Suárez, & Bajo, 2009).

Prima facie, retrieval-induced forgetting appears consistent with interference mechanisms like cue dependence and occlusion. From an interference perspective, the category label serves as the retrieval cue, and subsumes the studied exemplars. In Nrp categories, the cue is, on average, equally associated to each of the exemplars; in Rp categories, the cue is more strongly associated to Rp+ exemplars than Rp− exemplars, and thus the former occlude the latter. Yet, two important findings challenge the interference account of retrieval-induced forgetting.[4] First, it has been demonstrated that retrieval-induced forgetting is retrieval-dependent—that is, simply restudying the Rp+ exemplars, rather than overtly retrieving them, does not result in an Rp− decrement

4. In addition to the two discussed here, Anderson (2003) described two other indices of inhibition: strength independence and competition dependence. However, as these indices do not bear on the issue of the list-strength effect, I do not discuss them here.

(Ceranni & Shimamura, 1999; Hulbert et al., 2012; but see Jonker et al., 2013; Verde, 2013). This is inconsistent with most interference models: Although it is reasonable to assume that retrieval practice results in greater strengthening than restudy, restudying should still result in stronger Rp+ traces than Rp− traces. Therefore, the finding that Nrp and Rp− exemplars are remembered at the same rate, even though Rp+ exemplars exceed Rp− exemplars, is inconsistent with occlusion. From an inhibition perspective, the act of overt retrieval involves suppression of a target's competitors. For example, after studying "kiwi", "pineapple", "lemon", and "orange" in the "fruit" category, completing the category-exemplar-stem "fruit"-"p___" involves suppressing "kiwi", "lemon", and "orange" to retrieve "pineapple". This suppression is necessary because the "fruit" cue activates all four exemplars, not just "pineapple"—coactivation results in suppression. From an inhibition perspective, then, whether or not retrieval practice results in strengthening is secondary—its relative strength is increased by virtue of the fact that competitors are weakened. Therefore, while interference theory attributes retrieval-induced forgetting to occlusion of Rp− exemplars by Rp+ exemplars, inhibition theory attributes retrieval-induced forgetting to a mechanism that actively weakens the strength of Rp− exemplars.

The second property of retrieval-induced forgetting that favours an inhibition explanation is that the mnemonic cost to Rp− exemplars is cue independent. In other words, Rp− exemplars are impaired for the original cue (e.g., "fruit"-"a___" for "apple") and for all other cues (e.g., "red"-"a___" for "apple") (Anderson & Bell, 2001; Anderson & Spellman, 1995; Hulbert et al., 2012; Saunders & MacLeod, 2006; Shivde & Anderson, 2001; Weller, Anderson, Gómez-Ariza, & Bajo, 2013). It is important to emphasize how

extraordinary cue independence is, given the substantial support for the cue-dependence principle (Surprenant & Neath, 2009, Chapt. 3). It is difficult to see how an interference theory of forgetting could incorporate cue-independence (but see Huber, Tomlinson, Jang, & Hopper, 2015; Tomlinson, Huber, Rieth, & Davelaar, 2009, for an attempt using SAM), but the phenomenon is consistent with inhibitory accounts (Anderson, 2003; Anderson & Spellman, 1995). It should be noted that several studies have failed to replicate cue-independent forgetting (Perfect et al., 2004; Williams & Zacks, 2001; see also Bulevich, Roediger, Balota, & Butler, 2006). For present purposes, when I discuss inhibition theory, I will assume that cue-independent forgetting is a real phenomenon; otherwise, however, I remain agnostic regarding its veracity.

Paradigms used to investigate the list-strength effect bear important similarities to the retrieval-practice paradigm. In both, a subset of to-be-remembered stimuli is strengthened, and performance on the unstrengthened items is compared to a baseline, control list. When present, the list-strength effect and retrieval-induced forgetting result in a mnemonic cost for the unstrengthened items. From an inhibition perspective, the most important difference between the paradigms is that, in the retrieval-practice paradigm, inhibition takes place before the test phase, but in list-strength paradigms (i.e., the mixed-pure, *A*+2*B*, and strong-interference paradigms), inhibition only occurs during the test phase. This is because restudy is insufficient to cause inhibition; rather, overt retrieval is required.

Bäuml (1997) was the first to test an inhibition account of the list-strength effect. He argued that the presence of the list-strength effect in free recall, but not cued recall or recognition, was best explained by subjects deploying an inhibitory mechanism during the

test phase. Wike and Wike (1970) and Wixted et al. (1997) have demonstrated that, in mixed-strength lists, subjects, on average, output strong items earlier and faster than weak items. From an inhibition perspective, the retrieval of any item during a free-recall test results in the weakening of the remaining traces. So, by the time subjects begin recalling mixed-weak items, the mnemonic traces of these items have been subjected to substantial inhibition from the recall of mixed-strong items. In pure lists, all items are equally (or, at least, roughly equally) associated to the retrieval cue. As a result, pure-weak items are subjected to less inhibition than mixed-weak items, but pure-strong items are subjected to more inhibition than mixed-strong items.

Bäuml (1997) used a novel variant of the mixed-pure paradigm to test this account. Subjects in his experiment studied lists of items from several distinct categories. There were six exemplars in each category, each of which started with a unique letter to allow for unambiguous word-stem completion. There were three category types: pure-weak categories, for which each exemplar was studied for 2 s; pure-strong categories, for which each exemplar was studied for 6 s; and mixed categories, for which half of the exemplars were studied for 2 s and the other half of the exemplars were studied for 6 s. Because memory was tested using category-exemplar-stem completion (e.g., "animal"-"g___" for "giraffe"), Bäuml was able to control output order. Importantly, for half of the mixed categories, strong exemplars were tested first; for the other half of the mixed categories, weak exemplars were tested first. When Bäuml ignored test position, he found a positive list-strength effect; crucially, however, when he restricted his analysis to first-half exemplars, he found a null list-strength effect. This pattern is consistent with the inhibition account: If overt retrieval is required for inhibition, then this should manifest

later in the test phase (as shown by the overall positive list-strength effect) but not early in the test phase (as shown by the first-half null list-strength effect).

Subsequent to Bäuml's (1997) experiment, Malmberg and Shiffrin (2005) investigated different strengthening techniques on the list-strength effect in free recall. They found that, although spaced repetitions produced the standard list-strength effect, massed repetitions, increased study time, and levels-of-processing manipulations did not. Verde (2009) noted that Malmberg and Shiffrin's results make interpretation of Bäuml's findings difficult, since Bäuml strengthened items using increased study time. Therefore, the null list-strength effect observed in the first half of his test may have been due to a suboptimal strengthening technique rather than a lack of inhibition during the early part of the test phase. Verde performed several conceptual replications of Bäuml's experiment using a spaced strengthening technique. Ignoring test position, Verde found positive list-strength effects, consistent with the one reported by Bäuml. Critically, however, when Verde considered recall of only the first tested item in each category, there was still a positive list-strength effect, contradicting Bäuml's first-half analysis. Verde's results cannot be explained by inhibition, since no overt retrieval took place prior to recall of the first item.

A final important difference between the list-strength effect and retrieval-induced forgetting casts doubt on the viability of the inhibition account. As described above, the list-strength effect does not occur in recognition (Ratcliff et al., 1990); however, retrieval-induced forgetting does (Hicks & Starns, 2004; Rupprecht & Bäuml, 2016). Inhibition theorists explain recognition-based retrieval-induced forgetting by appealing to cue independence. Recall that inhibition is not limited to the original cue; rather, it decreases

the probability of successful retrieval using any possible cue. For example, if subjects studied "vegetable"-"asparagus" during the study phase, "vegetable" is the cue and "asparagus" is the target. In a recognition test, "asparagus" becomes the cue, with subjects asked to make an old/new decision. Crucially, "asparagus" is a different cue from "vegetable"—that is, it is an independent cue. If cue independence is responsible for retrieval-induced forgetting in recognition, then the same rationale should lead to a list-strength effect in recognition. Since the list-strength effect is not found in recognition, it is difficult to see how inhibition theory can simultaneously account for retrieval-induced forgetting in recognition along with a null or negative list-strength effect.

Although I am skeptical of the inhibition interpretation of the list-strength effect (Anderson, 2003; Bäuml, 1997), I think the data leveraged in support of it make a persuasive case for the role of output interference. Output interference is the well-documented phenomenon wherein performance on a memory test declines as testing proceeds (Arbuckle, 1967; Aue, Criss, & Prince, 2015; J. Brown, 1954; Dalezman, 1976; Jahnke, 1969; Kang & Oh, 2016; Roediger, 1974; Roediger & Schmidt, 1980; Runquist & Horton, 1977; A. D. Smith, 1971, 1973; Smith, D'Agostino, & Reid, 1970; Tulving & Arbuckle, 1963, 1966). Particularly in free recall, one might suspect that output interference simply stems from subjects exhausting the set of items they successfully encoded at study. However, item recognition—in which, importantly, the experimenter controls output order rather than the subject—produces output interference, with discrimination of targets from distractors decreasing as test position increases (Annis, Malmberg, Criss, & Shiffrin, 2013; Criss, Malmberg, & Shiffrin, 2011; Koop, Criss, & Malmberg, 2015; Norman & Waugh, 1968; Osth, Jansson, Dennis, & Heathcote, 2018;

Peixotto, 1947; Ratcliff & Hockley, 1980; Ratcliff & Murdock, 1976; Schulman, 1974; Underwood & Freund, 1970).

Why does output interference occur? From an interference perspective, successfully retrieving an item increases its association to the retrieval cue. This increase in strength exerts interference on other, related items through occlusion. The fact that, on mixed-strength lists, strong items are output earlier than weak items (Wike & Wike, 1970; Wixted et al., 1997), may explain the list-strength effect. On pure lists, all output items were necessarily encoded at the same strength, so output interference "washes out". In contrast, on mixed lists, the early recall of mixed-strong items protects them from output interference, and the later recall of mixed-weak items subjects them to output interference. This is the pattern produced in Bäuml's (1997) experiment.

An important strength of the output-interference account of the list-strength effect is its parsimony. Recall that Verde (2009) found positive list-strength effects for both the entire test list and for just the first tested item. Although inhibition can explain the former finding, it cannot explain the latter. An inhibition account, therefore, would have to appeal to different mechanisms for the two list-strength effects. Conversely, interference appeals to the same mechanism—occlusion—for both patterns.

3.3. Summary of the List-Strength Effect

None of the accounts reviewed above provide a complete explanation of the list-strength effect. Hirshman's (1995) criterion-placement account is consistent with some data, but its universality is difficult to assess given that many experiments only analyze measures of discrimination, not hit and false-alarm rates. Although BCDMEM (Dennis & Humphreys, 2001) and the continuous-memory version of TODAM (Murdock & Kahana,

1993a, 1993b) can explain a null list-strength effect in recognition, they cannot explain negative list-strength effects (e.g., Ratcliff et al., 1990). Norman's (1999, 2002) insufficient-strengthening account is consistent with results from the strong-interference paradigm, but the account cannot explain why much weaker strengthening manipulations can produce positive list-strength effects in free recall (e.g., the 2:1 strength ratio used by Tulving & Hastie, 1972).[5] Similarly, although Norman's (1999, 2002) dual-process account is consistent with some empirical findings, it cannot explain why cued recall, which should rely on recollection and not familiarity, yields a null list-strength effect (Ratcliff et al., 1990; Wilson & Criss, 2017). Output interference and inhibition are consistent with Bäuml's (1997) finding that the list-strength effect is present when test position is ignored but absent when controlled, but inconsistent with Verde's (2009) experiments, which, while replicating Bäuml's full-list analysis, also found positive list-strength effects with test position controlled.

The differentiation account (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997; Shiffrin et al., 1990) may provide the most complete explanation of the list-strength effect. It is capable of explaining null and negative list-strength effects without appealing to criterion placement, and it is able to predict positive list-strength effects when the differentiation process is disrupted (Murnane & Shiffrin, 1991b; Sahakyan & Malmberg, 2018). Yet, one finding it cannot explain is the positive list-strength effect observed in the strong-interference paradigm (Diana & Reder, 2005; Norman, 1999, 2002; Norman et al., 2008).

5. I thank Amy Criss for this idea.

In the following chapter, I present a detailed description and analysis of the Retrieving Effectively from Memory model (Shiffrin & Steyvers, 1997). In particular, I use version 3 of this model (REM.3) which, to date, has received very little attention in the literature.

Chapter 4: The Simulation Test of REM.3

The discoveries of the mirror effect (Glanzer & Adams, 1985, 1990; Stretch & Wixted, 1998) and the null or negative list-strength effect (Ratcliff et al., 1990, 1992, 1994) led to the abandonment of many first-generation global-matching models (for reviews, see Clark & Gronlund, 1996; Shiffrin et al., 1990). One of the earliest second-generation global-matching models was REM, introduced by Shiffrin and Steyvers (1997). Since its inception, REM has successfully accounted for a number of human-memory phenomena, including the word-frequency effect (Malmberg, Holden, & Shiffrin, 2004; Malmberg & Murnane, 2002), the letter-frequency effect (Malmberg, Steyvers, Stephens, & Shiffrin, 2002), effects of midazolam on memory (Malmberg et al., 2004), associative recognition (Criss & Shiffrin, 2005), the list-length effect (Criss & Shiffrin, 2004a, 2004b; Lehman & Malmberg, 2013), the strength-based mirror effect (Criss, 2006), output interference (Criss et al., 2011), the list-strength effect (Malmberg & Shiffrin, 2005; Osth, Fox, et al., 2018), context-dependent memory (Klein et al., 2007), retrieval-induced forgetting (Verde, 2013), intentional forgetting (Lehman & Malmberg, 2011), and some implicit-memory phenomena (Schooler et al., 2001).

Shiffrin and Steyvers (1997) presented several versions of REM, but the one most often utilized is called REM.1. A hallmark of the REM framework is differentiation: Like its SAM predecessor (Shiffrin et al., 1990; see also Raaijmakers & Shiffrin, 1992), item strengthening in REM.1 always results in single-image storage. This assumption, however, renders REM.1 unable to account for two empirical phenomena: the spacing effect (i.e., better memory for spaced repetitions than massed repetitions; see Delaney et

al., 2010, for a review) and positive list-strength effects observed with the strong-interference paradigm (Norman, 1999, 2002).

The REM variant Shiffrin and Steyvers (1997) called REM.3 relaxes REM.1's strict differentiation assumption. In REM.3, item repetitions are only stored in the originally generated image if subjects recognize that the repeated item was previously studied. The purpose of the simulations in the present chapter is to determine whether REM.3 is a viable alternative to REM.1. In particular, I assess whether REM.3 can account for the spacing effect and results from the strong-interference paradigm. I begin with a review of REM's historical underpinnings, followed by a description of the basic model (i.e., REM.1). I then describe REM.3 and how it differs from REM.1. Finally, I test REM.3's viability by simulating phenomena for which REM.1 cannot account.

4.1. A Brief Review of REM's Antecedents

REM is a direct descendant of the SAM model; in turn, SAM is a descendant of the Atkinson-Shiffrin model (Atkinson & Shiffrin, 1968). As such, both SAM and REM share characteristics with the Atkinson-Shiffrin model. In particular, both SAM and REM divide memory into short- and long-term stores. When an item is studied, it first enters the short-term store, where it is rehearsed along with other recently presented items. Once the short-term store reaches capacity, a random item is transferred to the long-term store, from which it can later be retrieved. Memories in the long-term store are assumed to be permanent, with forgetting attributed to interference (e.g., occlusion, cue overload, context change, etc.).

Raaijmakers and Shiffrin's (1980, 1981) original instantiation of SAM was only applied to free and cued recall; Gillund and Shiffrin (1984) later extended the model to

recognition, and Mensink and Raaijmakers (1988, 1989) extended it to a number of classic interference phenomena. Gillund and Shiffrin demonstrated that the recognition version of SAM could account for a number of benchmark phenomena, including the list-length effect (Strong, 1912), the word-frequency effect (Gorman, 1961), and higher recognition performance observed with increased study time (Ratcliff & Murdock, 1976). However, SAM was unable to simultaneously account for the mirror effect (Glanzer & Adams, 1985, 1990; Stretch & Wixted, 1998) and a null or negative list-strength effect (Ratcliff et al., 1990, 1992, 1994). It is worth noting, however, that, although the recognition version of SAM has been abandoned, its free- and cued-recall versions still enjoy considerable success (Huber et al., 2015; Kahana, 1996; Karlsen & Snodgrass, 2004; Kimball, Smith, & Kahana, 2007; Raaijmakers & Phaf, 1999; Sirotin, Kimball, & Kahana, 2005; Tomlinson et al., 2009).

4.2. Description of REM

4.2.1. The geometric distribution

REM relies heavily on the geometric distribution; therefore, it is prudent to review some of its properties first. The geometric distribution is the probability distribution of the number of Bernoulli trials required for a single, specific event to occur. A Bernoulli trial is a trial for which two mutually exclusive outcomes are possible. By convention, the terms “success” and “failure” are used to denote the outcome for which the probability is being computed and the other outcome, respectively. For example, if one is interested in the number of coin flips required to land on heads, heads is referred to as a success and tails is referred to as a failure. Provided the probability of flipping heads on any given trial can be determined, then the geometric distribution reflects the probability

distribution of the number of trials to reach one success. As another example, if one is interested in the number of rolls of a six-sided die that are needed before a four is rolled, then a roll of four is referred to as a success and rolls of one, two, three, five, and six are referred to as failures. Note that, although rolling a six-sided die has six possible outcomes, a die roll is still a Bernoulli trial if we are only interested in two possible outcomes, and if these outcomes, combined, exhaust all possible outcomes (e.g., six vs. anything else, four or five vs. anything else, etc.). In other words, Bernoulli trials are true dichotomies.

Let p denote the probability of a success and t denote the number of trials. Because there are only two outcomes, the probability of a failure is $1-p$. The probability mass function can then be written as:

$$f(t) = (1-p)^{(t-1)}p$$

where t is a positive integer and $0 < p < 1$.

For the purposes of REM, the most important feature of the geometric distribution is that varying the p parameter changes the sampling distribution. If a success is a common occurrence, then randomly sampling from the geometric distribution will produce a lot of low numbers. Consider, for example, an event that has a 90% chance of occurring on any given trial. Then, the probability of a success on the first trial is .9, the probability of a success on the second trial is .09, the probability of success on the third trial is .009, and so on. These probabilities are additive, such that the probability of success on any of the first three trials is .999. A property of the geometric distribution is that p approaches 0 as t increases; however, the speed at which this occurs goes up as the probability of a success on any given trial increases. In the present case, then, randomly

sampling from the geometric distribution will usually produce mostly 1s with some 2s. Sampling a 3 will be rare, and sampling very large numbers will be extraordinarily unlikely.

What about the distribution when *p* is set to a small probability? Consider the situation in which the probability of success on a given trial is .1. Here, the probability of success on the first trial is .1, the probability of success on the second trial is .09, the probability of success on the third trial is .081, and so on. In this case, although the probability still approaches 0 as *t* increases, it does so at a much slower rate than when the probability of success was higher. Randomly sampling from the geometric distribution when p = .1, then, will produce a more variable result than when *p* = .9.

4.2.2. Stimuli in REM

In REM, stimuli are represented as vectors of lexical and semantic features. For simplicity, all stimulus vectors are equal in length—*w* is the parameter that denotes the number of features per stimulus. Each stimulus feature is a positive integer drawn from the geometric distribution with parameter *g*. In Bernoulli trial terms, g represents the probability of success on any given trial, so higher values of *g* result, on average, in smaller feature values. Although Shiffrin and Steyvers (1997) left the notion of features relatively vague, features can be thought of as representing the individual components of the studied stimuli that subjects could potentially notice. So, some features represent orthographic features, some represent contextual features, some represent semantic features, and so on. The distinctiveness of a given feature is indexed by its value, with larger feature values denoting more distinctive features. This is due to the fact that, when

sampling from the geometric distribution, lower values are more common than higher values.

In most REM simulations, w is set to 20. The value of g is chosen according to the normative frequency of the word stimuli being simulated—that is, $g_l < g_h$, where g_l is the value of g used for low-frequency words and g_h is the value of g used for high-frequency words. In Shiffrin and Steyvers' (1997) original simulations, g_l was set to .325 and g_h was set to .45. As shown in the previous section, sampling from the geometric distribution with a lower probability event produces more variable results—as such, vectors for low-frequency words will, on average, have more unique integers and more large integers when compared to vectors for high-frequency words. The idea underpinning this assumption is that low-frequency words are composed of more distinctive features than high-frequency words.

Prima facie, making $g_l < g_h$ as an a priori assumption may appear arbitrary. Indeed, without this assumption, REM cannot predict the word-frequency mirror effect. Yet, there is empirical support for the assumption. Zechmeister (1969) asked subjects to rate the orthographic and graphemic distinctiveness of low- and high-frequency words, and found that low-frequency words were rated as more distinctive than high-frequency words. Landauer and Streeter (1973) supplemented these subjective ratings with an objective analysis of low- and high-frequency words. They found that high-frequency words have more orthographic neighbours than low-frequency words (an orthographic neighbour is defined as the words that a given word can become by changing a single letter; e.g., orthographic neighbours of "cat" include "eat", "cut", and "car"). Additionally, the neighbours of high-frequency words had higher normative frequencies

than the neighbours of low-frequency words. Landauer and Streeter also showed that the phonemes and graphemes of low-frequency words are less common than those found in high-frequency words. Given these findings, there is empirical justification for REM's assumption that $g_l < g_h$.

4.2.3. Mnemonic traces in REM

Mnemonic traces in REM are called images. When an item is studied, an image corresponding to the studied item is generated. Images are rarely exact copies of studied stimuli; rather, images usually contain a combination of correctly copied features, incorrectly copied features, and missing features. Like stimuli, images are vectors of length w. Positions for which features were copied (correctly or incorrectly) contain positive integers; positions with missing information (i.e., positions for which no information was copied) are set to 0.

Shiffrin and Steyvers (1997) presented three encoding parameters: t, u, and c. The t parameter represents the number of time units for which each stimulus is studied. The u parameter represents the probability of copying a feature on any given unit of study time. Finally, the c parameter represents the probability that a feature, if copied to the image, will be copied correctly. In Shiffrin and Steyvers' simulations, u was set to .04, c was set to .7, and t varied according to the paradigm being simulated. For example, in simulating the mixed-pure paradigm, they set t_1 to 7 and t_2 to 10, with t_1 representing the number of time units for which weak items were studied and t_2 representing the number of time units for which strong items were studied.

Although Shiffrin and Steyvers (1997) treated t and u as separate parameters, subsequent papers noted that they are actually a single parameter, representing the

probability of storing a value for a given feature when an item is studied. Notice that the probability of storing information for a feature across the entire study trial is *tu*. For a more concrete example, Shiffrin and Steyvers' values of 7, 10, and .04 for t_1, t_2, and u, respectively, can be condensed into u_{weak} and u_{strong}: $u_{weak} = t_1 \times u = .28$ and $u_{strong} = t_2 \times u = .4$. Therefore, for the remainder of the present paper, I will use u to refer to *tu* from Shiffrin and Steyvers' original presentation.

When a feature from a studied stimulus is copied to its corresponding image, it is copied correctly with probability c and incorrectly with probability $1-c$. When copied incorrectly, a new value is drawn from the geometric distribution. However, REM assumes that subjects are unaware of experimental manipulations like word frequency. Therefore, incorrect feature values are drawn with the value of g assumed to reflect subjects' beliefs about environmental base rates of features. Shiffrin and Steyvers set this to .4.[6] For ease of exposition, I differentiate between two g parameters: g_{draw}, which refers to the value of g used in filling stimulus vectors, and g_{base}, which refers to the value of g used when incorrectly copying a feature.[7]

---

6. It is not entirely clear why Shiffrin and Steyvers (1997) chose .4 as the value of g known to subjects. However, recognition discrimination increases as g decreases, so .4 was likely chosen because it produced reasonable performance levels.

7. My division of g into g_{draw} and g_{base} should not be interpreted to mean that these are two independent, free parameters. Values of g_{draw}—inasmuch as they are taken to be indices of normative feature frequency—necessarily depend on the value of g_{base}. Most REM implementations set g_{base} to .4, and g_l and g_h to values such as .325 and .45,

To summarize, images in REM are vectors of length w. Values in the image are 0s or positive integers. On average, wu of the image values will be positive integers and $w(1-u)$ of the image values will be 0s. Of the non-zero values in the image, $100c$% of them will be the values contained in the studied stimulus, and $100(1-c)$% of them will be values drawn from the geometric distribution with g_{base}. Notice that this means that, by coincidence, some of the incorrectly stored values will be identical to the feature from the studied stimulus, and the probability of this increases as g_{draw} increases (i.e., as word frequency increases) and as the specific feature value decreases (e.g., if a feature value of 1 is incorrectly copied, the probability of the correct value being stored by sampling from the geometric distribution is equal to g_{base}).

4.2.4. Evaluating recognition probes

REM is a global-matching model, so recognition decisions are made by matching probes to all of the images in memory. The purpose of the matching process is to determine whether or not an image in memory was generated by studying the probe. This matching process involves comparing each feature in the probe to the corresponding feature in each image, skipping positions for which the image contains no information (i.e., positions for which the image vector is 0). The matching process takes feature diagnosticity into account. Recall that, regardless of the value of g_{base}, lower values are

---

respectively. If g_{base} were changed, for example, to .6, g_l and g_h would likewise need to increase; otherwise, it would be incoherent to argue that g_l reflects words composed of environmentally uncommon features and that g_h reflects words composed of environmentally common features.

more commonly sampled from the geometric distribution than higher values—that is, 1 is always more common than 2, which is always more common than 3, and so on. Consequently, the probability of a probe and an image sharing a feature by chance decreases as the stored feature value increases.

In REM, recognition decisions are made according to the ratio between the probability that an image in memory was generated by the probe and the probability that an image in memory was generated by a different stimulus (denoted φ). As mentioned above, this process takes environmental base rates into account, such that rarer feature matches increase the probability to a greater extent than do common feature matches. φ is computed according to the following equation

$$\phi = \frac{1}{n}\sum_{j=1}^{\infty}\lambda_j$$

where n is the number of images in memory. λ can be computed as follows

$$\lambda_j = (1-c)^{n_{jq}}\prod_{1}^{\infty}\left[\frac{c+(1-c)g_{base}(1-g_{base})^{i-1}}{g_{base}(1-g_{base})^{i-1}}\right]^{n_{ijm}}$$

where n_{jq} denotes the number of nonzero values in the jth image n_{ijm} denotes the number of nonzero values in the jth image that match the probe and have value i.

Recognition tests introduce a final REM parameter: *criterion*. The *criterion* parameter represents the odds ratio required for a probe to be called "old" (i.e., an "old" response is made if $\varphi >$ *criterion*, and a "new" response is made if $\varphi \leq$ *criterion*). Shiffrin and Steyvers (1997) set *criterion* to 1, which results in a probe being called "old" if the probability that the probe generated an image in memory exceeds the probability that the probe did not generate an image in memory. Note that an odds ratio of 1 reflects

equivalent evidence that the probe is a target and that the probe is a distractor. Shiffrin and Steyvers argued that, in the absence of experimental pressure to use a more liberal or conservative criterion, a value of 1 for the *criterion* parameter is the most reasonable.

4.2.5. Basic simulations

Shiffrin and Steyvers (1997) showed that REM.1—that is, the version of REM described above—is capable of predicting a number of benchmark recognition findings. In the following four subsections, I replicate REM simulations of the word-frequency mirror effect, the strength-based mirror effect, the negative list-strength effect, and the list-length effect. Combined, these simulations illustrate three properties of the REM framework: feature diagnosticity, differentiation, and interitem interference. Later in the paper, I run simulations using REM.3 instead of REM.1. However, my purpose in the present section is twofold: (1) to replicate simulation results from the literature, thereby establishing that I have correctly implemented the REM framework, and (2) to demonstrate the utility of the simplest version of REM.

Before proceeding to the simulations, a note on the analysis of simulated data is warranted. Historically, researchers have not performed inferential statistical tests on simulated data. Although I know of no study where the rationale for this is explicitly articulated, this is likely because, owing to the large number of simulated subjects, any observed difference is bound to be statistically significant. If it is not, a researcher could simply increase the number of simulations until the desired *p* value is achieved. I believe this reasoning is sound and, therefore, remain consistent with previous studies: Although I present descriptive statistics in figures and give a verbal description of the findings, I do not perform inferential tests.

4.2.5.1. The word-frequency mirror effect. As briefly described in the previous chapter, the word-frequency mirror effect occurs when low- and high-frequency words are included on a recognition test. The standard finding is that the hit rate is higher and the false-alarm rate is lower for low-frequency words compared to high-frequency words (Glanzer & Adams, 1985, 1990).

Ten thousand simulated subjects completed a recognition test consisting of 16 targets and 16 distractors, evenly divided between low- and high-frequency words. Consistent with Shiffrin and Steyvers (1997), I set g_l to .325 and g_h to .45. The remaining parameters were set using values common in the REM literature: $w = 20$, $u = .28$, $c = .7$, $g_{base} = .4$, and $criterion = 1$.

The simulation results are shown in Figure 4.1. There was a word-frequency mirror effect: The hit rate was higher for low-frequency words ($M = .872$, $SE = .001$) than high-frequency words ($M = .863$, $SE = .001$), but the false-alarm rate was higher for high-frequency words ($M = .352$, $SE = .002$) than low-frequency words ($M = .259$, $SE = .002$). Naturally, this resulted in a discrimination advantage—as indexed by d'—for low-frequency words ($M = 2.253$, $SE = 0.011$) over high-frequency words ($M = 1.869$, $SE = 0.010$).[8] Note that, although the hit-rate portion of the mirror effect is quite small, it is consistent across simulation runs.

REM explains the word-frequency mirror effect in terms of feature diagnosticity or feature distinctiveness. As described above, there is empirical support for the claim

8. In the present paper, d' was computed by changing hit and false-alarm rates of 1 and 0 to .995 and .005, respectively.

that low-frequency words are composed of more distinctive features than high-frequency words (Landauer & Streeter, 1973) and that this difference is noticeable to subjects (Zechmeister, 1969). REM simulates this feature difference by using less common features for low-frequency words than high-frequency words.

In the present simulations, the encoding parameters—*u* and *c*—were equivalent for low- and high-frequency words. Consequently, images from low- and high-frequency words, on average, consisted of the same number of copied features. Indeed, because of the more common features contained in the high-frequency vectors, images are more likely to have correct feature values for high- than low-frequency words. This is because, when correct copying fails, which occurs with probability $1-c$, features from high-frequency words are more likely to be sampled from the geometric distribution with parameter g_{base} than features from low-frequency words. However, because features from low-frequency words are more unique, a match between a probe and an image from a low-frequency word provides more evidence for an "old" response than a match between a probe and an image from a high-frequency word. In effect, fewer feature matches between a probe and image were required for the odds ratio to exceed *criterion* for a low-frequency word than a high-frequency word—put another way, a low-frequency word could be recognized with fewer features encoded compared to a high-frequency word. Therefore, feature diagnosticity explains the hit-rate portion of the word-frequency mirror effect.

The false-alarm portion of the mirror effect is also due to feature diagnosticity. On average, high-frequency distractors will have more features in common with high-frequency targets when compared to features shared between low-frequency distractors

and targets. Therefore, erroneous matches between high-frequency distractors and an image are more common than erroneous matches between low-frequency distractors and an image in memory, thereby producing a higher false-alarm rate.

4.2.5.2. The strength-based mirror effect. The strength-based mirror effect is the finding that the hit rate is higher and the false-alarm rate is lower following study of a strong list than study of a weak list (Stretch & Wixted, 1998). Ten thousand simulated subjects completed two study-test cycles, with stimuli presented for a longer study duration in one list than the other. Each recognition test had 25 targets and 25 distractors. I set u_{weak} to .28 and u_{strong} to .4, consistent with the values used by Shiffrin and Steyvers (1997). The remaining parameters were set to values common in the REM literature: w = 20, c = .7, g_{base} = .4, g_{draw} = .35, and *criterion* = 1.

Simulation results are shown in Figure 4.2. The results produced a strength-based mirror effect: The hit rate was higher following study of the strong list (M = .899, SE = .001) than following study of the weak list (M = .852, SE = .001), but the false-alarm rate was higher following study of the weak list (M = .300, SE = .001) than following study of the strong list (M = .175, SE = .001). As one would expect, d' was higher for the strong list (M = 2.366, SE = 0.006) than the weak list (M = 1.654, SE = 0.005).

Recall that the strength-based mirror effect has been attributed to differentiation (Criss, 2006) or to between-list criterion shifts (Stretch & Wixted, 1998). Critically, REM provides an account of the strength-based mirror effect that does not require a criterion shift. The *criterion* parameter was equal between list conditions; yet, despite the distractors being equivalent between lists, the false-alarm rate was higher on the weak list than the strong list. REM explains this pattern by appealing to differentiation: With

additional study time, more features are encoded for strong items than weak items, thus increasing their dissimilarity to other images. Obviously, the greater encoding of strong items compared to weak items explains the hit-rate portion of the strength-based mirror effect. Importantly, better encoding also reduces the similarity of the resulting images to distractors. The false-alarm portion of the mirror effect, then, stems from fewer erroneous matches between distractor features and features encoded for strong items.

4.2.5.3. The list-strength effect. Ratcliff et al. (1990) showed that the list-strength effect is slightly negative in recognition. I simulated the mixed-pure paradigm using 10000 simulated subjects. This represented a 2 (strength: weak vs. strong) × 2 (list type: pure vs. mixed) repeated-measures design. On the pure lists, 24 targets were encoded at the same rate; on the mixed list, half of the items were strengthened. Consistent with Shiffrin and Steyvers (1997), I set u_{weak} to .28 and u_{strong} to .4. The remaining parameters were set to values common in the REM literature: $w = 20$, $g_{\text{draw}} = .35$, $c = .7$, $g_{base} = .4$, and *criterion* = 1.

As expected, the simulations yielded a negative list-strength effect (see Figure 4.3): Pure-strong items ($M = 2.407$, $SE = 0.006$) were slightly better recognized than mixed-strong items ($M = 2.405$, $SE = 0.008$), but mixed-weak items ($M = 1.868$, $SE = 0.007$) were better recognized than pure-weak items ($M = 1.665$, $SE = 0.005$). The strong-to-weak ratios in the mixed and pure lists were 1.288 and 1.445, respectively, yielding a negative list-strength effect ($R_r = 0.891$).

Like the differentiation version of SAM (Shiffrin et al., 1990), strengthened items in REM are always accumulated in a single image. As an item becomes better learned (i.e., as the number of features correctly copied to the corresponding image increases), it

is less likely to erroneously match a distractor. This differentiation is responsible for REM's prediction of a negative list-strength effect in recognition.

Somewhat counterintuitively, stronger images exert less interitem interference than weaker images. Therefore, overall interference is highest for the pure-weak list, lowest for the pure-strong list, and intermediate for the mixed list. This leads to the negative list-strength effect. Discrimination is better for mixed-weak items than pure-weak items because half of the competitors on the mixed list are strong, and therefore produce less interference than the competitors on the pure-weak list. In contrast, discrimination is better for pure-strong items than mixed-strong items because the mixed-weak competitors produce more interference than the competitors on the pure-strong list. Notably, the effect is much smaller for strong items since, due to differentiation, strong images are less susceptible to interference, even from weak images.

4.2.5.4. The list-length effect. Despite recent controversy (Annis et al., 2015; Criss & Shiffrin, 2004a; Dennis & Humphreys, 2001; Dennis et al., 2008), the list-length effect was considered a well-established phenomenon when Shiffrin and Steyvers (1997) first presented REM (e.g., Strong, 1912). Moreover, the reason that REM predicts a list-length effect is informative. I therefore performed a simulation of the list-length effect. Forty thousand simulated subjects completed a single study-test cycle consisting of 10, 20, 30, or 40 targets and distractors. Because the manipulation was run between simulated subjects, there were 10000 simulated subjects per cell, rather than 10000 total simulated subjects. Because memory is cleared between study-test cycles in REM.1, the results would have been the same with a within-subjects manipulation of list length. Parameters

were set to values common in the REM literature: $w = 20$, $g_{draw} = .35$, $u = .28$, $c = .7$, $g_{base} = .4$, and *criterion* = 1.

Simulation results are shown in Figure 4.4. Inspection of the graph shows that d' decreased as list length increased, confirming that REM predicts a list-length effect in recognition.

The reason REM predicts a list-length effect lies in how the odds ratio is calculated. The probe is matched sequentially to each image in memory, and subjects' old/new decision reflects this. If there is only a single image in memory, then a match of a common feature value (e.g., 1) between the probe and image is fairly diagnostic. However, if the probe is matched to 40 separate images, then one would expect some erroneous feature matches, particularly for common feature values. The amount of evidence required for an "old" response therefore increases as list length increases, reducing the hit rate. At the same time, the possibility of an erroneous match between a distractor and image increases as list length increases, resulting in a higher false-alarm rate for long lists than short lists.

4.3. REM.3

Almost all implementations of REM make an important, simplifying assumption: namely, that strengthened items are always stored in a single image. The alternative approach, wherein strengthened items are stored in multiple images, renders list-strength manipulations de facto list-length manipulations, as in early instantiations of first-generation global-matching models (Gillund & Shiffrin, 1984; Shiffrin et al., 1990), and thus causes the model to predict a list-strength effect in recognition. However, it would be circular to assume that item strengthening accumulates in a single image simply because

this allows the model to fit experimental data. Instead, it is critical to consider the strengthening techniques experimenters have used, and determine whether single-image storage is supported for these techniques.

Experimenters have used a variety of strengthening techniques, including increased study time, massed repetitions (i.e., repeating strengthened items without any intervening items), spaced strengthening (i.e., only repeating strong items following one or more intervening items), and levels-of-processing manipulations (for a comparison in free recall, see Malmberg & Shiffrin, 2005). Increasing study time or inducing elaborative processing of strengthened items constrains each strong item to a separate episode, and thus single-image storage is reasonable. Otherwise, one would have to assume that the weak items in these experiments could also be stored in separate images. For example, consider two list-strength experiments, one of which presents weak and strong items for 1 and 2 s, respectively, and one of which presents weak and strong items for 2 and 4 s, respectively. To argue that the strong items in the first experiment are stored in separate images, one would also have to argue that the weak items in the second experiment are stored in separate images.

What about strengthening via item repetitions? In the case of massed strengthening, single-image storage seems reasonable. Massed repetitions are very similar to increased study time, save that an interstimulus interval is interpolated during item presentation. In REM, items are rehearsed in the short-term store before being transferred to the long-term store. As such, the immediate re-presentation of a strong item should occur while the image is still in the short-term store, therefore resulting in a single study episode.

Spaced-strengthening techniques present a challenge to the assumption of single-image storage. When items are strengthened through spaced repetitions, at least one other item intervenes between presentations. For example, Malmberg and Shiffrin (2005, Experiment 2) used nine-item lags in their list-strength experiments, such that nine items were interpolated between the first and second presentations of a strong item. Here, an assumption of automatic, single-image storage seems misplaced. Imagine, for example, that the first presentation of the strengthened item was poorly encoded, or that a large number of items appeared between presentations. In these cases, it is possible that the re-presentation of the strengthened item will not be recognized by subjects as previously studied and a second image could plausibly be created. Of course, if the original presentation was well-encoded, subjects would recognize the re-presentation as previously studied, and thus update the original image.

Shiffrin and Steyvers (1997) recognized the potential problems that arise for single-image storage with spaced repetitions. They therefore described a version of REM—REM.3—that took spaced repetitions into account. REM.3 incorporates a mechanism by which spaced repetitions can be encoded in the originally generated image—what Shiffrin and Steyvers termed superimposition of similar images. Crucially, in REM.3, a repetition must be recognized as previously studied for it to be superimposed on the original image. In REM.3, each target during the study phase is treated like a test probe. The simulated subject matches the target to all images in memory and, if the odds that the target was previously studied exceeds a given criterion, the most-similar image to the target is updated. If the criterion is not exceeded, a new image is generated. REM.3 thus has two criteria parameters, which I will term $criterion_{study}$ and $criterion_{test}$. The

former represents the odds required to identify a study item as previously studied; the latter represents the threshold required to call a test probe "old".

Shiffrin and Steyvers (1997) noted that a number of complexities arise with REM.3, not least of which is that the time needed to run the simulations increases considerably. Given that the qualitative predictions of REM.1 and REM.3 were the same—at least for the phenomena Shiffrin and Steyvers examined—they did not pursue REM.3 further. To the best of my knowledge, no subsequent REM papers have implemented REM.3.

We know that at least one phenomenon, the spacing effect, cannot be accommodated by REM.1. The spacing effect refers to the finding that spaced repetitions have a mnemonic advantage over massed repetitions (for reviews, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Delaney et al., 2010; Dempster, 1988; Dempster & Farris, 1990; Donovan & Radosevich, 1999; Janiszewski, Noel, & Sawyer, 2003; Noel & Vallen, 2009; Ruch, 1928; Son & Simon, 2012). The spacing effect is ubiquitous in the memory literature, occurring in free recall (Delaney & Knowles, 2005; Delaney, Spirgel, & Toppino, 2012; Delaney & Verkoeijen, 2009; Foos & Smith, 1974; Glanzer, 1969; Glenberg, 1977; Kahana & Howard, 2005; Madigan, 1969; Melton, 1967, 1970; Shaughnessy, Zimmerman, & Underwood, 1972; Siegel & Kahana, 2014; Underwood, 1969, 1970; Zimmerman, 1975), cued recall (Challis, 1993; D'Agostino & DeRemer, 1972; Delaney, Godbole, Holden, & Chang, 2018; Morehead, Dunlosky, Rawson, Bishop, & Pyc, 2018), recognition (Allen & Garton, 1970; Braun & Rubin, 1998; Kahana & Greene, 1993; Maddox, Pyc, Kauffman, Gatewood, & Schonhoff, 2018; Ross & Landauer, 1978; Russo, Parkin, Taylor, & Wilks, 1998; Strong, 1916; Xue et al., 2011),

frequency judgments (Greene & Stillwell, 1995; Hintzman, 1969; Hintzman & Rogers, 1973; Hintzman, Summers, Eki, & Moore, 1975; Rose & Rowe, 1976), implicit-memory tasks (Greene, 1990), second-language retention (Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bercovitz, Bell, Simone, & Wiseheart, 2017), and workplace training (Kim, Wong-Kee-You, Wiseheart, & Rosenbaum, 2019). Indeed, the spacing effect has even been documented in nonhuman populations, including mice (Aziz et al., 2014), chicks (C. P. Brown, 1976), honeybees (Menzel, Manz, Menzel, & Greggers, 2001), and sea slugs (Carew, Pinsker, & Kandel, 1972). Critically, if REM treats massed and spaced repetitions identically, then they will necessarily be recognized at the same rate, and therefore the model cannot account for the spacing effect.

In Chapter 3, I described Norman's (1999, 2002) results from the strong-interference paradigm. Unlike the mixed-pure paradigm, the strong-interference paradigm produces a positive list-strength effect in recognition. According to the insufficient-strengthening account, this is because, unlike the mixed-pure paradigm, the strong-interference paradigm strengthens strong items to ceiling. From a REM perspective, because this represents significant strong-item differentiation, performance should be better on the strong-interference list than the weak-interference list, contrary to empirical results (Diana & Reder, 2005; Norman, 1999, 2002; Norman et al., 2008; Osth et al., 2014). Interestingly, all published experiments using the strong-interference paradigm have strengthened strong items through widely spaced repetitions. It is conceivable, therefore, that the predictions derived from REM.1 may not extend to REM.3.

4.4. REM.3 Simulations

The purpose of the simulations that follow was to determine whether REM.3 is a viable alternative to REM.1. REM.1 cannot explain the spacing effect or the results from the strong-interference paradigm. Here, I assess whether REM.3 can predict these phenomena.

4.4.1. Simulation 1: Setting $criterion_{study}$

REM assumes that, when making old/new decisions on a recognition test, subjects optimize performance by making an "old" decision whenever the probability that the probe is a target exceeds the probability that the probe is a distractor. In REM.3, each item on the study list is subjected to the same evaluative process. However, is a criterion of 1 still reasonable in this case? Two factors suggest that it may not be.

First, on a recognition test, subjects are asked to make a binary decision: Did this item appear on the study list? Here, even a small amount of evidence one way or the other is sufficient to tip the proverbial scales. During a study list, in contrast, subjects are not overtly assessing whether each item was previously studied. Instead, motivated subjects are actively trying to commit each item to memory and, presumably, recognizing that an item was studied earlier only occurs when the corresponding image is particularly strong. In REM.3, then, if the first presentation of an item was poorly encoded, it does not make sense for the study item to be superimposed on the original image.

There is a second reason that a stricter value for $criterion_{study}$ is desirable. In REM.3, when a study item is identified as previously studied, it is superimposed on the most similar image. Superimposition can go wrong in two ways: First, the study item may not have actually appeared on the study list. This is akin to a false alarm on a recognition test. In this case, the image on which the study item is superimposed was generated by a

different item. This results in an image formed from two (or more) targets. Second, even if the item was previously studied, it may, by chance, match a different image to a greater extent than the image from the item that was actually studied earlier. In both cases, the item is superimposed on the wrong image, leading to images containing features from multiple items.

To test the degree to which superimposition is a problem in REM.3, I simulated study phases on which no items were repeated. The dependent variable was the proportion of targets that were superimposed on other images. Because no targets were repeated, all instances of superimposition were errors.

As reviewed earlier, in REM, the false-alarm rate increases as g_{draw} increases. A reasonable prediction for REM.3, then, is that the proportion of superimposition errors will increase as g_{draw} increases. I therefore varied g_{draw} in the present simulation.

4.4.1.1. Method. I varied three parameters: g_{draw} (.3, .35, .4, .45, .5), list length (2, 4, 8, 16, 32, 64, 128, 256), and $criterion_{study}$ (1, 2, 3). I ran 1000 simulations for each combination of these parameters, resulting in 120,000 simulated subjects.

Parameters not central to the present simulation were set to values common in the REM literature: $g_{base} = .4$, $w = 20$, and $c = .7$. In REM.3, two separate u parameters are required, which I will call u_1 and u_2. u_1 denotes the probability of storing features from a study item when a new image is generated (i.e., when superimposition does not occur). u_2 denotes the probability of storing features for a study item when it is superimposed on a previously generated image. Note that, during superimposition, only item features in positions where the image contains no information (i.e., 0s) can be encoded. In REM.3, it is standard for u_2 to be less than u_1, in the same way that u_{strong} is generally less than 2 ×

u_{weak}. This is because subjects are assumed to devote fewer resources to information that has already been well learned. In the present simulations, I set u_1 to .28 and u_2 to .12. This matches Shiffrin and Steyvers' (1997) values of .28 and .4 for u_{weak} and u_{strong}, respectively. This means that, when a study item is identified as previously studied, item features from vector positions for which the image has a value of 0 are copied with probability .12.

Two algorithmic decisions bear mentioning. First, when an item was identified as previously studied, it was superimposed on the most similar image in memory. If a tie occurred between two or more images, the item was superimposed on the more recently generated image. Second, none of our simulated study lists contained any repeated items, even by chance. As each study item was generated, it was checked against the current list of study items and, if it matched a previously generated item, it was replaced. These algorithmic precautions were used in this and all subsequent simulations.

4.4.1.2. Results and discussion. The simulation results are presented in Figure 4.5. As is clear in the figure, the probability of superimposition errors increased as word frequency (i.e., g_{draw}) increased. This is consistent with the higher false-alarm rate observed for high-frequency words than low-frequency words (Glanzer & Adams, 1985, 1990). It is also clear that the probability of superimposition errors increases with list length. This is intuitively reasonable: As the number of images in memory increases, the probability of a study item mistakenly matching a previously studied item increases. This is akin to the higher false-alarm rates observed following study of a long list compared to a short list (e.g., Gillund & Shiffrin, 1984). Of greatest importance for present purposes is

that, for all g_{draw} values and list lengths, the probability of superimposition errors decreases as $criterion_{study}$ increases.

None of the results from Simulation 1 are particularly surprising; indeed, they are necessary outcomes of the math underlying the model. However, what they demonstrate is that a value of 1 for $criterion_{study}$ is far too error-prone to be useful. For example, collapsing across g_{draw}, the probability of superimposition errors with $criterion_{study} = 1$ was .138 for 32-item lists and .164 for 64-item lists. By increasing $criterion_{study}$ to 2, mean error rates decrease to .064 and .070, respectively, and by increasing $criterion_{study}$ to 3, mean error rates decrease to .039 and .041, respectively. These error rates seem more reasonable, given that subjects are usually not asked to make old/new decisions during the study phase.

It is not possible to collect behavioural data on the probability of superimposition errors during a study phase. Asking subjects to report whether each item on a study list was studied earlier renders the study phase a de facto continuous-recognition test; subjects in such an experiment would presumably use the criterion they would use during an old/new recognition test. Therefore, Simulation 1 is necessarily speculative with regard to subjects' actual criterion placement at study. However, I find it difficult to believe that subjects are incorrectly identifying 10%-20% of novel study items as previously studied, and thus I believe that a higher value for $criterion_{study}$ is more realistic.

Interestingly, it may be possible to test subjects' criterion placement using neuroscientific methods. Recognition that the re-presentation of an item was previously studied—or incorrectly believing an item was previously studied—would presumably be accompanied by a feeling of knowing. Feelings of knowing produce distinct brain activity

measurable with electroencephalography (Paynter, Reder, & Kieffaber, 2009) and functional magnetic resonance imaging (Chua, Schacter, & Sperling, 2009). Physiological measures are also possible, as feelings of knowing are accompanied by heart-rate increases (Fiacconi, Kouptsova, & Köhler, 2017) and shorter viewing time as measured by eye tracking (Chua & Solinger, 2015). Therefore, neuroscientific and/or physiological tests of study-phase criterion placement may be a fruitful avenue for research.

In the remainder of my REM.3 simulations, I systematically varied $criterion_{study}$. By and large, changes in $criterion_{study}$ rarely affect the qualitative pattern of the results. However, in the situations in which differences in predictions across values of $criterion_{study}$ are observed, higher values of $criterion_{study}$ are more in line with empirical data than lower values.

4.4.2. Simulation 2: The spacing effect

In REM.1, the simplifying assumption that spaced repetitions are always accumulated in a single image renders it unable to predict the spacing effect. Given the robustness of the spacing effect (Delaney et al., 2010), REM.3 must be able to predict an advantage for spaced over massed repetitions to remain viable. The purpose of Simulation 2 was to assess whether REM.3 results in better memory for spaced (i.e., probabilistically superimposed) or massed (i.e., automatically superimposed) images.

4.4.2.1. Method. I simulated a 2 (strengthening method: massed vs. spaced) × 3 (list length: 16, 32, 64) design. Each study list consisted of an equal number of massed and spaced stimuli, all of which were presented twice.

I used the following fixed parameters: $g_{draw} = .35$, $g_{base} = .4$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$. I varied $criterion_{study}$ as in Simulation 1 (1, 2, 3).

Massed repetitions were always stored in a single image, with the assumption that the immediate re-presentation of an item would be obvious to subjects. Features from massed items were thus copied with probability u_1+u_2. This is identical to how strong items are treated in REM.1. When subjects recognized the spaced repetition as previously studied, it was superimposed on the most similar image in memory, with features copied with probability u_2. When a spaced repetition was not recognized as previously studied, its features were copied to a new image with probability u_1. Notice that superimposed spaced items were functionally identical to massed items, with features from both copied to an image with probability u_1+u_2. Spaced and massed items only differed when the second presentation was stored in a new image rather than in the originally generated image, or when a superimposition error occurred.

4.4.2.2. Results and discussion. The simulation results are displayed in Figure 4.6. It is evident that REM.3 correctly predicts a discrimination advantage for spaced over massed items, regardless of $criterion_{study}$.

The results of Simulation 2 confirm that REM.3 can account for the spacing effect. In the remaining simulations, I turn to two phenomena from experiments on the list-strength effect: REM.3 needs to predict a null or negative list-strength effect with the mixed-pure paradigm, but a positive list-strength effect with the strong-interference paradigm. Although REM.1 predicts the former result, it cannot predict the latter.

4.4.3. Simulation 3: The mixed-pure paradigm

As discussed throughout the present paper, the mixed-pure paradigm generally produces a null or negative list-strength effect in recognition (Ratcliff et al., 1990). As a second test of REM.3's viability, I investigated whether it can predict this result.

4.4.3.1. Method. Each simulated subject studied five lists: a pure-weak list, on which all items were presented once; a massed pure-strong list, on which all items were studied multiple times, with strengthening accomplished through immediate repetitions; a spaced pure-strong list, on which all items were presented multiple times, with strengthening accomplished through spaced repetitions; a massed mixed list, on which half of the items were presented once and half of the items were presented multiple times in a massed fashion; and a spaced mixed list, on which half of the items were presented once and half of the items were presented multiple times in a spaced fashion. In the spaced pure-strong list and the spaced mixed list, all items were presented once before any items were presented a second time, all items were presented twice before any items were presented a third time, and so on. The dependent variable of interest was d', although I also computed R_r on the mean d' scores to remain consistent with previous reports (e.g., Ratcliff et al., 1990).

The following variables were systematically manipulated across simulations: the number of strong-item presentations (2, 3, 4, 5, 6), $criterion_{study}$ (1, 2, 3), and list length (28 vs. 84 unique targets per list). The full design can thus be conceived of as a 2 (item strength: weak vs. strong) × 2 (list type: pure vs. mixed) × 2 (strengthening technique: massed vs. spaced) × 2 (list length: 28 vs. 84) × 5 (strong-item presentations: 2, 3, 4, 5, 6) × 3 ($criterion_{study}$: 1, 2, 3) mixed design, with item strength, list type, and strengthening technique manipulated within simulated subjects and the remaining factors manipulated between simulated subjects. I ran 1000 simulations per cell, yielding 30,000 simulated subjects.

I used the following fixed parameters: $g_{base} = .4$, $g_{draw} = .35$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$. Algorithmically, strong items were treated as in Simulation 2.

4.4.3.2. Results and discussion. Figure 4.7 shows *d'* results for the pure-weak (top row), pure-strong (second row), mixed-weak (third row), and mixed-strong (bottom row) conditions, and Figure 4.8 shows R_r results. As expected, massed strengthening produces a negative list-strength effect, with a mean R_r across simulations of 0.937. Surprisingly, spaced strengthening produced a positive list-strength effect, with a mean R_r across simulations of 1.123. The R_r was slightly larger with an 84-item list ($R_r = 1.133$) than a 28-item list ($R_r = 1.113$) and increased as $criterion_{study}$ increased ($R_r = 1.083$, 1.136, and 1.150 for $criterion_{study}$ values of 1, 2, and 3, respectively).

Do these results falsify REM.3? The prevalent view in the literature is that the list-strength effect is negative or null in recognition. Yet, careful examination of the R_r values reported with massed and spaced strengthening suggests that the list-strength effect tends to be null or negative with massed strengthening and null or slightly positive with spaced strengthening. For example, Ratcliff et al. (1990, Experiment 5) manipulated strengthening technique (massed vs. spaced) within subjects and between lists using the mixed-pure paradigm. They found a negative list-strength effect with massed strengthening that did not quite reach their adopted significance level ($R_r = 0.89$, $p = .055$, two tailed). In contrast, the R_r was slightly positive in the spaced-strengthening condition ($R_r = 1.03$, NS). As another example, Ratcliff et al. (1992) used massed strengthening in Experiment 1 and obtained an R_r of 1.04; in Experiment 2, they used spaced

strengthening and obtained an R_r of 1.21, broadly consistent with the predictions of REM.3.

Table 4.1 shows the R_r values in published reports of the list-strength effect in recognition. I excluded experiments that used very fast presentation rates (Ratcliff et al., 1994; Yonelinas et al., 1992), as these are known to produce positive list-strength effects because of rehearsal borrowing (i.e., weak items are presented too quickly to process, so, on mixed lists, subjects continue rehearsing strong items during the presentation of weak items; for a discussion, see Yonelinas et al., 1992). I also excluded studies that used stimuli other than words or word pairs (e.g., Malmberg & Shiffrin, 1991a, 1991b). Such stimuli have not been modelled in REM, and there is evidence that they are more susceptible to interitem interference than words (Kinnell & Dennis, 2012; Osth et al., 2014). Finally, I only included experiments using the full mixed-pure paradigm—that is, I excluded experiments that omitted the pure-strong list or the pure-weak list (e.g., Hirshman, 1995). Across these experiments, the mean R_r for massed strengthening is 0.916, but the mean R_r for spaced strengthening is 1.062.

It is critical to note that REM.1 cannot account for even a slightly positive list-strength effect in recognition; therefore, REM.1 cannot account for the difference documented in Table 4.1. In contrast, Simulation 3 shows that REM.3 can simultaneously account for a positive list-strength effect with spaced strengthening and a negative list-strength effect with massed strengthening.

4.4.4. Simulation 4: Norman (2002, Experiment 1)

As described in Chapter 3, Norman (1999, 2002) introduced the strong-interference paradigm as an alternative to the mixed-pure paradigm. In the strong-interference

paradigm, only weak items are tested, thereby allowing experimenters to strengthen strong items to ceiling. Another difference between the two paradigms is that in the strong-interference paradigm, interference items are presented on both lists. The only difference is that on the weak-interference list, the interference items are given fewer presentations than on the strong-interference list.

Because it assumes that overall interference decreases as the strength of some items on a list increases, REM.1 cannot account for the positive list-strength effect observed in the strong-interference paradigm (Norman, 1999, 2002). The purpose of Simulation 4 was to investigate whether REM.3 can predict the pattern. In particular, I applied REM.3 to Experiment 1 of Norman (2002).

4.4.4.1. Description of the experiment. Norman (2002, Experiment 1) used the strong-interference paradigm to test for a list-strength effect in recognition, operationalized as better discrimination in the weak-interference list than the strong-interference list. Stimuli were unrelated, medium-frequency words. List type was manipulated within subjects, with order of conditions counterbalanced. On the weak-interference list, subjects studied five untested primacy buffers, 50 target and 50 interference items randomly combined, and five untested recency buffers. Subjects were unaware of the buffer/target/interference item distinction, and thus attempted to learn all items. The strong-interference list was identical to the weak-interference list, save that, following the five recency buffers, the 50 interference items were presented five more times. Therefore, the interference:target strength ratio was 1:1 on the weak-interference list and 6:1 on the strong-interference list. Note, as well, that interference items were all presented a second time before any were presented a third time, all were presented a third

time before any were presented a fourth time, and so on. Each repetition of the interference items was randomized; consequently, the minimum lag between interference-item repetitions was 49, and the maximum lag was 98.

Two methodological precautions bear mentioning: First, to attenuate the likelihood of rehearsal borrowing, on each study trial, subjects were asked to indicate whether the word's referent could fit into a banker's box. Second, to equate the study-test lag between conditions, a longer distractor task was interpolated between the study and test phases in the weak-interference condition than the strong-interference condition. Neither of these features were implemented in my simulations: To my knowledge, no one has incorporated different encoding tasks in REM. In addition, because I did not include context features in my simulations, equating study-test lag is not necessary.

Test lists consisted of the 50 targets and 50 distractors. At test, subjects rated their confidence that probes were "old" on a six-point scale. If they believed that a probe was "old" (i.e., if they selected 4, 5, or 6 on the confidence scale), they then made a remember/know judgment (Tulving, 1985). Norman (2002) computed d' by changing confidence ratings of 4, 5, and 6 to "old" responses and changing confidence ratings of 1, 2, or 3 to "new" responses. I did not simulate confidence-scale responding; rather, I simply had simulated subjects make old/new recognition decisions.

Norman's (2002) results produced a list-strength effect: d' was higher on the weak-interference list (M = 2.35, SE = 0.12) than the strong-interference list (M = 2.22, SE = 0.10). The hit rate was higher in the weak-interference list (M = .91) than the strong-interference list (M = .77). Interestingly, the false-alarm rate was also higher in the weak-

interference list (M = .22) than the strong-interference list (M = .11).[9] Hit and false-alarm rates, then, produced a concordant effect (i.e., hit and false-alarm rates increasing together) rather than a mirror effect (i.e., an increasing hit rate accompanied by a decreasing false-alarm rate).

4.4.4.2. Simulation 4A: REM.1. I began by verifying that REM.1 could not predict Norman's (2002, Experiment 1) results. Although the reasoning presented above is intuitively reasonable, to the best of my knowledge, there are no published reports simulating the strong-interference paradigm with REM.1. If REM.1 predicts Norman's results, then the use of a more complex model like REM.3 would violate the principle of parsimony.

4.4.4.2.1. Method. Recall that the purpose of the strong-interference paradigm is to strengthen strong items to ceiling on the strong-interference list. In Shiffrin and Steyvers (1997) simulations u_{strong} was set to .4. However, this keeps strong-item performance well below ceiling. Here, I varied u_{strong} (.4, .6, .8) to simulate increasingly strong interference items. I set u_{weak} to .28 as in Shiffrin and Steyvers' simulations.

Because Norman (2002) used medium-frequency words, I set g_{draw} to .38. The remaining parameters were fixed at values used in previous simulations: g_{base} = .4, w = 20, c = .7, and $criterion_{test}$ = 1 (recall that REM.1 does not have a $criterion_{study}$ parameter).

4.4.4.2.2. Results and discussion. Figure 4.9 shows hit rates, false-alarm rates, and d' as a function of u_{strong}. Examination of the hit and false-alarm rates reveals that REM.1

9. I do not have standard errors for hit and false-alarm rates, as Norman (2002) presented means as a function of confidence-scale responses.

correctly predicts the concordant effect observed by Norman (2002, Experiment 1). However, REM.1 predicts a negative list-strength effect, with better discrimination on the strong-interference list than the weak-interference list. Indeed, as u_{strong} increases, the magnitude of the strong-list advantage increases. Therefore, REM.1 is inconsistent with Norman's empirical results.

4.4.4.3. Simulation 4B: REM.3

4.4.4.3.1. Method. In my REM.3 simulations, I only varied the $criterion_{study}$ parameter (1, 2, 3, 4). As I describe below, for the present simulation, $criterion_{study}$ affected the qualitative pattern of results. In particular, a $criterion_{study}$ value of 1 failed to replicate the empirical pattern, but $criterion_{study}$ values of 2 and 3 did. Therefore, I also simulated the paradigm with a $criterion_{study}$ of 4 to ensure the pattern would continue with stricter values of $criterion_{study}$. The following fixed parameters were used: $g_{draw} = .38$, $g_{base} = .4$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$.

4.4.4.3.2. Results and discussion. Figure 4.10 shows hit rates, false-alarm rates, and d' as a function of $criterion_{study}$. Qualitatively, the results for $criterion_{study}$ values of 2, 3, and 4 replicate Norman's (2002, Experiment 1) results, with better discrimination on the weak-interference list than the strong-interference list. However, when $criterion_{study}$ is 1, REM.3 predicts equivalent discrimination for the weak- and strong-interference lists. Note, as well, that the hit and false-alarm rates replicate the concordant effect from Norman's experiment: Both hit and false-alarm rates are higher on the weak-interference list than the strong-interference list.

I am not concerned that a value of 1 for $criterion_{study}$ fails to produce a list-strength effect in REM.3. Based on the results of Simulation 1, it is clear that 1 is not a reasonable

value for the $criterion_{study}$ parameter. The fact that higher values of $criterion_{study}$ produce results that are more consistent with empirical data further bolsters this contention.

Prima facie, the concordant effect observed here and in Norman (2002, Experiment 1) is puzzling. Typically, strength manipulations produce a mirror effect (Criss, 2006; Stretch & Wixted, 1998). However, the way differentiation works in REM.3 provides an explanation for this pattern. Because of the probabilistic nature of superimposition in REM.3, memory will consist of more images following study of the strong-interference list than the weak-interference list. This is because, when a repeated item is not recognized as previously studied, a new image is generated. Thus, there is more interitem interference on the strong-interference list than the weak-interference list, producing a lower hit rate. Distractors, however, are more likely to erroneously match an image on the weak-interference list than the strong-interference list. This is because there is a greater degree of differentiation for interference items on the strong-interference list than the weak-interference list, and images from interference items are therefore less likely to erroneously match a distractor. As such, the false-alarm rate is also lower on the strong-interference list than the weak-interference list.

4.4.5. Simulation 5: Norman (1999, Experiment 4)

Norman (2002, Experiment 1) used relatively long study phases, with 360 trials in the strong-interference condition. To ensure that the REM.3 results from Simulation 4B were not an artifact of the large number of study trials, I also simulated Norman (1999, Experiment 4). This experiment had two between-subjects conditions, which I simulate in Simulations 5A and 5B. One condition was very similar to Norman (2002, Experiment 1), with the major difference being that the study list was shorter. In the other condition,

distractors were semantically related to targets (e.g., alligator-crocodile). Both experiments yielded a list-strength effect.

4.4.5.1. Simulation 5A: Unrelated distractors.

4.4.5.1.1. Description of the experiment. The procedure of Norman (1999, Experiment 4) was very similar to the procedure of Norman (2002, Experiment 1). The strong-interference paradigm was used, with interference items presented once in the weak-interference list and six times in the strong-interference list. However, there were 16 targets instead of 50, 48 interference items instead of 50, and three primacy and recency buffers instead of five. In addition, the test list contained only half of the targets from the study phase, along with eight distractors. Unlike the confidence judgments from Norman (2002, Experiment 1), Norman (1999, Experiment 4) used standard old/new recognition.

Norman (1999, Experiment 4) uncovered a list-strength effect as measured by d': Discrimination was better on the weak-interference list ($M = 2.28$, $SE = 0.08$) than the strong-interference list ($M = 1.90$, $SE = 0.09$). The hit and false-alarm rates produced the concordant effect observed in Norman (2002, Experiment 1): The hit rate was higher in the weak-interference list ($M = .91$, $SE = .02$) than the strong-interference list ($M = .66$, $SE = .03$), and the false-alarm rate was higher in the weak-interference list ($M = .12$, $SE = .01$) than the strong-interference list ($M = .03$, $SE = .01$).

4.4.5.1.2. Method. Other than the number of targets, interference items, primacy buffers, recency buffers, and distractors, the algorithm and parameters for this simulation were the same as those used in Simulation 4B.

4.4.5.1.3. Results and discussion. Figure 4.11 shows hit rates, false-alarm rates, and *d'* as a function of $criterion_{study}$. As in Simulation 4B, $criterion_{study}$ values of 2, 3, and 4 produced a list-strength effect, whereas a $criterion_{study}$ value of 1 produced a null list-strength effect. Simulation 5A also replicated the concordant effect observed in Simulation 4B and the empirical data.

4.4.5.2. Simulation 5B: Semantically related distractors.

4.4.5.2.1. Description of the experiment. This condition was identical to the condition from Simulation 5A, save that each distractor was semantically related to a target from the study phase. Results matched those from the unrelated-distractors condition: There was a list-strength effect, with higher discrimination (as measured by *d'*) in the weak-interference list ($M = 1.87$, $SE = 0.08$) than the strong-interference list ($M = 1.48$, $SE = 0.09$). The concordant effect was also replicated: The hit rate was higher in the weak-interference list ($M = .85$, $SE = .02$) than the strong-interference list ($M = .58$, SE = .03), and the false-alarm rate was higher in the weak-interference list ($M = .18$, $SE = .02$) than the strong-interference list ($M = .07$, $SE = .01$).

4.4.5.2.2. Method. To simulate semantic relatedness in REM, I set 8 features for each distractor equal to 8 features of one of the targets. Otherwise, Simulation 5B was identical to Simulation 5A.

4.4.5.2.3. Results and discussion. Figure 4.12 shows hit rates, false-alarm rates, and *d'* as a function of $criterion_{study}$. REM.3 replicates the concordant effect regardless of the value of $criterion_{study}$. Similarly, values of 3 or 4 for $criterion_{study}$ produce a list-strength effect, consistent with Norman's (1999, Experiment 4) results. However, a value

of 1 for $criterion_{study}$ produces a negative list-strength effect, and a value of 2 for $criterion_{study}$ produces a null list-strength effect.

4.5. Discussion of REM.3

The purpose of the simulations reported in the present chapter was to determine whether, when stripped of a simplifying assumption, REM is capable of explaining phenomena with which it was previously believed inconsistent. In REM.1, strengthened items are always stored in the same image; in REM.3, repetitions are superimposed on the original image only if the subject recognizes the repetition as previously studied. In Simulation 1, I showed that, although a value of 1 for $criterion_{test}$ is reasonable, it is not reasonable for $criterion_{study}$. Because subjects are not explicitly attempting to make old/new judgments at study, the threshold for superimposing a studied item on a previously generated image needs to be higher. In Simulation 2, I demonstrated that REM.3 correctly predicts an advantage for spaced over massed repetitions. In Simulation 3, I examined whether REM.3 predicts a negative or null list-strength effect with the mixed-pure paradigm in recognition. REM.3 actually predicts a slightly positive list-strength effect with spaced strengthening, but examination of published reports of the list-strength effect in recognition provides empirical support for this prediction. Finally, in Simulations 4 and 5, I showed that REM.3 can account for positive list-strength effects observed with the strong-interference paradigm.

REM.3 provides a more complete picture than REM.1. REM.1's simplifying assumption that repetitions are always accumulated in a single image, regardless of the time or number of intervening items between presentations, renders it unable to account for the spacing effect. Obviously, if massed and spaced repetitions are algorithmically

identical, they will be recognized at the same rate. Instead, if superimposition is made dependent on subjects' recognition that the item was previously studied (as in REM.3), then spaced repetitions obtain a mnemonic advantage over massed repetitions.

The list-strength effect is one of the phenomena responsible for the emergence of REM. Broadly, REM.1 is consistent with experiments on the list-strength effect: Strengthening some items on a list protects the weaker items from interitem interference, thus resulting in a negative list-strength effect. Yet, although REM.1 always predicts a negative list-strength effect, this is inconsistent with null or slightly positive list-strength effects observed in the literature with spaced strengthening (Ratcliff et al., 1990, 1992, 1994). Similarly, the prediction of an increasingly negative list-strength effect as strong-item strength increases renders REM.1 unable to explain positive list-strength effects found with the strong-interference paradigm (Norman, 1999, 2002). I have demonstrated that REM.3 is consistent with these results. In the following chapter, I empirically test a novel prediction generated by REM.3.

Chapter 5: The Empirical Test of REM.3

As described in the previous chapter, REM.3 is able to explain two findings that REM.1 cannot: the spacing effect and positive list-strength effects with the strong-interference paradigm. Chapter 4 also demonstrated that REM.3 predicts a slightly positive list-strength effect in the mixed-pure paradigm when strengthening is accomplished through spaced repetitions, a prediction that receives empirical support in the literature. REM.3's success over REM.1 in accounting for these phenomena stems from making superimposition of repetitions on originally generated images dependent on subjects' recognition that the item was previously studied.

The superimposition process during REM.3's study phase is similar to an REM test-phase algorithm introduced by Criss et al. (2011). In traditional instantiations of REM, the test phase does not involve the addition of new images to memory. This assumption is made on the basis of computational efficiency, rather than because of any commitment to the idea that test probes are not encoded.

In Criss et al.'s (2011) test-phase algorithm, when a test probe is called "old", the most similar image in memory is updated—that is, the probe is superimposed on the most similar image. When a probe is called "new", its features are copied to a new image. Consequently, false alarms produce incorrect superimposition—that is, storage of multiple items in a single image, and misses produce storage of one item across multiple images. Criss et al. demonstrated that this version of REM was capable of predicting output interference in old/new and forced-choice recognition (Norman & Waugh, 1968; Peixotto, 1947).

Given the similarities between REM.3 and Criss et al.'s (2011) test-phase algorithm, it makes sense to combine them (i.e., to use the study-phase algorithm from REM.3 and the test-phase algorithm from Criss et al.'s REM variant). This extension of REM.3 generates novel predictions regarding the mixed-pure paradigm, which are described below. First, however, it is necessary to discuss why REM predicts output interference.

Output interference in REM has an important, somewhat counterintuitive, property: namely, output interference arises entirely from "new" responses in recognition. "New" responses increase the size of the search set by adding images to memory. In contrast, the search set remains the same size when an "old" response is made, since new features are copied to an extant image. Of course, if the "old" response is a false alarm, the distractor's features are copied to an image that was generated by a different item at study; yet, this still attenuates interference, since the addition of any features to an image renders it "more complete", and therefore less similar to other images or test probes.

As demonstrated in Chapter 4, REM.3 predicts a positive list-strength effect when strengthening is accomplished via spaced repetitions. If the test-phase algorithm presented by Criss et al. (2011) is added to REM.3, then REM.3 makes a novel prediction concerning the list-strength effect in the mixed-pure paradigm: If weak items are tested before strong items on the mixed list, there will be a positive list-strength effect; if strong items are tested before weak items on the mixed list, there will be a null list-strength effect.

Why does the combination of REM.3 and Criss et al.'s (2011) output-interference version of REM produce this prediction? I will start by addressing this with respect to the

weak-tested-first condition, as this is the condition producing the counterintuitive prediction of a positive list-strength effect in recognition. Recall that, in REM, interference arises from additional images in memory, and differentiation is capable of offsetting interference. In REM.3, because superimposition of similar images is probabilistic, the number of images in memory increases from the pure-weak list to the mixed list to the pure-strong list. Therefore, interitem interference actually increases with list strength, but the degree to which images are differentiated in stronger lists compensates for this.

When weak items are tested first on the mixed list, they are evaluated with respect to a larger search set than pure-weak items, thereby giving first-half pure-weak items an advantage over the mixed-weak items. Even though some of the mixed-weak items' competitors are strong, and therefore differentiated, this is not enough to offset the larger search set. In contrast, mixed-strong items are compared to second-half pure-strong items. Prima facie, one might suspect that the mixed-weak items from the first half of the test list would produce more interference than first-half pure-strong items, thereby yielding a pure-strong advantage. However, recall that hits result in updating the most similar image, thereby increasing differentiation of that trace and reducing its interference. Hits to mixed-weak items, therefore, increase mixed-strong performance by reducing interitem interference. Why does this not hold for the pure-strong list which, after all, will likely have more hits in the first half of the list than the mixed list? The reason is that increases in differentiation yield diminishing returns. pure-strong images are already well differentiated by the time they are tested, so updating them does not produce much more protection. When coupled with the fact that there are more images in the pure-strong

condition than the mixed-strong condition, performance ends up being better for mixed-strong items than second-half pure-strong items.

Next, consider the strong-tested-first condition, which REM.3 and Criss et al.'s (2011) output-interference version of REM predict will produce a null list-strength effect. Here, mixed-weak items are compared to second-half pure-weak items. As before, pure-weak items will be evaluated with respect to a smaller search set than mixed-weak items. However, because of the higher hit rate in the first half of the mixed list, the second half of the test lists will have search sets closer in size. This is because the larger number of misses in the first half of the pure-weak list will cause the search set to increase in size more rapidly than the mixed list. This results in roughly equal performance for mixed-weak items and second-half pure-weak items.

For strong items, although the pure-strong list will have more images than the mixed list, the interference caused by the mixed-weak items will offset this difference, resulting in roughly comparable performance. Notice that, when strong items are tested before weak items, the protection from increasing the differentiation of mixed-weak items does not occur until after the mixed-strong items are tested, and thus the model's predictions are similar to standard REM instantiations.

5.1. Simulating Output Interference in the Mixed-Pure Paradigm

5.1.1. Simulation 1: Weak-tested first

5.1.1.1. Method. I simulated the mixed-pure paradigm using the study-phase algorithm from REM.3 and the test-phase algorithm from Criss et al.'s (2011) output-interference version of REM. On the pure-weak list, 84 targets were studied once; on the pure-strong list, 84 targets were studied twice; and on the mixed list, 42 targets were

studied once and 42 targets were studied twice. List construction was modelled after the "spaced repetitions in the same order" lists from Experiment 3 of Malmberg and Shiffrin (2005). In free recall, Malmberg and Shiffrin showed that this form of strengthening produced the largest list-strength effect. For the pure-strong list, targets were first divided into 12, nonoverlapping, seven-item sublists. After a sublist was presented, it was re-presented immediately in the same order. Therefore, all repetitions had a lag of six. Strong items in the mixed list were strengthened in the same fashion, with the 42 strong targets divided into six sublists. Note that, on the mixed list, all strong items and their repetitions were presented before any weak items. This is a methodological precaution taken with human subjects to prevent the possibility of rehearsal borrowing (i.e., studying mixed-weak items during the presentation of mixed-strong items; see Malmberg & Shiffrin, 2005; Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1990).

Test lists consisted of the 84 targets along with 84 distractors. For the mixed list, the first 84 test probes consisted of the 42 mixed-weak targets and 42 distractors; the second half of the test list consisted of the 42 mixed-strong targets and 42 distractors. The pure test lists were matched to the mixed list, such that the last 42 targets from the study list appeared in the first half of the test list, and the first 42 items from the study list appeared in the second half of the test list. This allows for item-strength comparisons between mixed and pure list without confounding study and test position: The first 42 study items on the mixed list are mixed-strong items, and these are tested in the second half of the test list. These can be compared to the first 42 targets from the pure-strong list, which are also tested in the second half of the test list. Similarly, the second 42 study items on the mixed list are pure-weak items, and these are tested in the first half of the

test list. Performance on these items can be compared to performance on the second 42 study items from the pure-weak list, which also appear in the first half of the test list.

As in the REM.3 simulations from Chapter 4, I varied $criterion_{study}$ (1, 2, 3). The remaining parameters were fixed as follows: $g_{base} = .4$, $g_{draw} = .35$, $u_1 = .28$, $u_2 = .12$, $c = .7$, and $criterion_{test} = 1$.

The study-phase algorithm was the same as that used in Chapter 4. As each item was studied, it was matched to all of the images in memory. If the odds that the item was studied earlier exceeded $criterion_{study}$, the study item was superimposed on the most similar image, with features copied to the old image with probability u_2. As before, only positions for which the image had no information stored could be updated—that is, nonzero values were skipped in updating the image.

The algorithm for the test phase followed that used by Criss et al. (2011) in their output-interference simulations. When a probe was identified as "old", it was superimposed on the most similar image, with features copied with probability u_2. When a probe was identified as "new", its features were copied to a new image with probability u_1.

The simulations represent a 2 (strength: weak vs. strong) × 2 (list type: pure vs. mixed) × 3 ($criterion_{study}$: 1, 2, 3) mixed design, with the first two factors within subjects and the last factor between subjects. There were 3000 simulated subjects, with 1000 in each of the $criterion_{study}$ conditions.

5.1.1.2. Results. Hit and false-alarm rates are shown in Figure 5.1 and d' is shown in Figure 5.2. Like the majority of the REM.3 simulations in Chapter 4, the qualitative pattern of results was unaffected by $criterion_{study}$. Therefore, I collapsed across

$criterion_{study}$ to consider the results. Examination of performance confirms that output interference occurred: The first half of the pure-weak list had a higher hit rate (M = .773, SE = .001), lower false-alarm rate (M = .349, SE = .001), and higher d' (M = 1.163, SE = 0.006) compared to the second half (hit rate: M = .711, SE = .001; false-alarm rate: M = .350, SE = .001; d': M = 0.961, SE = 0.005). Output interference also occurred on the pure-strong list: The first half of the pure-strong list had a higher hit rate (M = .903, SE = .001), lower false-alarm rate (M = .325, SE = .001), and higher d' (M = 1.830, SE = 0.007) compared to the second half (hit rate: M = .847, SE = .001; false-alarm rate: M = .332, SE = .001; d': M = 1.510, SE = .006).

Of greater importance for present purposes is the list-strength effect. In Bäuml's (1997) cued-recall experiment on the list-strength effect, he controlled test position like I have here. He performed two analyses: one for which test position was ignored, and one for which test position was controlled. I take the same approach. Ignoring test position (i.e., using d' from the entire pure-weak and pure-strong lists), there is a null list-strength effect: d' is slightly higher for pure-strong items (M = 1.629, SE = 0.005) than mixed-strong items (M = 1.622, SE = 0.007), and d' is slightly higher for mixed-weak items (M = 1.063, SE = 0.006) than pure-weak items (M = 1.046, SE = 0.004). The strong-to-weak ratios were 1.526 and 1.557 for mixed and pure lists, respectively, yielding a slightly negative list-strength effect (R_r = 0.980). Therefore, ignoring test position produces the standard finding of a null list-strength effect in recognition.

The above analysis confounds list type with test position. Test probes from the first half of the pure-strong list are evaluated with less output interference than any of the mixed-strong items, since the testing of mixed-strong items is preceded by the testing of

mixed-weak items. Similarly, test probes from the second half of the pure-weak list are evaluated with more output interference than any mixed-weak items, since all mixed-weak items are tested in the first half of the mixed test list. To control for this, mixed-strong performance can be compared to second-half pure-strong performance, and mixed-weak performance can be compared to first-half pure-weak performance. This analysis produces a positive list-strength effect: d' is higher for mixed-strong items than second-half pure-strong items (M = 1.510, SE = 0.006), and d' is higher for pure-weak items (M = 1.163, SE = 0.005) than mixed-weak items. The strong-to-weak ratios were 1.526 and 1.298 for mixed and pure lists, respectively, yielding a positive list-strength effect (R_r = 1.175). Therefore, when controlling for test position, REM.3 predicts a positive list-strength effect, contrary to most results in the literature (Ratcliff et al., 1990).

5.1.1.2. Discussion. The present simulation demonstrates that REM.3 predicts a positive list-strength effect when weak items are tested before strong items on the mixed list. This is a function of output interference and REM.3's differentiation process. Study lists have 84, 126, and 168 trials in the pure-weak, mixed, and pure-strong lists, respectively. Because single-image storage in REM.3 is probabilistic rather than inevitable, some of the repeated items in the mixed and pure-strong lists are stored in separate images. In effect, then, the search set is smallest for the pure-weak list, largest for the pure-strong list, and intermediate for the mixed list. When sufficient differentiation occurs (i.e., when a sufficient number of repeated items are stored in a single image), the decreased probability of interference from differentiated images offsets the costs associated with larger search sets. Here, the search set also grows as the test list proceeds, from the addition of probes identified as "new" to memory. So, although pure-

strong items are more protected from interitem interference than mixed-strong items, this advantage is offset by the larger search set on the pure-strong list than the mixed-strong list (stemming from adding "new" items from the test list to memory and from strong-item repetitions that were inadvertently stored in separate images). Similarly, although mixed-weak items are shielded from interitem interference by the presence of mixed-strong items, they nevertheless are more poorly recognized than pure-weak items because of the larger search set on the mixed list compared to the pure-weak list. In other words, REM.3 predicts a positive list-strength effect because cue overload (i.e., a larger search set) overwhelms occlusion.

5.1.2. Simulation 2: Strong-tested first

5.1.2.1. Method. Simulation 2 was identical to Simulation 1, save that first-half study items were tested in the first half of the test list and second-half study items were tested in the second half of the test list. This means that, on mixed lists, strong items were tested before weak items.

5.1.2.2. Results. Hit and false-alarm rates are shown in Figure 5.3 and *d′* is shown in Figure 5.4. The qualitative pattern of results was unaffected by the $criterion_{study}$ parameter, so I collapsed across $criterion_{study}$ to consider the results. On the pure-weak list, output interference was fairly modest: Although the first half of the test list produced a higher hit rate (M = .749, SE = .001) and better discrimination (M = 1.081, SE = 0.006) than the second half (hit rate: M = .732, SE = .001; *d′*: M = 1.022, SE = 0.006), the false-alarm rate was slightly higher on the second half of the pure-weak test (M = .352, SE = .001) than the first half (M = .350, SE = .001). On the pure-strong list, output interference was reversed: The second half of the test list produced a higher hit rate (M = .880, SE =

.001) and higher d' (M = 1.662, SE = 0.007) than the first half (hit rate: M = .870, SE = .001; d': M = 1.634, SE = 0.006). However, the false-alarm rate did increase from the first half of the test list (M = .328, SE = .001) to the second half (M = .334, SE = .001), consistent with output interference.

Ignoring test position, there was a null list-strength effect: d' was higher for mixed-strong items (M = 1.692, SE = 0.007) than pure-strong items (M = 1.618, SE = 0.005), and d' was higher for mixed-weak items (M = 1.048, SE = 0.005) than pure-weak items (M = 1.039, SE = 0.004). The strong-to-weak ratios were 1.615 and 1.558 for mixed and pure lists, respectively, yielding a slightly positive list-strength effect (R_r = 1.037).

To unconfound list type from test position, I compared mixed-strong items to first-half pure-strong items, and mixed-weak items to second-half pure-weak items. With test position controlled, there was still a null list-strength effect: Mixed-strong items were better recognized than pure-strong items (M = 1.634, SE = 0.006), and mixed-weak items were better recognized than pure-weak items (M = 1.022, SE = 0.004). The strong-to-weak ratios were 1.615 and 1.599 for mixed and pure lists, respectively, yielding a slightly positive list-strength effect (R_r = 1.010).

5.1.2.3. Discussion. Concerning the list-strength effect, the results from Simulation 2 are unambiguous and broadly consistent with REM.1 (Shiffrin & Steyvers, 1997): Whether or not test position is controlled, REM.3 predicts a null list-strength effect when strong items are tested before weak items on the mixed list.

A more surprising finding from Simulation 2 is that output interference was not observed on the pure-strong list. Yet, reduced or absent output interference on a strongly encoded list is consistent with differentiation: In the output-interference version of REM,

images are updated when "old" responses are made, thereby increasing image differentiation and reducing interitem interference for other items. So, as subjects move through a test list on which the hit rate is relatively high, later-tested targets are easier to identify as "old" because there is less interitem interference. REM simulations and empirical data consistent with this claim are provided by Kılıç et al. (2017).

If no output interference with strong lists is the norm, then why was output interference observed on the pure-strong list in Simulation 1? Recall that, in Simulation 1, second-half pure-strong items were tested first, but that in Simulation 2, first-half pure-strong items were tested first. In an REM.3 study phase, items repeated later in a study list are more likely to be stored in separate images than items repeated early in the list. This is because, with each successive study trial, the evidence needed for superimposition of similar images to occur increases as a result of the size of the search set increasing. On the pure-strong list, the first repeated item occurred in Study Position 8. When this item was studied, subjects needed to identify that it was earlier studied in a search set of seven items. This is akin to successfully identifying that a probe is "old" following study of a seven-item list. In contrast, the first repeated item from the second half of the study list occurred in Study Position 98. By this point, the search set will be much larger, so more evidence is needed to identify that the item was previously studied. Consequently, successful superimposition follows patterns observed with the list-length effect (Strong, 1912).

In Chapter 4, I showed that REM.3 predicts the spacing effect because spaced repetitions are sometimes stored in separate images. The same phenomenon is operating for second-half pure-strong items in Simulation 1: They are more likely to be in multiple

traces, and thus have a mnemonic benefit over first-half pure-strong items, which are more likely to be in single images. Recall that, when comparing massed and spaced repetitions in REM.3, spaced repetitions that are correctly superimposed are functionally identical to massed items. Therefore, the output interference observed in Simulation 1 was primarily driven by study-phase mechanisms rather than test-phase mechanisms.

5.2. Experiment 1

The purpose of Experiment 1 was to test the prediction from Simulation 1: namely, does a positive list-strength effect occur when weak items are tested before strong items on the mixed list? In Chapter 3, I reviewed several theoretical accounts of the list-strength effect. Next, I briefly consider what predictions these various accounts make regarding the present experiment.

5.2.1. Predictions

5.2.1.1. Differentiation. As shown in Simulation 1, REM.3 predicts a positive list-strength effect when weak items are tested first on the weak list. However, this prediction differs from classic differentiation models (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997; Shiffrin et al., 1990), wherein item repetitions are always accumulated in a single mnemonic trace. From the perspective of classic differentiation accounts, Experiment 1 should produce a negative list-strength effect.

5.2.1.2. Criterion placement. Recall that, according to the criterion-placement account (see Hirshman, 1995), the null list-strength effect arises from subjects adopting increasingly conservative response criteria as the average strength of study items increases. Although this account predicts a hit rate consistent with a positive list-strength effect, the false-alarm rate is in the direction of a negative list-strength effect, thereby

yielding a null list-strength effect as assessed by discrimination. The criterion-placement account therefore predicts a null list-strength effect in the present case.

5.2.1.3. BCDMEM and the continuous-memory version of TODAM. Both of these accounts predict a null list-strength effect for the present experiment. In BCDMEM (Dennis & Humphreys, 2001), interference is only caused by the previous contexts in which items have been encountered. Interitem interference does not occur in BCDMEM, and therefore recognition of strong and weak items should be unaffected by list composition.

Although the continuous-memory version of TODAM (Murdock & Kahana, 1993a, 1993b) does not reject the notion of interitem interference, it assumes that this interference arises from both experimental and pre-experimental memories. The strong items in a study list are not strong enough to overcome all of a subject's pre-experimental memories, and thus the continuous-memory version of TODAM predicts a null list-strength effect.

5.2.1.4. Insufficient-strengthening and dual-process accounts. According to Norman (1999, 2002), strengthening in the mixed-pure paradigm is insufficient to produce interference. In the present case, strong items are presented twice and weak items are presented once—this is a considerably smaller strength ratio than has been used in the strong-interference paradigm. Therefore, from the perspective of the insufficient-strengthening account, Experiment 1 should produce a null list-strength effect.

What about dual-process accounts? According to dual-process accounts of the list-strength effect (Diana & Reder, 2005; Norman, 2002), positive list-strength effects are driven by recollection. If a recognition test forces subjects to adopt memory strategies

they would use in a free-recall test, there should be a positive list-strength effect; otherwise, the list-strength effect should be negative.

Experiment 1 does not provide a direct test of the dual-process account, inasmuch as the remember/know procedure is not included. There is no mechanism in Experiment 1 for separating recollection from familiarity. However, if the positive list-strength effect relies on recall-like processes, then one should not expect a positive list-strength effect in Experiment 1. In free recall, strong items are output earlier than weak items on mixed strength lists (Wike & Wike, 1970; Wixted et al., 1997); in Experiment 1, this is reversed, such that weak items are tested before strong items. According to dual-process accounts, recollection-based phenomena should occur in recognition to the degree that the recognition test mimics processes observed in free recall. This being the case, one could tentatively predict a negative list-strength effect in Experiment 1.

5.2.1.5. Inhibition. According to inhibition theory (Anderson, 2003; Norman et al., 2007), forgetting occurs as a result of active remembering. When a target is retrieved, its competitors are made less accessible, even with independent cues. What triggers inhibition? According to inhibition theory, inhibition is retrieval dependent—simply restudying an item does not result in the inhibition of its competitors. In his cued-recall experiment, Bäuml (1997) found that controlling test position resulted in a null list-strength effect, but that confounding strength with test position produced a positive list-strength effect. This is because, in list-strength experiments, retrieval does not occur until the test phase. Therefore, inhibition is absent at the start of the test list, since nothing has yet been retrieved. However, as subjects move through the test list, identifying targets as "old" causes other targets to be suppressed, thereby reducing performance. Therefore,

consistent with Bäuml's cued-recall experiment, inhibition theory predicts a null list-strength effect when controlling test position but a positive list-strength effect when confounding test position. Notice, then, that the prediction from inhibition is opposite the prediction from REM.3, which predicts a positive list-strength effect when test position is controlled but a null list-strength effect when test position and list type are confounded.

5.2.1.6. Output interference. The REM.3 predictions presented in Simulation 1 are based on output interference, and thus it is difficult to disentangle the REM account from the output-interference account. However, it is worth noting that, to the degree that output interference has been used to explain the list-strength effect (e.g., Tulving & Hastie, 1972; Wixted et al., 1997), the explanations have focused on mixed-strong items interfering with mixed-weak items. In other words, output-interference accounts have made a different prediction from that made by REM: There should be a positive list-strength effect when mixed-strong items are tested before mixed-weak items, but not when mixed-weak items are tested before mixed-strong items.

5.2.2. Method

5.2.2.1. Subjects. Twenty-five Memorial University of Newfoundland students participated in exchange for course credit or a $10.00 honourarium.

5.2.2.2. Materials. The experiment was programmed in Psychopy (version 1.82.01; Peirce, 2007) and presented on Macintosh desktop computers located in individual, sound-attenuating cubicles. Stimuli were 516 words obtained from the MRC Psycholinguistic Database (Coltheart, 1981). All words were 5 to 12 letters in length (M = 7.023, SE = 0.081) and had Kucera-Francis written frequencies (Kucera & Francis, 1967) of 30 to 200 occurrences per million (M = 72.194, SE = 1.770). Words had mean

familiarity, concreteness, and imageability ratings of 530.630 (*SE* = 1.686, range: 389-590), 390.108 (*SE* = 4.135, range: 217-589), and 428.091 (*SE* = 3.521, range: 224-589), respectively. For each subject, stimuli were sampled at random without replacement to create the study and test lists.

All study and test lists began and ended with two untested serial-position buffers, and consisted of 84 targets. On the pure-weak list, all targets were presented once; on the pure-strong list, all targets were presented twice; and on the mixed list, half of the targets were presented once and half were presented twice. In the mixed list, all presentations of strong items occurred before any weak items were presented. Strong items were repeated with a six-item lag, as in Simulations 1 and 2 (modelled after Malmberg & Shiffrin, 2005). The first 42 targets studied always appeared in the second half of the test list, and the second 42 targets studied always appeared in the first half of the test list. This meant that, for the mixed list, weak items were tested before strong items. Note that each half of the test list had 42 targets and 42 distractors.

5.2.2.3. Procedure. Subjects cycled through three phases three times: study, distraction, and test. The order of lists was randomized for each subject.

5.2.2.3.1. Study phase. During the study phase, targets appeared one at a time in the centre of the computer screen. Words were presented in black, Times New Roman font on a white background. Stimuli were presented at a 2-s rate with no interstimulus interval. Subjects were asked to learn the words for an upcoming memory test, the nature of which was explained.

5.2.2.3.2. Distraction phase. During the distraction phase, subjects solved math problems of the form $AB+C = D$? where A, B, and C were integers from 2 to 10 and D

was always equal to $AB+C$ or was displaced by ±1. Subjects were instructed to press the "y" key if an equation was correct or the "n" key if the equation was incorrect. The probability of $AB+C = D$ was .5, the probability of $AB+C = D+1$ was .25, and the probability of $AB+C = D-1$ was .25. Values for A, B, and C were drawn randomly with replacement; therefore, there was a chance that subjects could see the same equation more than once during an experimental session.

The purpose of the math task was to equate the study-test lag between list types, an approach used by Malmberg and Shiffrin (2005). Subjects solved math equations for 76, 160, and 244 s in the pure-strong, mixed, and pure-weak conditions, respectively.

5.2.2.3.3. Test phase. The old/new recognition test was subject paced. For each probe, subjects pressed the "y" key if they believed it was a target and the "n" key otherwise. Feedback was not provided.

5.2.3. Results

In the present paper, generalized eta squared (η_g^2; Olejnik & Algina, 2003) is the effect size reported for F tests and Cohen's d (Cohen, 1992) is the effect size reported for t tests. Statistical analyses were carried out using R (version 3.4.1; R Core Team, 2017). Analyses of variance were performed using the ez package (version 4.4-0; Lawrence, 2016), and Cohen's d was computed using the effsize package (version 0.7.1; Torchiano, 2018). I used the plyr (version 1.8.4; Wickham, 2011) and reshape2 (version 1.4.3; Wickham, 2007) packages for computing descriptive statistics and structuring data frames, respectively.

Across the experimental session, subjects saw an average of 101.640 (*SE* = 7.988, range: 43-184) math problems. Performance was high, with a mean of 95.246% (*SE* = 0.565) answered correctly.

Hit and false-alarm rates are shown in Table 5.1, and *d′* is shown in Table 5.2. The dependent variable of interest was *d′*.

I first analyzed *d′* without regard to test position (e.g., Bäuml, 1997; Verde, 2009). Inspection of the means shows that pure-strong items (*M* = 1.291, *SE* = 0.116) were better recognized than mixed-strong items (*M* = 0.997, *SE* = 0.117), but that mixed-weak items (*M* = 0.915, *SE* = 0.111) were better recognized than pure-weak items (*M* = 0.788, *SE* = 0.116). The strong-to-weak ratios were 1.089 and 1.639 for mixed and pure lists, respectively, yielding a negative list-strength effect (R_r = 0.665). A 2 (strength: weak vs. strong) × 2 (list type: pure vs. mixed) repeated-measures analysis of variance produced a main effect of strength, $F(1, 24) = 28.260$, $p < .001$, $\eta_g^2 = .064$, with discrimination better for strong items (*M* = 1.144, *SE* = 0.092) than weak items (*M* = 0.852, *SE* = 0.098). The main effect of list type was not significant, $F(1, 24) = 0.790$, $p = .383$, $\eta_g^2 = .005$. The interaction between strength and list type was significant, $F(1, 24) = 5.728$, $p = .025$, $\eta_g^2 = .034$, supporting the reliability of the negative R_r.

Planned contrasts compared weak- and strong-item performance in mixed and pure lists. Pure-strong performance was better than mixed-strong performance, although this result did not reach the adopted significance level, $t(24) = 2.043$, $p = .052$, $d = 0.407$. The discrimination advantage for pure-weak items over mixed-weak items was not significant, $t(24) = 1.158$, $p = .258$, $d = 0.232$.

I next conditionalized performance on list half, thereby deconfounding test position and list type. Second-half pure-strong items (M = 1.163, SE = 0.139) were better recognized than mixed-strong items, but mixed-weak items were better recognized than first-half pure-weak items (M = 0.814, SE = 0.141). The strong-to-weak ratios were 1.089 and 1.427 for mixed and pure lists, respectively, yielding a negative list-strength effect (R_r = 0.763). A 2 (strength: weak vs. strong) × 2 (list type: pure vs. mixed) repeated-measures analysis of variance produced a main effect of strength, $F(1, 24) = 9.009$, p = .006, η_g^2 = .029, with discrimination better for strong items (M = 1.080, SE = 0.099) than weak items (M = 0.865, SE = 0.113). Neither the main effect of list type, $F(1, 24)$ = 0.080, p = .779, η_g^2 = .001, nor the interaction between strength and list type, $F(1, 24)$ = 2.606, p = .120, η_g^2 = .011, were significant.

Planned contrasts produced a nonsignificant advantage for second-half pure-strong items over mixed-strong items, $t(24) = 1.018$, p = .319, d = 0.204, and a nonsignificant advantage for mixed-weak items over first-half pure-weak items, $t(24) = 0.886$, p = .385, d = 0.177.

5.2.4. Discussion

The purpose of Experiment 1 was to test a prediction derived from REM.3 (see Figures 5.1 and 5.2). According to these simulations, structuring the mixed list such that all of the weak items were tested before any of the strong items should have produced a positive list-strength effect when performance was conditionalized on list half and a null list-strength effect when confounding test position and list type. Although Experiment 1 produced a null list-strength effect when test position and list type were confounded, this is not a surprising finding, given the plethora of null list-strength effects in the literature

(Ratcliff et al., 1990, 1992, 1994). More importantly for present purposes, Experiment 1 failed to find evidence for the simulation predictions; indeed, there was a statistically significant negative list-strength effect instead of the predicted positive list-strength effect.

The results of Experiment 1 are consistent with classic differentiation models like REM.1 (Shiffrin & Steyvers, 1997), the differentiation version of SAM (Shiffrin et al., 1990), and SLIM (McClelland & Chappell, 1998). However, as shown in Chapter 4, several findings in the literature—the spacing effect, positive list-strength effects with spaced strengthening techniques, and list-strength effects found with the strong-interference paradigm—render classic differentiation models untenable.

The fact that Experiment 1 produced a significant negative list-strength effect is also inconsistent with BCDMEM (Dennis & Humphreys, 2001) and the continuous-memory version of TODAM (Murdock & Kahana, 1993a, 1993b). This is because both of these models predict null list-strength effects rather than negative list-strength effects.

Interestingly, the results of Experiment 1 are consistent with a free-recall-based output-interference account (Tulving & Hastie, 1972; Wixted et al., 1997). According to this account, the list-strength effect occurs because subjects recall strong items earlier than weak items on mixed lists, thereby subjecting mixed-weak items to more output interference than pure-weak items. In Experiment 2, I more directly tested the output-interference account by testing strong items before weak items on the mixed list.

In addition to testing the output-interference account, Experiment 2 offers a direct test of the REM.3 simulations presented in Figures 5.3 and 5.4. REM.3 predicts null list-

strength effects when strong items are tested before weak items, regardless of analysis method (i.e., confounding or unconfounding test position and list type).

5.3. Experiment 2

5.3.1. Method

5.3.1.1. Subjects. Twenty-five Memorial University of Newfoundland students participated in exchange for course credit or a $10.00 honourarium. Computer error led to the loss of three subjects' data, resulting in a sample size of 22.

5.3.1.2.Materials and procedure. Experiment 2 was identical to Experiment 1, save that the first 42 targets from the study list were tested in the first half of the test list, and the last 42 targets from the study list were tested in the second half of the test list. This meant that, for mixed lists, strong items were tested before weak items.

5.3.2. Results

Across the experimental session, subjects saw an average of 112.727 (SE = 9.905, range: 52-224) math problems. Performance was high, with a mean of 94.700% (SE = 1.234) answered correctly.

Mean hit and false-alarm rates are reported in Table 5.3, and mean d' is reported in Table 5.4.

I first analyzed results without regard to test position. Inspection of the means shows that mixed-strong items (M = 1.437, SE = 0.131) were better recognized than pure-strong items (M = 1.137, SE = 0.125), but that pure-weak items (M = 0.985, SE = 0.143) were better recognized than mixed-weak items (M = 0.628, SE = 0.113). The strong-to-weak ratios were 2.288 and 1.154, respectively for mixed and pure lists, yielding a positive list-strength effect (R_r = 1.983). A 2 (strength: weak vs. strong) × 2 (list type:

pure vs. mixed) repeated-measures analysis of variance produced a main effect of strength, $F(1, 21) = 24.967$, $p < .001$, $\eta_g^2 = .143$, with discrimination better for strong items ($M = 1.287$, $SE = 0.106$) than weak items ($M = 0.806$, $SE = 0.095$). The main effect of list type was not significant, $F(1, 21) = 0.043$, $p = .838$, $\eta_g^2 = .001$. The interaction between strength and list type was significant, $F(1, 21) = 17.002$, $p < .001$, $\eta_g^2 = .072$, supporting the reliability of the R_r.

Planned contrasts compared weak- and strong-item performance in mixed and pure lists. Performance was significantly better for mixed-strong items than pure-strong items, $t(21) = 2.100$, $p = .048$, $d = 0.448$. There was also an advantage for pure-weak items over mixed-weak items, although this advantage did not reach the adopted significance level, $t(21) = 2.058$, $p = .052$, $d = 0.439$.

I next analyzed d' conditionalized on list half. Inspection of the means shows that mixed-strong items were better recognized than first-half pure-strong items ($M = 1.277$, $SE = 0.163$), but that performance was better for second-half pure-weak items ($M = 0.874$, $SE = 0.159$) than mixed-weak items. The strong-to-weak ratios were 2.288 and 1.465 for mixed and pure lists, respectively, yielding a positive list-strength effect ($R_r = 1.562$). A 2 (strength: weak vs. strong) × 2 (list type: pure vs. mixed) repeated-measures analysis of variance produced a main effect of strength, $F(1, 21) = 25.476$, $p < .001$, $\eta_g^2 = .177$, with discrimination better for strong items ($M = 1.357$, $SE = 0.116$) than weak items ($M = 0.750$, $SE = 0.109$). The main effect of list type was not significant, $F(1, 21) = 0.072$, $p = .790$, $\eta_g^2 = .001$. The interaction between strength and list type was significant, $F(1, 21) = 6.051$, $p = .023$, $\eta_g^2 = .023$, supporting the reliability of the R_r.

Planned contrasts compared weak- and strong-item performance across mixed and pure lists. There was a nonsignificant advantage for mixed-strong items over pure-strong items, $t(21) = 0.883$, $p = .387$, $d = 0.188$. There was also a nonsignificant advantage for pure-weak items over mixed-weak items, $t(21) = 1.448$, $p = .163$, $d = 0.308$.

5.3.3. Discussion

Experiment 2 was designed to test contrasting predictions from REM.3 and the output-interference account of the list-strength effect. According to REM.3, testing strong items before weak items on the mixed list should produce a null list-strength effect. In contrast, according to the output-interference account, testing strong items before weak items should produce a positive list-strength effect. The results of Experiment 2 clearly support the output-interference account: Regardless of analysis method, there was a significant, positive list-strength effect as measured by the R_r.

Combined, the results of Experiments 1 and 2 provide support for the output-interference account and contradict predictions from REM.3. I discuss the implications of these findings in more detail in Chapter 6. First, however, I wanted to verify that my stimuli could produce a typical null list-strength effect using the standard mixed-pure paradigm, wherein test lists are randomized without regard to study position or item strength.

5.4. Experiment 3

5.4.1. Method

5.4.1.1. Subjects. Twenty-five Memorial University of Newfoundland students participated in exchange for course credit or a $10.00 honourarium.

5.4.1.2. Materials and procedure. Experiment 3 was identical to Experiments 1 and 2, save that test lists were randomized as in the standard mixed-pure paradigm (Ratcliff et al., 1990). This meant that, unlike Experiments 1 and 2, the distribution of targets and distractors could differ between list halves (i.e., there was no longer a guarantee that the first and second halves of the test list contain 42 targets and 42 distractors).

5.4.2. Results

Across the experimental session, subjects saw an average of 107.720 (*SE* = 8.084, range: 40-222) math problems. Performance was high, with a mean of 93.767% (*SE* = 1.128) answered correctly.

Mean hit and false-alarm rates are reported in Table 5.5, and mean d' is reported in Table 5.6. Note that, because mixed-weak and mixed-strong items were not blocked at test, the same false-alarm rate is used to compute d' for these conditions. Although separate false-alarm rates is preferable, a common false-alarm rate for mixed-strong and mixed-weak items is standard in the list-strength literature (e.g., Ratcliff et al., 1990).

Because test lists were randomized, there is no way to control for test position. Therefore, I performed a single analysis using the entirety of the pure lists. Inspection of the means shows that mixed-strong items (M = 1.108, SE = 0.139) were better recognized than pure-strong items (M = 1.041, SE = 0.138), and that mixed-weak items (M = 0.746, SE = 0.113) were recognized at the same rate as pure-weak items (M = 0.746, SE = 0.111). The strong-to-weak ratios were 1.484 and 1.396 in mixed and pure lists, respectively, yielding a slightly positive list-strength effect (R_r = 1.063). A 2 (strength: weak vs. strong) × 2 (list type: pure vs. mixed) repeated-measures analysis of variance

produced a main effect of strength, $F(1, 24) = 21.907$, $p < .001$, $\eta_g^2 = .066$, with discrimination better for strong items ($M = 1.074$, $SE = 0.122$) than weak items (M = 0.746, SE = 0.101). Neither the main effect of list type, $F(1, 24) = 0.126$, $p = .726$, $_{\eta g}^2 = .001$, nor the interaction between strength and list type, $F(1, 24) = 0.262$, $p = .614$, $\eta_g^2 = .001$, were significant.

Planned contrasts compared performance on strong and weak items across list type. Mixed-strong items had a nonsignificant advantage over pure-strong items, $t(24) = 0.511$, $p = .614$, $d = 0.102$, and mixed-weak items had a nonsignificant advantage over pure-weak items, $t(24) = 0.007$, $p = .995$, $d = 0.001$.

5.4.3. Discussion

Experiment 3 had a straightforward purpose: to verify that the stimuli used in Experiments 1 and 2 would produce a standard list-strength effect when test lists were randomized. Although the list-strength effect was slightly positive, neither the interaction between strength and list type, nor the pairwise comparisons across list type, produced evidence that the positive R_r was reliable. Notably, the fact that the list-strength effect was slightly positive is consistent with the massed/spaced discrepancy documented in Chapter 4 (see Table 4.1). Using a spaced strengthening technique, Experiment 3 produced a slightly positive list-strength effect rather than the more common negative list-strength effect (Ratcliff et al., 1990).

In Chapter 6, I consider the theoretical implications of the REM.3 simulations presented in Chapter 4 and the experimental findings presented in this chapter.

Chapter 6: General Discussion

6.1. Summary of Findings

The REM model—introduced by Shiffrin and Steyvers (1997)—has provided an account of a wide variety of memory phenomena, ranging from the word-frequency mirror effect (Malmberg et al., 2004) to implicit-memory phenomena (Schooler et al., 2001). A simplifying assumption in most versions of REM is that item strengthening always accumulates in a single image. Although this simplification drastically reduces the complexity of the simulations, it renders REM unable to predict a number of important phenomena, including the spacing effect (e.g., Strong, 1916), slightly positive list-strength effects with spaced strengthening (e.g., Ratcliff et al., 1990, Experiment 5), and positive list-strength effects observed with the strong-interference paradigm (e.g., Norman, 2002). In Chapter 4, I demonstrated that a REM version Shiffrin and Steyvers termed REM.3—in which repetitions are only superimposed on the originally generated image when subjects recognize the repetition as previously studied—can handle these phenomena.

In Chapter 5, I combined REM.3 with Criss et al.'s (2011) output-interference version of REM. This hybrid REM model made an important, novel prediction: namely, that testing mixed-weak items before mixed-strong items would produce a list-strength effect in recognition. I tested this prediction in Chapter 5, but the results produced a statistically significant, negative list-strength effect (i.e., the results were opposite to the REM predictions). I tested a second REM prediction: namely, that testing mixed-strong items before mixed-weak items would produce a null list-strength effect. However, Experiment 2 produced a statistically significant, positive list-strength effect, again

inconsistent with REM's predictions. Finally, in Experiment 3, I verified that, under normal testing conditions, the stimuli used in Experiments 1 and 2 produce the standard null list-strength effect.

Combined, the experimental results contradict the REM.3/output-interference hybrid. However, they are consistent with the output-interference account of the list-strength effect proposed by Tulving and Hastie (1972; see also Wixted et al., 1997).

In the remainder of the present chapter, I discuss the theoretical implications of the results, both for REM.3 in particular and for accounts of the list-strength effect more generally. Although the experimental results are inconsistent with REM, I provide a framework based on mnemonic strategies and output interference to account for the findings. First, however, I address whether the list-strength effects observed in Experiments 1 and 2 stem from methodological artifacts.

6.2. Methodological Considerations

In Chapter 2, I reviewed the literature on the list-length effect. Although the list-length effect was long considered a hallmark of recognition (Gillund & Shiffrin, 1984; Ratcliff & Murdock, 1976; Strong, 1912), Dennis and Humphreys (2001; see also Murdock & Kahana, 1993b) suggested that extant list-length effects may have been artifactual. In particular, Dennis and Humphreys argued that four factors could have artificially produced or inflated previous list-length effects: differences in study-test lag, attentional lapse, rehearsal borrowing, and counterproductive context reinstatement. Debate still exists concerning the status of the list-length effect in recognition (Annis et al., 2015; Criss & Shiffrin, 2004a, 2004b; Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, 2011), and I will not attempt to resolve that issue here.

Nevertheless, given the similarities between list-length and list-strength manipulations, it is worthwhile to consider the degree to which Dennis and Humphreys' arguments apply to the present experiments.

Like list-length manipulations, list-strength manipulations encounter issues with study-test lag. As average test strength increases from the pure-weak list to the mixed list to the pure-strong list, the time required to present the study list increases. Given the 2-s presentation rate in the present experiments, the pure-weak list lasted for 176 s, the mixed list lasted for 260 s, and the pure-strong list lasted for 344 s. To address the study-test lag, I interpolated a distractor task between study and test, with the duration of the distractor task increasing as list length decreased. The result was that the time that elapsed between the presentation of the first primacy buffer and the end of the math task was equated across list types.

The math task equated the start of the study and test lists, but it did not fully address the issue of the study-test lag. In Experiments 1 and 2, test lists were blocked according to study position: In Experiment 1, second-half study items were tested in the first block of the test list and first-half study items were tested in the second block of the test list; in Experiment 2, first-half study items were tested in the first block of the test list and second-half study items were tested in the second block of the test list. For illustrative purposes, consider the study-test lag between second-half study items from the pure-weak and pure-strong lists in Experiment 1. In the pure-weak list, study of these items began 89 s after study-phase onset and ended 172 s after study-phase onset. In contrast, study of second-half pure-strong items began 173 s after study-phase onset and ended 340 s after study-phase onset. Assuming subjects spent an average of 1 s on each test probe, then the

midpoint of the first block of the test phase occurred 462 s after the start of the study-test cycle. As such, second-half pure-strong items had a recency advantage over second-half pure-weak items.

The issue of the study-test lag is only problematic if it artificially produced the negative list-strength effect in Experiment 1 and/or the positive list-strength effect in Experiment 2. For example, although second-half pure-strong items had a recency advantage over second-half pure-strong items, comparison of performance on these items is not relevant in assessing whether a list-strength effect occurred. Note that, because test lists were always equal in length across list types, we need only consider the lag between when the items were last studied and the start of the test list (i.e., calculating the lag between the last time the item was studied and its likely test position within the specified block will produce the same difference between conditions). Since the study phases were identical in Experiments 1 and 2, this allows simultaneous consideration of the study-test-lag concern between experiments. Mixed-strong items were always compared to pure-strong items from the first half of the study list, and mixed-weak items were always compared to pure-weak items from the second half of the study list.[10] For the strong-item comparison, then, the study-test lag was equivalent: Both mixed-strong items and first-half pure-strong items were studied between 5 and 172 s following study-phase onset. There was, however, a recency advantage for mixed-weak items over second-half pure-weak items. Second-half pure-weak items were studied between 89 and 172 s after study-

10. It does not make sense to consider the whole-list analysis because, by definition, that analysis confounded study and test position.

phase onset, and mixed-weak items were studied from 173 to 256 s after study-phase onset. Therefore, the study-test lag was shorter for mixed-weak items than pure-weak items, presumably giving them a mnemonic benefit.

The recency advantage for mixed-weak items over pure-weak items cannot explain the presence of a positive list-strength effect in Experiment 2. In that case, despite the recency advantage afforded mixed-weak items, pure-weak items were recognized at a higher rate. If anything, then, rather than inflating the list-strength effect, the weak-item study-test-lag confound in Experiment 2 may have underestimated it. The confound leaves open the possibility, however, that the negative list-strength effect in Experiment 1 partially stemmed from a recency advantage for mixed-weak items. I find this argument unpersuasive for three reasons: First, it is unclear why the mixed-weak items would benefit from the recency advantage in Experiment 1 but not Experiment 2. Second, the difference in study-test lag cannot explain why strong-item performance went in different directions between Experiments 1 and 2. Finally, although it is likely that time-based recency plays some role in memory, the argument in the present case relies heavily on the decay theory of forgetting (Thorndike, 1913). As reviewed in Chapter 1, there are reasons to be skeptical of decay as a mechanism of forgetting—generally, it is intervening experiences, rather than the mere passage of time, that causes forgetting (e.g., Jenkins & Dallenbach, 1924; see McGeoch, 1932a, for an early review).

Despite the above rationale, one might reasonably ask why I did not use an experimental design that was immune to study-test-lag criticisms. I used the design that I did for four reasons. First, Malmberg and Shiffrin (2005) showed that, in free recall, spaced strengthening with relatively short lags produced the largest, positive list-strength

effects. This type of strengthening cannot be used if study-test lag is controlled. If experimenters want to control study-test lag, they need to present all to-be-remembered items once in a random order, followed by blocks of strong-item repetitions. This results in a substantially longer lag between repetitions than I wanted to use, given Malmberg and Shiffrin's findings. If, for example, I had used that approach, and Experiments 1 and 2 had both yielded positive list-strength effects, a plausible interpretation of the results would have been that the lag between repetitions was too long for single-image storage.

The second reason I chose to ignore the study-test lag is that, in studies that have controlled this aspect of the mixed-pure paradigm, the same pattern of results is observed as when it is ignored. Osth et al. (2014) and Osth, Fox, et al. (2018) reported null list-strength effects when study-test lag was controlled, although it should be noted that none of their experiments included a pure-strong list.

The third reason to structure the study lists in the way that I did concerns another confound noted by Dennis and Humphreys (2001): rehearsal borrowing. When the null list-strength effect was first documented (Ratcliff et al., 1990), researchers wondered whether rehearsal borrowing was contributing to the result—that is, were subjects on mixed lists using strong-item presentations to rehearse weak-item presentations? If this occurred, it could mask the presence of a list-strength effect. The simplest way to avoid this concern is to present all of the strong items before any weak items are presented. In this way, subjects cannot rehearse weak items during strong-item presentations, since strong-item presentations end before any weak items have been studied.

The final reason I did not implement study-test-lag controls is that I wanted to remain as consistent as possible with traditional list-strength experiments. Although

Malmberg and Shiffrin (2005) equated the start of the study and test lists by increasing the duration of their distractor task for shorter lists, they did not consider study position. Indeed, in early list-strength experiments (Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1990, 1992, 1994; Yonelinas et al., 1992), study-test lag was ignored entirely; if a distractor task was interpolated between study and test, its duration was the same across list types.

Before concluding my discussion of study-test lag, it is worth emphasizing one final point: Had I controlled study-test lag and found positive list-strength effects, two methodological explanations would have become plausible. First, one could argue that the use of a long repetition lag rendered subjects unable to identify repeated items as previously studied. Second, controlling study-test lag would have made my design far more similar to the strong-interference paradigm than the mixed-pure paradigm. We already know that the strong-interference paradigm produces a positive list-strength effect (Norman, 1999, 2002), so finding one with a similar procedure would not have been surprising. In balance, then, the benefits of ignoring study-test lag outweighed the disadvantages.

I addressed rehearsal borrowing above, but what about attentional lapse? The argument for attentional lapse is that the greater attentional demands imposed by longer lists reduces subjects' focus, thereby decreasing test performance (Henmon, 1917; Underwood, 1978). Although attentional lapse is persuasive in the case of the list-length effect, its relevance to the list-strength effect is less clear. If we accept the premise that longer study lists reduce subjects' attention, then we would expect encoding to suffer for mixed-weak items and second-half pure-strong items, since these items were studied

following 86 study trials (two primacy buffers and 42 twice-presented targets). Yet, in Experiment 1, mixed-weak items were better recognized than second-half pure-weak items, which followed only 44 study items (two primacy buffers and 42 once-presented targets). From an attentional-lapse perspective, we would also expect second-half pure-strong items to be more poorly recognized than first-half pure-strong items. Numerically, there is support for this idea: d' was higher for first-half pure-strong items ($M = 1.277$, $SE = 0.163$) than second-half pure-strong items ($M = 1.106$, $SE = 0.137$), but this difference was not significant, $t(21) = 1.271$, $p = .218$, $d = 0.271$. So, there may have been a slight effect of attentional lapse on the pure-strong list. Yet, since first-half pure-strong items were compared to mixed-strong items, this attentional lapse cannot have contributed to the list-strength effect.

Dennis and Humphreys' (2001) final confound concerns context reinstatement. In list-length experiments, context reinstatement poses a problem in the retroactive design, wherein start-of-list items from the long list are tested but end-of-list items are not. Following the distractor task, if subjects try to reinstate study context, it is easier to reinstate end-of-study context than start-of-study context, since end-of-study context will be more similar to start-of-test context than start-of-study context. So, a strategy that appears intuitively reasonable is actually counterproductive. While it is conceivable that this affected performance in Experiment 2—where test lists began with start-of-study targets—it should have affected performance equally across lists. In other words, because test lists were blocked according to study block, context reinstatement is not a viable interpretation of the results.

6.3. Theoretical Considerations

In this section, I attempt to square the results from Experiments 1 and 2 with extant accounts of the list-strength effect. To foreshadow, no account is capable of explaining the findings. Therefore, I conclude by describing an account that appeals to output interference and subjects' mnemonic strategies.

6.3.1. Differentiation

Classic differentiation models include the differentiation version of SAM (Shiffrin et al., 1990), REM (Shiffrin & Steyvers, 1990; although REM.3 is an exception), and SLIM (McClelland & Chappell, 1998). In classic differentiation models, differentiation occurs automatically: When an earlier studied item is re-presented, the mnemonic trace generated on its first presentation is updated, rather than a new trace being formed. Classic differentiation accounts necessarily predict a negative list-strength effect in recognition, unless some feature of the experiment disrupts the differentiation process (e.g., Murnane & Shiffrin, 1991b; Sahakyan & Malmberg, 2018). Consequently, classic differentiation models cannot explain the positive list-strength effect observed in Experiment 2.

What about more realistic differentiation models? REM.3 (Shiffrin & Steyvers, 1997; Chapt. 4 of the present work) is a well-specified differentiation model wherein single-trace storage is probabilistic rather than automatic. However, REM.3 cannot explain the present findings (see Figures 5.1-5.4 in Chapter 5): Indeed, REM.3 predicts the opposite pattern—that is, a positive list-strength effect in Experiment 1 and a negative list-strength effect in Experiment 2. I have tested several parameter values with REM.3, but none have been able to predict the present findings. I will have more to say about

REM.3 below; for now, it suffices to say that the present results are inconsistent with both classic differentiation models and more complex versions.

6.3.2. Criterion placement

According to the criterion-placement account of the list-strength effect (Hirshman, 1995), recognition does not produce a list-strength effect because subjects calibrate their test criteria based on the average strength of the study list. So, response criteria become increasingly conservative from the pure-weak list to the mixed list to the pure-strong list. The criterion-placement account, then, cannot explain the contrasting results from Experiments 1 and 2.

Rather than setting response criteria according to study-list strength, is it possible that subjects set their response criterion according to the perceived difficulty of the first few test items? Using mixed-strength lists, Verde and Rotello (2007) demonstrated that this is the case—that is, subjects use a more conservative response criterion when strong items are tested first and a more liberal response criterion when weak items are tested first. What implications does this have for Experiments 1 and 2? The mixed list in Experiment 1 began with weak items whereas the mixed list in Experiment 2 began with strong items. This implies that subjects would have used a more liberal response criterion in Experiment 1 than in Experiment 2. Yet, the data do not support this contention: The proportion of probes identified as "old" (i.e., hits and false alarms) was actually lower in Experiment 1 (M = .490, SE = .025) than Experiment 2 (M = .542, SE = .033), suggesting that a slightly more liberal response criterion was used when the first half of the test list was more difficult.

In Chapter 3, I described why I believe differentiation accounts provide a more parsimonious explanation of the list-strength effect and strength-based mirror effect than criterion-placement accounts. Briefly, more conservative responding is a natural outcome of differentiation, and thus does not require an appeal to strategic deployment of response criteria on the part of subjects. In any case, the present results are inconsistent with the criterion-placement account, inasmuch as such an account cannot predict a negative list-strength effect in Experiment 1 and a positive list-strength effect in Experiment 2.

6.3.3. The continuous-memory version of TODAM

Murdock and Kahana (1993a) proposed the continuous-memory version of TODAM in response to Shiffrin et al.'s (1990) claim that TODAM could not predict a null list-strength effect in recognition. Murdock and Kahana showed that, when simulated subjects' memories were initialized with pre-experimental noise, interference from the pre-experimental memories overwhelmed any effect of strong-item interference on the mixed list, thereby negating the list-strength effect predicted by simpler versions of TODAM.

Although the assumption—present in many memory models—that subjects' memories are blank at the beginning of a study list is an obvious simplification, it is unclear whether Murdock and Kahana's (1993a) model provides an adequate solution. Pre-experimental memories are presumably isolated from experimental memories by virtue of the fact that they occurred in different contexts. Most people do not routinely participate in memory experiments, and therefore one would expect that the cue corresponding to a current experiment would be fairly diagnostic. In any case, the

continuous-memory version of TODAM cannot explain the negative list-strength effect from Experiment 1 and the positive list-strength effect from Experiment 2.

6.3.4. BCDMEM

Dennis and Humphreys' (2001) BCDMEM makes the provocative claim that recognition is unaffected by interitem interference. Instead, BCDMEM assumes that all forgetting in a recognition test is either a product of poor encoding or interference from previous contexts in which the probes have been encountered. Because BCDMEM excludes interitem interference, it easily predicts a null list-strength effect in recognition.

It should also be noted that BCDMEM can account for output interference—a phenomenon on which the present experiments were based. Although interitem interference does not affect memory, BCDMEM includes context drift (Estes, 1955a, 1955b; Mensink & Raaijmakers, 1988, 1989), such that context becomes increasingly different as time elapses. So, in BCDMEM, discrimination declines as subjects move through a recognition test list because, with every test probe, the study context moves further into the past, thus becoming less similar to present context (see Osth & Dennis, 2015; Osth, Jansson, et al., 2018). However, although this mechanism is consistent with the positive list-strength effect in Experiment 2 when test position was ignored, it cannot explain the presence of a list-strength effect when performance was conditionalized on list half. Notice that, by conditionalizing performance on list half, comparisons between the mixed and pure lists involve items for which study context is equally distant in the past. Therefore, BCDMEM cannot explain the present findings.

6.3.5. Insufficient strengthening

Norman (1999, 2002) argued that experimenters' emphasis on keeping strong-item performance below ceiling was obscuring the list-strength effect. He therefore strengthened strong items to ceiling and tested only weak items. As reviewed earlier, this generally yields a positive list-strength effect (Diana & Reder, 2005; Norman, 1999, 2002; Norman et al., 2008; but see Osth, Fox, et al., 2018).

The present experiments do not directly assess the merit of the insufficient-strengthening account, inasmuch as I did not manipulate the degree of strengthening (i.e., all experiments used a 2:1 strengthening ratio). Yet, what is clear from the present findings is that, although strengthening to ceiling is sufficient to produce a positive list-strength effect, it is not necessary. In Experiment 2, I found a positive list-strength effect when strong items were only presented twice; this is a much less potent strength manipulation than typical list-strength experiments have used, where a 3:1 or 4:1 strength ratio is far more common (Malmberg & Shiffrin, 2005; Ratcliff et al., 1990).

6.3.6. Dual-process accounts

Dual-process accounts of the list-strength effect (Diana & Reder, 2003; Norman, 1999, 2002; Norman & O'Reilly, 2003; Reder et al., 2000) contend that the positive list-strength effect is an inherently recollection-based phenomenon, and thus it can only be detected in recognition when contributions from familiarity are minimized. I did not include the remember/know procedure (Tulving, 1985) in my experiments, and thus I cannot speculate on the degree to which familiarity and recollection contributed to the results. Yet, the fact that testing strong items before weak items produced a positive list-strength effect is broadly consistent with the notion that the list-strength effect is recall-based. In free recall, subjects generally output strong items earlier than weak items (Wike

& Wike, 1970; Wixted et al., 1997). I will have more to say about this in the following Section.

6.4. A Strategy-Disruption Account of the List-Strength Effect

Upon first documenting the list-strength effect in free recall, Tulving and Hastie (1972) described it as a puzzling finding, saying that it "appears to violate both common sense and current theoretical notions" (p. 302). The present results seem to return us to this early position, given that no extant theoretical account is capable of predicting the pattern. Although an overarching explanation of list-strength effects in recognition remains elusive, the free-recall list-strength effect is well understood. It therefore may be informative to consider the dominant account of the list-strength effect in free recall when searching for a recognition explanation.

6.4.1. The list-strength effect in free recall

As described in Chapter 3, positive list-strength effects are generally found in free recall (Malmberg & Shiffrin, 2005; Tulving & Hastie, 1972). The free-recall list-strength effect is often explained in terms of sampling and recovery probabilities. Here, I use the SAM model (Raaijmakers & Shiffrin, 1980, 1981) to illustrate the list-strength effect in free recall. Note that, although REM has replaced the recognition version of SAM (Gillund & Shiffrin, 1984), SAM and REM are nearly indistinguishable in free recall (see, e.g., Malmberg & Shiffrin, 2005). The only difference has to do with complexity—for example, SAM represents items as single numbers, whereas REM uses feature vectors.

In SAM, studied items are first placed in the short-term store, where they are rehearsed along with other recently studied items. When the short-term store reaches capacity, a random item is transferred to the long-term store. At test, images are sampled

with either the general context cue or with a context-plus-item cue, where the item is a previously recalled target. The probability of sampling a given image is proportional to the associative strength between the cue and image, weighed against the associative strength between the cue and all images. The associative strength between an item and context increases linearly while the item is rehearsed in the short-term store, and the associative strength between two items increases linearly while the items are rehearsed together in the short-term store. Because strong items are presented multiple times, the context cue will be more strongly associated to strong images than weak images. Similarly, assuming a spaced strengthening technique was used, strong items will be associatively connected to a greater number of images than weak items.

When a cue samples an image, it must be recovered. The sampled image can only be output if recovery is successful. The probability of successful recovery increases as the associative strength between the cue and image increases.

For the sake of completeness, I next briefly describe SAM's parameters as presented by Raaijmakers and Shiffrin (1980). There are four strength parameters: *a*, *b*, *c*, and *d*. These are, respectively, the context-to-image strength parameter, the image-to-image strength parameter, the self strength parameter, and the residual strength parameter. The context-to-image strength parameter represents the rate at which strength between the context cue and image increases while the image is in the short-term store, the image-to-image strength parameter represents the rate at which the associative strength between two images increases while they occupy the short-term store together, the self strength parameter represents the degree to which an image's associative strength to itself increases while the image is in the short-term store, and the residual strength parameter

represents the associative strength between images that did not occupy the short-term store together. Of these, the self strength parameter is probably the most obscure: Note, however, that because sampling with a cue in SAM occurs with replacement, this leaves open the possibility that using a recently recalled item as a cue will result in that cue sampling its own image.

In SAM, sampling an image with a cue causes the associative strength between the cue and image to increase, a process Raaijmakers and Shiffrin (1980) termed incrementing. SAM has three incrementing parameters: e, f, and g. These are, respectively, the context-to-image incrementing parameter, the image-to-image incrementing parameter, and the self incrementing parameter. The context-to-image incrementing parameter represents the degree to which the associative strength between context and an image increases when an image is retrieved, the image-to-image incrementing parameter represents the degree to which the associative strength between two images increases when one image is used to retrieve another, and the self incrementing parameter represents the degree to which the associative strength between an image and itself increases when a cue retrieves its own image.

Raaijmakers and Shiffrin (1980) included two sampling parameters: k_{max} and l_{max}. The k_{max} parameter represents the number of failed sampling attempts that simulated subjects make before terminating the search. In SAM, because all cues have some degree of associative strength to all images in memory, if search is permitted to continue indefinitely, then everything in memory will eventually be sampled (although not necessarily recovered). The l_{max} parameter represents the number of sampling attempts

that can be made with a given item cue before the simulated subject abandons the cue and returns to the context cue.

The last two parameters in SAM are *t* and *r*. The *t* parameter represents the number of time units for which each item is presented. Note that the values of the *a*, *b*, and *c* parameters reflect the amount of associative strengthening that occurs per unit of study time.[11] The *r* parameter represents the capacity of the short-term store. As items are studied, they are added one by one to the short-term store. When *r* images are being rehearsed in the short-term store, studying a new item causes a random item in the short-term store to be dropped. This discontinues the accumulation of associative strength between the dropped image and context, and between the dropped image and the other images in the short-term store.

Why does SAM predict a positive list-strength effect? I begin by considering why strong items have a mnemonic advantage over weak items on pure lists. Consider that, although the average associative strength between the context cue and a given image is higher on the pure-strong list than the pure-weak list, the probability of sampling any given image is, on average, the same. Let *S* denote the number of times strong items are presented, *W* denote the number of times weak items are presented, and n denote the total number of items on the study list. If we make the simplifying assumption that all items spend the same amount of time in the short-term store (i.e., that all items are encoded

---

11. Like the t_1 and t_2 parameters from REM, the *t* parameter in SAM is superfluous. To remove it, one would simply have to multiply *a*, *b*, and *c* by *t*.

equally well), then the probability of sampling a pure-weak item's image—denoted i_{PW}—with the general context cue is:

$$p(i_{PW}) = W/(Wn)$$

Similarly, the probability of sampling a pure-strong item's image—denoted i_{PS}—is:

$$p(i_{PS}) = S/(Sn)$$

Therefore, $p(i_{PW}) = p(i_{PS})$. As a result, the advantage of pure-strong items over pure-weak items is a function of the recovery process, not the sampling process. On pure-strong lists, the probability of successfully outputting an item, given that it has been sampled, is higher than on a pure-weak list.

On the mixed list, the strong-item advantage is a function of both sampling and recovery. Using the same simplifying assumptions used for the pure lists, the probability of the general context cue sampling a mixed-weak item's image—denoted i_{MW}—is:

$$p(i_{MW}) = W/(W(n/2)+S(n/2))$$

and the probability of the general context cue sampling a mixed-strong item's image—denoted i_{MS}—is:

$$p(i_{MS}) = S/(W(n/2)+S(n/2))$$

Since $S > W$, $p(i_{MS}) > p(i_{MW})$, and $p(i_{MW}) < p(i_{PW}) = p(i_{PS}) < p(i_{MS})$. So, in SAM, the list-strength effect is an occlusion phenomenon: the mixed-strong items block access to the mixed-weak items, thereby enhancing strong-item memory on mixed lists and depressing weak-item memory on mixed lists, relative to pure lists.

I confirmed the above predictions by simulating the mixed-pure paradigm in SAM. Ten thousand simulated subjects performed free recall of 28-item lists. On the pure-weak list, all items were presented once; on the pure-strong list, all items were

presented twice; and on the mixed list, half of the items were presented once and half were presented twice.

Simulations used the following parameter values, consistent with those used in most of Raaijmakers and Shiffrin's (1980, 1981) simulations: $t = 2$, $a = 0.1$, $b = 0.1$, $c = 0.1$, $d = 0.02$, e = 0.7, $f = 0.7$, $g = 0.7$, $r = 4$, $k_{max} = 30$, and $l_{max} = 3$. The simulations produced a positive list-strength effect: The proportion of mixed-strong items recalled ($M = .515$) was higher than the proportion of pure-strong items recalled ($M = .442$), but the proportion of pure-weak items recalled ($M = .301$) was higher than the proportion of mixed-weak items recalled ($M = .232$). The strong-to-weak ratios were 2.220 and 1.468 for mixed and pure lists, respectively, yielding a positive list-strength effect ($R_r = 1.512$).

SAM provides a straightforward account of the free-recall list-strength effect: Because of the sampling bias afforded mixed-strong items, they tend to be output earlier in the test phase than mixed-weak items or pure-strong items. At the same time, mixed-weak items tend to be output later in the test phase than other items. To verify this, I made a slight modification to Raaijmakers and Shiffrin's (1980) original SAM algorithm. In SAM, one way to operationalize output latency is by the total number of sampling attempts made prior to an image's recovery. If we assume that the number of sampling attempts is relatively uniform across time, then this provides a useful measure of response latency. I therefore eliminated the stopping parameter, $k_{max;}$ instead, simulated subjects made sampling attempts until all items were recalled or until a maximum number of sampling attempts were made (i.e., until a specified amount of time had elapsed, where time is operationalized in terms of the number of sampling attempts since test-phase onset). I will term this new parameter t_{max}. Although k_{max} and t_{max} appear similar, there are

two important differences: First, the t_{max} counter is incremented even when a sampling attempt results in the recovery of a new item; conversely, the k_{max} counter is only incremented when a sampling attempt fails. Second, the t_{max} parameter is set to an unrealistically high value, at least for comparison with human subjects. The point is to measure the boundaries of the SAM model, not to make comparisons to real data. Notice that, given a sufficiently large value for t_{max}, there should be no missing data in the simulation results—that is, all subjects should recall at least one mixed-weak item, even if doing so requires an unrealistically large number of sampling attempts.

This second simulation was identical to the previous simulation, save that k_{max} was replaced with t_{max}, which was set to 1000. This meant that a simulated subject's test phase ended when 1000 sampling attempts had been made or when all 28 targets had been recovered. The results are displayed in Table 6.1. Results were consistent with the sampling interpretation described above: The average output latency was earlier for mixed-strong items (M = 258.60, MDN = 39) than pure-strong items (M = 282.95, MDN = 55), but the average output latency was earlier for pure-weak items (M = 516.18, MDN = 551) than mixed-weak items (M = 546.16, MDN = 617.5).

So, in SAM, two related mechanisms account for the free-recall list-strength effect: First, mixed-strong items have a privileged output position relative to pure-strong items—that is, strong items are recalled earlier in the test phase in mixed lists than pure lists. Second, because of their later output position, mixed-weak items are subjected to a greater degree of output interference than pure-weak items. In SAM, incrementing produces output interference, since increasing the associative strength between a cue and

an image reduces the probability of sampling a different image with that cue on subsequent sampling attempts.

6.4.2. A strategy-disruption account of the list-strength effect

Free recall reliably produces a positive list-strength effect (Fritzen, 1975; Hastie, 1975; Malmberg & Shiffrin, 2005; Mueller & Brown, 1977; Sahakyan et al., 2014; Tulving & Hastie, 1972; Wixted et al., 1997). Although Tulving and Hastie considered it a puzzling phenomenon when first documented, it has since been effectively explained by sampling/recovery models of memory such as SAM. Following study of a mixed list, subjects tend to output strong items early and weak items late (Wike & Wike, 1970; Wixted et al., 1997). This gives mixed-strong items a sampling advantage over pure-strong items, and mixed-weak items an output-interference disadvantage compared to pure-weak items.

Other than free recall, only one paradigm produces a positive list-strength effect: *A*-*B*/*A*-*C* cued recall (Bäuml, 1997; Verde, 2009). By *A*-*B*/*A*-*C* cued recall, I mean a cued-recall experiment in which each cue subsumes several targets (e.g., “walnut-scissors”, “walnut-envelope”, “walnut-diamond”, “highway-icicle”, “highway-farm”, etc.). This can be contrasted with *A*-*B*/*C*-*D* cued recall, in which each cue subsumes a single target (e.g., “triangle-chicken”, “sand-helicopter”, “string-island”, etc.). Importantly, although *A*-*B*/*A*-*C* cued recall produces a positive list-strength effect (Bäuml, 1997; Verde, 2009), *A*-*B*/*C*-*D* cued recall produces a null list-strength effect (Wilson & Criss, 2017).

The explanation for the *A*-*B*/*A*-*C* cued-recall list-strength effect is identical to that used for free recall: Compared to pure cues (i.e., cues subsuming all equally strengthened

targets), mixed cues yield earlier output of strong targets than weak targets. In SAM terms, the *A* cue simply takes the place of the general context cue used in free recall.

Paradigms that produce a null or negative list-strength effect include *A*-*B*/*C*-*D* cued recall (Wilson & Criss, 2017), two-alternative forced-choice recognition (Yonelinas et al., 1992), and old/new recognition (Ratcliff et al., 1990). An important difference between the paradigms producing a positive list-strength effect and the paradigms producing a null or negative list-strength effect is that, in the former, subjects control output order, but, in the latter, experimenters control output order. When left up to experimenters, output order is generally randomized—this means that, in mixed-strength lists, weak and strong items are equally likely to appear early in the test phase. In contrast, when subjects control output order of mixed lists, strong items appear earlier than weak items (Wike & Wike, 1970; Wixted et al., 1997).

To what degree does changing the output order that subjects would have otherwise used matter? A phenomenon known as part-list cuing suggests that interfering with a subject's mnemonic strategy can have detrimental effects. In a part-list-cuing experiment, subjects study a list of words and are then tested using free recall. In one condition, the experimenter provides subjects with a subset of the items from the study list, allegedly to aid their recall. In the control condition, subjects complete the free-recall test without cues. The standard finding is that the provision of cues harms memory, with the proportion of items recalled higher in the uncued condition compared to the cued condition (Aslan, Bäuml, & Grundgeiger, 2007; Andrés & Howard, 2011; Basden, Basden, Church, & Beaupre, 1991; J. Brown, 1968; Peynırcıoğlu, 1989; Slamecka, 1968;

for reviews, see Bovee, Fitz, Yehl, Parrott, & Kelley, 2009; Nickerson, 1984; Roediger & Neely, 1982).

One dominant account attributes part-list cuing to strategy disruption (Basden, Basden, & Stephens, 2002; Basden & Basden, 1995; Basden, Basden, & Galloway, 1977; Sloman, Bower, & Rohrer, 1991; Serra & Nairne, 2000). According to the strategy-disruption account, providing subjects with cues forces them to modify the mnemonic strategy they otherwise would have used. This disruption leads them to adopt a suboptimal mnemonic strategy, resulting in poorer recall of the remaining items. Strategy disruption is widely accepted as a cause of part-list cuing: Even accounts that rely on other mechanisms tend to incorporate a role for strategy disruption (Aslan & Bäuml, 2007; Bäuml & Aslan, 2006; Cole, Reysen, & Kelley, 2013; Crescentini, Shallice, Del Missier, & Macaluso, 2010; Kelley, Parasiuk, Salgado-Benz, & Crocco, 2016).

How might strategy disruption affect recognition results from the mixed-pure paradigm? From a strategy-disruption account, a subject's mnemonic strategy should be affected on both pure and mixed lists, by virtue of the fact that both types of list randomize test order. Yet, there is reason to believe that the degree of disruption will be greater on mixed lists than pure lists. There are regularities in how subjects recall items from a pure list, such as primacy and recency (Rundus & Atkinson, 1970); by and large, however, mnemonic strategies from pure lists will be idiosyncratic across subjects. In contrast, although idiosyncrasies will certainly exist from subject to subject, the structure of the mixed list is such that we know what strategy, in broad terms, subjects will employ if left to their own devices: They will output strong memories prior to weak memories

(Bousfield & Barclay, 1950; Bousfield & Sedgewick, 1944; Bousfield et al., 1954, 1956, 1958; W. Brown, 1915; Wike & Wike, 1970; Wixted et al., 1997).

So, according to the strategy-disruption account of the list-strength effect, the list-strength effect should be positive when the memory test is consistent with subjects' mnemonic strategies, but null or negative when the memory test interferes with subjects' mnemonic strategies. Experiments 1, 2, and 3 provide direct support for this account: There was a positive list-strength effect in Experiment 2 in which, crucially, strong items were tested before weak items on the mixed list. Conversely, there was a negative list-strength effect in Experiment 1, where weak items were tested before strong items on the mixed list. Finally, Experiment 3 produced a null list-strength effect (i.e., between the list-strength effects from Experiments 1 and 2) when test probes were randomized without regard to item strength. In strategy-disruption terms, Experiment 1 went out of its way to disrupt subjects mnemonic strategies, Experiment 2 went out of its way to remain consistent with subjects' mnemonic strategies, and Experiment 3 ignored subjects' mnemonic strategies altogether.

Crucially, the strategy-disruption account I have proposed here is also consistent with discrepancies in the literature. In free recall and *A*-*B*/*A*-*C* cued recall, subjects determine output order, and thus are free to utilize any mnemonic strategy they believe will be most effective. Both of these paradigms produce positive list-strength effects (Bäuml, 1997; Tulving & Hastie, 1972). In contrast, *A*-*B*/*C*-*D* cued recall and recognition randomize test order, thereby interfering with subjects' mnemonic strategies. Both of these paradigms produce null or negative list-strength effects (Ratcliff et al., 1990; Wilson & Criss, 2017). Consequently, the strategy-disruption account provides an

explanation of the present data specifically, as well as an explanation of data in the literature more generally.

6.5. Final Thoughts

Two goals guided the present study: First, I sought to test the viability of REM.3, which provides for a more realistic differentiation process than REM.1. Second, I sought to test a novel prediction of REM.3. REM.3 was capable of accounting for several phenomena for which simpler versions of REM cannot: the spacing effect, negative list-strength effects with massed strengthening but slightly positive list-strength effects with spaced strengthening, and results from the strong-interference paradigm. However, REM.3's predictions regarding output interference in the mixed-pure paradigm were not supported. Instead, I proposed a strategy-disruption account of the list-strength effect, according to which subjects' mnemonic strategies are responsible for positive list-strength effects. Future work should further investigate the viability of more complex REM versions. In addition, the present findings highlight the need to consider mnemonic strategies when designing experiments.

References

Abernethy, E. M. (1940). The effect of changed environmental conditions upon the results of college examinations. *Journal of Psychology, 10,* 293-301. doi:10.1080/00223980.1940.9917005

Allen, L. R., & Garton, R. F. (1968). The influence of word-knowledge on the word-frequency effect in recognition memory. *Psychonomic Science, 10,* 401-402. doi:10.3758/BF03331581

Allen, L. R., & Garton, R. F. (1970). Manipulation of study trials in recognition memory. *Perception & Psychophysics, 7,* 215-217. doi:10.3758/BF03209362

Anderson, K. J., & Revelle, W. (1994). Impulsivity and time of day: Is rate of change in arousal a function of impulsivity? *Journal of Personality and Social Psychology, 67,* 334-344. doi:10.1037/0022-3514.67.2.334

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49,* 415-445. doi:10.1016/j.jml.2003.08.006

Anderson, M. C. (2007). Inhibition: Manifestation in long-term memory. In H. L. Roediger, Y. Dudai, & S. M. Fitzpatrick (Eds.), Science of Memory: Concepts (pp. 295-300). Oxford, England: Oxford University Press.

Anderson, M. C., & Bell, T. (2001). Forgetting our facts: The role of inhibitory processes in the loss of propositional knowledge. *Journal of Experimental Psychology: General, 130,* 544-570. doi:10.1037/0096-3445.130.3.544

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 20, 1063-1087. doi:10.1037/0278-7393.20.5.1063

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature, 410,* 366-369. doi:10.1038/35066572

Anderson, M. C., & Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in Cognitive Sciences, 18,* 279-292. doi:10.1016/j.tics.2014.03.002

Anderson, M. C., & Levy, B. (2002). Repression can (and should) be studied empirically. *Trends in Cognitive Sciences, 6,* 502-503. doi:10.1016/S1364-6613(02)02025-9

Anderson, M. C., & Levy, B. J. (2007). Theoretical issues in inhibition: Insights from research on human memory. In D. S. Gorfein & C. M. MacLeod (Eds.), Inhibition in cognition (pp. 81-102). Washington, DC: American Psychological Association. doi:10.1037/11587-005

Anderson, M. C., & Levy, B. J. (2009). Suppressing unwanted memories. *Current Directions in Psychological Science, 18,* 189-194. doi:10.1111/j.1467-8721.2009.01634.x

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review, 102,* 68-100. doi:10.1037/0033-295X.102.1.68

Andrés, P., & Howard, C. E. (2011). Part set cuing in older adults: Further evidence of intact forgetting in aging. *Aging, Neuropsychology, and Cognition, 18,* 385-395. doi:10.1080/13825585.2010.542892

Annis, J., Lenes, J. G., Westfall, H. A., Criss, A. H., & Malmberg, K. J. (2015). The list-length effect does not discriminate between models of recognition memory. *Journal of Memory and Language, 85,* 27-41. doi:10.1016/j.jml.2015.06.001

Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (2013). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1365-1376. doi:10.1037/a0032188

Arbuckle, T. Y. (1967). Differential retention of individual paired associates within an RTT "learning" trial. *Journal of Experimental Psychology, 74,* 443-451. doi:10.1037/h0024712

Aslan, A., & Bäuml, K.-H. T. (2007). Part-list cuing with and without item-specific probes: The role of encoding. *Psychonomic Bulletin & Review, 14,* 489-494. doi:10.3758/BF03194095

Aslan, A., Bäuml, K.-H. T., & Grundgeiger, T. (2007). The role of inhibitory processes in part-list cuing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 335-341. doi:10.1037/0278-7393.33.2.335

Atkinson, R. C., & Joula, J. F. (1973). Factors influencing speed and accuracy of word recognition. In S. Kornblum (Ed.), Attention and performance (vol. IV, pp. 583-612). San Diego, CA: Academic Press.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology: I. Learning, memory and thinking (pp. 243-293). Oxford, England: W. H. Freeman.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), The psychology of learning and motivation (vol. 2, pp. 89-195). Oxford, England: Academic Press. doi:10.1016/S0079-7421(08)60422-3

Aue, W. R., Criss, A. H., & Fischetti, N. W. (2012). Associative information in memory: Evidence from cued recall. *Journal of Memory and Language, 66,* 109-122. doi:10.1016/j.jml.2011.08.002

Aue, W. R., Criss, A. H., & Prince, M. A. (2015). Dynamic memory searches: Selective output interference for the memory of facts. *Psychonomic Bulletin & Review, 22,* 1798-1806. doi:10.3758/s13423-015-0840-5

Aziz, W., Wang, W., Kesaf, S., Mohamed, A. A., Fukazawa, Y., & Shigemoto, R. (2014). Distinct kinetics of synaptic structural plasticity, memory formation, and memory decay in massed and spaced learning. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 111,* E194-E202. doi:10.1073/pnas.1303317110

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1-29. doi:10.1146/annurev-psych-120710-100422

Baddeley, A. D. (1982). Domains of recollection. *Psychological Review, 89,* 708-729. doi:10.1037/0033-295X.89.6.708

Badham, S. P., Poirier, M., Gandhi, N., Hadjivassiliou, A., & Maylor, E. A. (2016). Aging and memory as discrimination: Influences of encoding specificity, cue overload, and prior knowledge. *Psychology and Aging, 31,* 758-770. doi:10.1037/pag0000126

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4,* 316-321. doi:10.1111/j.1467-9280.1993.tb00571.x

Basden, B. H., Basden, D. R., Church, B. A., & Beaupre, P. (1991). Setting boundary conditions on the part-set cuing effect. *Bulletin of the Psychonomic Society, 29,* 213-216. doi:10.3758/BF03342681

Basden, B. H., Basden, D. R., & Stephens, J. P. (2002). Part-set cuing of order information in recall tests. *Journal of Memory and Language, 47,* 517-529. doi:10.1016/S0749-596X(02)00016-5

Basden, D. R., & Basden, B. H. (1995). Some tests of the strategy disruption interpretation of part-list cuing inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1656-1669. doi:10.1037/0278-7393.21.6.1656

Basden, D. R., Basden, B. H., & Galloway, B. C. (1977). Inhibition with part-list cuing: Some tests of the item strength hypothesis. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 100-108. doi:10.1037/0278-7393.3.1.100

Bäuml, K.-H. T. (1997). The list-strength effect: Strength-dependent competition or suppression? *Psychonomic Bulletin & Review, 4,* 260-264. doi:10.3758/BF03209403

Bäuml, K.-H. T., & Aslan, A. (2006). Part-list cuing can be transient and lasting: The role of encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 33-43. doi:10.1037/0278-7393.32.1.33

Bäuml, K.-H. T., Pastötter, B., & Hanslmayr, S. (2010). Binding processes: Neurodynamics and functional role in memory and action. *Neuroscience & Biobehavioral Reviews, 34,* 1047-1054. doi:10.1016/j.neubiorev.2009.04.005

Bäuml, K.-H. T., Zellner, M., & Vilimek, R. (2005). When remembering causes forgetting: Retrieval-induced forgetting as recovery failure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1221-1234. doi:10.1037/0278-7393.31.6.1221

Beaman, C. P. (2006). The relationship between absolute and proportion scores of serial order memory: Simulation predictions and empirical data. *Psychonomic Bulletin & Review, 13,* 92-98. doi:10.3758/BF03193818

Bercovitz, K. E., Bell, M. C., Simone, P. M., & Wiseheart, M. (2017). The spacing effect in older and younger adults: Does context matter? *Aging, Neuropsychology, and Cognition, 24,* 703-716. doi:10.1080/13825585.2016.1251552

Beth, E. H., Budson, A. E., Waring, J. D., & Ally, B. A. (2009). Response bias for picture recognition in patients with Alzheimer's disease. *Cognitive and Behavioral Neurology, 22,* 229-235. doi:10.1097/WNN.0b013e3181b7f3b1

Bilodeau, I. M., & Schlosberg, H. (1951). Similarity in stimulating conditions as a variable in retroactive inhibition. *Journal of Experimental Psychology, 41,* 199-204. doi:10.1037/h0056809

Bjork, R. A. (2007). Inhibition: An essential and contentious concept. In H. L. Roediger, Y. Dudai, & S. M. Fitzpatrick (Eds.), Science of memory: Concepts (pp. 307-313). Oxford, England: Oxford University Press.

Bjork, R. A., & Richardson-Klavehn, A. (1989). On the puzzling relationship between environmental context and human memory. In C. Izawa (Ed.), Current issues in cognitive processes: The Tulane Flowerree Symposium on Cognition (pp. 313-344). Hillsdale, NJ: Erlbaum.

Bousfield, W. A., & Barclay, W. D. (1950). The relationship between order and frequency of occurrence of restricted associative responses. *Journal of Experimental Psychology, 40,* 643-647. doi:10.1037/h0059019

Bousfield, W. A., Cohen, B. H., & Silva, J. G. (1956). The extension of Marbe's law to the recall of stimulus-words. *American Journal of Psychology, 69,* 429-433. doi:10.2307/1419046

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 30,* 149-165. doi:10.1080/00221309.1944.10544467

Bousfield, W. A., Sedgewick, C. H., & Cohen, B. H. (1954). Certain temporal characteristics of the recall of verbal associates. *American Journal of Psychology, 67,* 111-118. doi:10.2307/1418075

Bousfield, W. A., Whitmarsh, G. A., & Esterson, J. (1958). Serial position effects and the Marbe effect in the free recall of meaningful words. *Journal of General Psychology, 59,* 255-262. doi:10.1080/00221309.1958.9710194

Bovee, J. C., Fitz, C., Yehl, G., Parrott, S., & Kelley, M. R. (2009). Applied part-set cuing. In M. R. Kelley (Ed.), Applied memory (pp. 73-87). Hauppauge, NY: Nova Science Publishers.

Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in recognition memory. *Memory & Cognition, 11,* 307-315. doi:10.3758/BF03196977

Bowyer, P. A., Humphreys, M. S., & Revelle, W. (1983). Arousal and recognition memory: The effects of impulsivity, caffeine and time on task. *Personality and Individual Differences, 4,* 41-49. doi:10.1016/0191-8869(83)90051-X

Brandt, M. (2007). Bridging the gap between measurement models and theories of human memory. *Zeitschrift für Psychologie, 215,* 72-85. doi:10.1027/0044-3409.215.1.72

Brandt, M., Zaiser, A.-K., & Schnuerch, M. (2019). Homogeneity of item material boosts the list length effect in recognition memory: A global matching perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45,* 834-850. doi:10.1037/xlm0000594

Braun, K., & Rubin, D. C. (1998). The spacing effect depends on an encoding deficit, retrieval, and time in working memory: Evidence from once-presented words. *Memory, 6,* 37-65. doi:10.1080/741941599

Brown, C. P. (1976). The role of short-term retention in passive avoidance by young chicks. *Behavioral Biology, 18,* 301-305. doi:10.1016/S0091-6773(76)92246-X

Brown, J. (1954). The nature of set-to-learn and of intra-material interference in immediate memory. *Quarterly Journal of Experimental Psychology, 6,* 141-148. doi:10.1080/17470215408416659

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10,* 12-21. doi:10.1080/17470215808416249

Brown, J. (1968). Reciprocal facilitation and impairment of free recall. *Psychonomic Science, 10,* 41-42. doi:10.3758/BF03331397

Brown, R., & McNeill, D. (1966). The 'tip of the tongue' phenomenon. *Journal of Verbal Learning & Verbal Behavior, 5,* 325-337. doi:10.1016/S0022-5371(66)80040-3

Brown, W. (1915). Incidental memory in a group of persons. *Psychological Review, 22,* 81-85. doi:10.1037/h0073959

Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory & Cognition, 37,* 807-818. doi:10.3758/MC.37.6.807

Bulevich, J. B., Roediger, H. L., Balota, D. E., & Butler, A. C. (2006). Failures to find suppression of episodic memories in the think/no-think paradigm. *Memory & Cognition, 34,* 1569-1577. doi:10.3758/BF03195920

Buratto, L. G., & Lamberts, K. (2008). List strength effect without list length effect in recognition memory. *Quarterly Journal of Experimental Psychology, 61,* 218-226. doi:10.1080/17470210701566713

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review, 106*, 551-581. doi:10.1037/0033-295X.106.3.551

Burri, C. (1931). The influence of an audience upon recall. *Journal of Educational Psychology, 22,* 683-690. doi:10.1037/h0070961

Calfee, R. C., & Atkinson, R. C. (1965). Paired-associate models and the effects of list length. *Journal of Mathematical Psychology, 2,* 254-265. doi:10.1016/0022-2496(65)90004-0

Camp, G., Pecher, D., & Schmidt, H. G. (2007). No retrieval-induced forgetting using item-specific independent cues: Evidence against a general inhibitory account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 950-958. doi:10.1037/0278-7393.33.5.950

Capaldi, E. J., & Neath, I. (1995). Remembering and forgetting as context discrimination. *Learning & Memory, 2,* 107-132. doi:10.1101/lm.2.3-4.107

Carew, T. J., Pinsker, H. M., & Kandel, E. R. (1972). Long-term habituation of a defensive withdrawal reflex in aplysia. *Science, 175,* 451-454. doi:10.1126/science.175.4020.451

Carriere, J. S. A., Seli, P., & Smilek, D. (2013). Wandering in both mind and body: Individual differences in mind wandering and inattention predict fidgeting. *Canadian Journal of Experimental Psychology, 67,* 19-31. doi:10.1037/a0031438

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49,* 231-248. doi:10.1016/S0749-596X(03)00061-5

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,* 354-380. doi:10.1037/0033-2909.132.3.354

Ceraso, J. (1967). The recall of long and short stimulus-lists. *American Journal of Psychology, 80,* 221-228. doi:10.2307/1420980

Ceraso, J., Bader, L., & Silverstein, M. (1970). Response latency as a function of list length in paired-associate learning. *Psychonomic Science, 19,* 239-240. doi:10.3758/BF03328796

Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 389-396. doi:10.1037/0278-7393.19.2.389

Cheng, N. Y. (1929). Retroactive effect and degree of similarity. *Journal of Experimental Psychology, 12,* 444-449. doi:10.1037/h0071397

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 106,* 8719-8724. doi:10.1073/pnas.0900234106

Chua, E. F., Schacter, D. L., & Sperling, R. (2009). Neural correlates of metamemory: A comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of Cognitive Neuroscience, 21,* 1751-1765. doi:10.1162/jocn.2009.21123

Chua, E. F., & Solinger, L. A. (2015). Building metamemorial knowledge over time: Insights from eye tracking about the bases of feeling-of-knowing and confidence judgments. *Frontiers in Psychology, 6,* 1206. doi:10.3389/fpsyg.2015.01206

Cinel, C., Cortis Mack, C., & Ward, G. (2018). Towards augmented human memory: Retrieval-induced forgetting and retrieval practice in an interactive, end-of-day review. *Journal of Experimental Psychology: General,* 147, 632-661. doi:10.1037/xge0000441

Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1403-1414. doi:10.1037/0278-7393.25.6.1403

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3,* 37-60. doi:10.3758/BF03210740

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159. doi:10.1037/0033-2909.112.1.155

Cole, S. M., Reysen, B., & Kelley, M. R. (2013). Part-set cuing facilitation for spatial information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1615-1620. doi:10.1037/a0032424

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A,* 497-505. doi:10.1080/14640748108400805

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). New York: Cambridge University Press.

Crescentini, C., Shallice, T., Del Missier, F., & Macaluso, E. (2010). Neural correlates of episodic retrieval: An fMRI study of the part-list cueing effect. *NeuroImage, 50,* 678-692. doi:10.1016/j.neuroimage.2009.12.114

Crespi, S., Robino, C., Silva, O., & de'Sperati, C. (2012). Spotting expertise in the eyes: Billiards knowledge as revealed by gaze shifts in a dynamic visual prediction task. *Journal of Vision, 12.* doi:10.1167/12.11.30

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55,* 461-478. doi:10.1016/j.jml.2006.08.003

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59,* 297-319. doi:10.1016/j.cogpsych.2009.07.003

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 484-499. doi:10.1037/a0018435

Criss, A. H., Aue, W., & Kılıç, A. (2014). Age and response bias: Evidence from the strength-based mirror effect. *Quarterly Journal of Experimental Psychology, 67,* 1910-1924. doi:10.1080/17470218.2013.874037

Criss, A. H., & Howard, M. W. (2015). Models of episodic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), The Oxford handbook of computational and mathematical psychology (pp. 165-183). New York, NY: Oxford University Press.

Criss, A. H., & Koop, G. J. (2015). Differentiation in episodic memory. In J. G. W. Raaijmakers, A. H. Criss, R. L. Goldstone, R. M. Nosofsky, and M. Steyvers (Eds.), Cognitive modeling in perception and memory: A festschrift for Richard M. Shiffrin (pp. 112-125). New York, NY: Psychology Press.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64,* 316-326. doi:10.1016/j.jml.2011.02.003

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLIM). *Journal of Memory and Language, 55,* 447-460. doi:10.1016/j.jml.2006.06.003

Criss, A. H., & Shiffrin, R. M. (2004a). Context noise and item noise jointly determine recognition memory: A comment on Dennis and Humphreys (2001). *Psychological Review, 111,* 800-807. doi:10.1037/0033-295X.111.3.800

Criss, A. H., & Shiffrin, R. M. (2004b). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition, 32,* 1284-1297. doi:10.3758/BF03206319

Criss, A. H., & Shiffrin, R. M. (2005). List discrimination in associative recognition and implications for representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1199-1212. doi:10.1037/0278-7393.31.6.1199

Criss, A. H., Wheeler, M. E., & McClelland, J. L. (2013). A differentiation account of recognition memory: Evidence from fMRI. *Journal of Cognitive Neuroscience, 25,* 421-435. doi:10.1162/jocn_a_00292

Curtis, E. T., Chubala, C. M., Spear, J., Jamieson, R. K., Hockley, W. E., & Crump, M. J. C. (2016). False recognition of instruction-set lures. *Memory, 24,* 32-43. doi:10.1080/09658211.2014.982657

D'Agostino, P. R., & DeRemer, P. (1972). Item repetition in free and cued recall. *Journal of Verbal Learning & Verbal Behavior, 11,* 54-58. doi:10.1016/S0022-5371(72)80059-8

Dalezman, J. J. (1976). Effects of output order on immediate, delayed, and final recall performance. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 597-608. doi:10.1037/0278-7393.2.5.597

Davis, G. A. (1966). Effects of list length and the number of response alternatives in a serially learned paired-associates task. *Journal of General Psychology, 75,* 29-33. doi:10.1080/00221309.1966.9710347

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17-22. doi:10.1037/h0046671

Delaney, P. F., Godbole, N. R., Holden, L. R., & Chang, Y. (2018). Working memory capacity and the spacing effect in cued recall. *Memory, 26,* 784-797. doi:10.1080/09658211.2017.1408841

Delaney, P. F., & Knowles, M. E. (2005). Encoding strategy changes and spacing effects in the free recall of unmixed lists. *Journal of Memory and Language, 52,* 120-130. doi:10.1016/j.jml.2004.09.002

Delaney, P. F., Spirgel, A. S., & Toppino, T. C. (2012). A deeper analysis of the spacing effect after "deep" encoding. *Memory & Cognition, 40,* 1003-1015. doi:10.3758/s13421-012-0207-3

Delaney, P. F., & Verkoeijen, P. P. J. L. (2009). Rehearsal strategies can enlarge or diminish the spacing effect: Pure versus mixed lists and encoding strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1148-1161. doi:10.1037/a0016380

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), The psychology of learning and motivation: Advances in research and theory (vol. 53, pp. 63-147). San Diego, CA: Elsevier Academic Press. doi:10.1016/S0079-7421(10)53003-2

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology, 79,* 162-170. doi:10.1037/0022-0663.79.2.162

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist, 43,* 627-634. doi:10.1037/0003-066X.43.8.627

Dennis, S., & Chapman, A. (2010). The inverse list length effect: A challenge for pure exemplar models of recognition memory. *Journal of Memory and Language, 63,* 416-424. doi:10.1016/j.jml.2010.06.001

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452-478. doi:10.1037/0033-295X.108.2.452

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59,* 361-376. doi:10.1016/j.jml.2008.06.007

Diana, R. A., & Reder, L. M. (2005). The list strength effect: A contextual competition account. *Memory & Cognition, 33,* 1289-1302. doi:10.3758/BF03193229

Dong, T. (1972). Cued partial recall of categorized words. *Journal of Experimental Psychology, 93,* 123-129. doi:10.1037/h0032488

Domhoff, G. W., & Fox, K. C. R. (2015). Dreaming and the default network: A review, synthesis, and counterintuitive research proposal. *Consciousness and Cognition, 33,* 342-353. doi:10.1016/j.concog.2015.01.019

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84,* 795-805. doi:10.1037/0021-9010.84.5.795

Dulsky, S. G. (1935). The effect of a change of background on recall and relearning. *Journal of Experimental Psychology, 18,* 725-740. doi:10.1037/h0058066

Ebbinghaus, H. (1913). Memory: A contribution to experimental psychology (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Teachers College Press. (Original work published 1885).

Eich, E. (1995). Mood as a mediator of place dependent memory. *Journal of Experimental Psychology: General, 124,* 293-308. doi:10.1037/0096-3445.124.3.293

Eich, J. E. (1980). The cue-dependent nature of state-dependent retrieval. *Memory & Cognition, 8,* 157-173. doi:10.3758/BF03213419

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89,* 627-661. doi:10.1037/0033-295X.89.6.627

Eich, J. M. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review, 92,* 1-38. doi:10.1037/0033-295X.92.1.1

Erber, J. T. (1974). Age differences in recognition memory. *Journal of Gerontology, 29,* 177-181. doi:10.1093/geronj/29.2.177

Erdelyi, M. H. (2006). The unified theory of repression. *Behavioral and Brain Sciences, 29,* 499-511. doi:10.1017/S0140525X06009113

Estes, W. K. (1955a). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369-377. doi:10.1037/h0046888

Estes, W. K. (1955b). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62,* 145-154. doi:10.1037/h0048509

Farley, J., Risko, E. F., & Kingstone, A. (2013). Everyday attention and lecture retention: The effects of time, fidgeting, and mind wandering. *Frontiers in Psychology, 4,* 619. doi:10.3389/fpsyg.2013.00619

Farnsworth, P. R. (1934). Examinations in familiar and unfamiliar surroundings. *Journal of Social Psychology, 5,* 128-129. doi:10.1080/00224545.1934.9921593

Fiacconi, C. M., Kouptsova, J. E., & Köhler, S. (2017). A role for visceral feedback and interoception in feelings-of-knowing. *Consciousness and Cognition, 53,* 70-80. doi:10.1016/j.concog.2017.06.001

Foos, P. W., & Smith, K. H. (1974). Effects of spacing and spacing patterns in free recall. *Journal of Experimental Psychology, 103,* 112-116. doi:10.1037/h0036828

Foroni, F., Vignando, M., Aiello, M., Parma, V., Paoletti, M. G., Squartini, A., & Rumiati, R. I. (2017). The smell of terroir! Olfactory discrimination between wines of different grape variety and different terroir. *Food Quality and Preference, 58,* 18-23. doi:10.1016/j.foodqual.2016.12.012

Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2013). Window to the wandering mind: Pupillometry of spontaneous thought while reading. *Quarterly Journal of Experimental Psychology, 66,* 2289-2294. doi:10.1080/17470218.2013.858170

Freud, S. (1964). The standard edition of the complete psychological works of Sigmund Freud. (J. Strachey, Trans.). Oxford, England: Macmillan. (Original work published 1895)

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation, 14,* 375-399. doi:10.3758/BF03203275

Fritzen, J. (1975). Intralist repetition effects in free recall. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 756-763. doi:10.1037/0278-7393.1.6.756

Gibson, E. J. (1940). A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review, 47,* 196-229. doi:10.1037/h0060582

Gibson, E. J., & Walk, R. D. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology, 49,* 239-242. doi:10.1037/h0048274

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review, 62,* 32-41. doi:10.1037/h0048826

Gillund, G., & Shiffrin, R. M. (1981). Free recall of complex pictures and abstract words. *Journal of Verbal Learning & Verbal Behavior, 20,* 575-592. doi:10.1016/S0022-5371(81)90192-4

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1-67. doi:10.1037/0033-295X.91.1.1

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13,* 8-20. doi:10.3758/BF03198438

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5-16. doi:10.1037/0278-7393.16.1.5

Glanzer, M., Hilford, A., & Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1176-1195. doi:10.1037/0278-7393.30.6.1176

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 500-513. doi:10.1037/0278-7393.25.2.500

Glenberg, A. M. (1977). Influences of retrieval processes on the spacing effect in free recall. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 282-294. doi:10.1037/0278-7393.3.3.282

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. British *Journal of Psychology, 66,* 325-331. doi:10.1111/j.2044-8295.1975.tb01468.x

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585-612. doi:10.1146/annurev.psych.49.1.585

Gómez-Ariza, C. J., Pelegrina, S., Lechuga, M. T., Suárez, A., & Bajo, M. T. (2009). Inhibition and retrieval of facts in young and older adults. *Experimental Aging Research, 35,* 83-97. doi:10.1080/03610730802545234

Gordon, S. K., & Clark, W. C. (1974). Application of signal detection theory to prose recall and recognition in elderly and young adults. *Journal of Gerontology, 29,* 64-72. doi:10.1093/geronj/29.1.64

Greene, R. L. (1990). Spacing effects on implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1004-1011. doi:10.1037/0278-7393.16.6.1004

Greene, R. L., & Stillwell, A. M. (1995). Effects of encoding variability and spacing on frequency discrimination. *Journal of Memory and Language, 34,* 468-476. doi:10.1006/jmla.1995.1021

Grenfell-Essam, R., Ward, G., & Tan, L. (2013). The role of rehearsal on the output order of immediate free recall of short and long lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 317-347. doi:10.1037/a0028974

Grenfell-Essam, R., Ward, G., & Tan, L. (2017). Common modality effects in immediate free recall and immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 1909-1933. doi:10.1037/xlm0000430

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1355-1369. doi:10.1037/0278-7393.20.6.1355

Haberlandt, K., Thomas, J. G., Lawrence, H., & Krohn, T. (2005). Transposition asymmetry in immediate serial recall. *Memory, 13,* 274-282. doi:10.1080/09658210344000297

Hamilton, M., & Rajaram, S. (2003). States of awareness across multiple memory tasks: Obtaining a "pure" measure of conscious recollection. *Acta Psychologica, 112,* 43-69.

Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: The role of active forgetting in memory. *Trends in Cognitive Sciences, 17,* 111-120. doi:10.1016/j.tics.2013.01.001

Hastie, R. (1975). Intralist repetition in free recall: Effects of frequency attribute recall instructions. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 3-12. doi:10.1037/0278-7393.1.1.3

Henmon, V. A. C. (1917). The relation between learning and retention and amount to be learned. *Journal of Experimental Psychology, 2,* 476-484. doi:10.1037/h0070292

Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomic Bulletin & Review, 11,* 125-130. doi:10.3758/BF03206471

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of repetitions. *Journal of Experimental Psychology, 80,* 139-145. doi:10.1037/h0027133

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers, 16,* 96-101. doi:10.3758/BF03202365

Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review, 93,* 411-428. doi:10.1037/0033-295X.93.4.411

Hintzman, D. L. (1987). Recognition and recall in MINERVA 2: Analysis of the 'recognition-failure' paradigm. In P. Morris (Ed.), Modelling cognition (pp. 215-229). Oxford, England: John Wiley & Sons.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528-551. doi:10.1037/0033-295X.95.4.528

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 667-680. doi:10.1037/0278-7393.18.4.667

Hintzman, D. L., & Hartry, A. L. (1990). Commensurability in memory for frequency. *Journal of Memory and Language, 29,* 501-523. doi:10.1016/0749-596X(90)90049-6

Hintzman, D. L., & Rogers, M. K. (1973). Spacing effects in picture memory. *Memory & Cognition, 1,* 430-434. doi:10.3758/BF03208903

Hintzman, D. L., Summers, J. J., Eki, N. T., & Moore, M. D. (1975). Voluntary attention and the spacing effect. *Memory & Cognition, 3,* 576-580. doi:10.3758/BF03197533

Hilford, A., Glanzer, M., & Kim, K. (1997). Encoding, repetition, and the mirror effect in recognition memory: Symmetry in motion. *Memory & Cognition, 25,* 593-605. doi:10.3758/BF03211302

Hipple, R. (1972). Retention of paired associates as a function of list length. *Journal of Experimental Psychology, 92,* 435-437. doi:10.1037/h0032365

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302-313. doi:10.1037/0278-7393.21.2.302

Hockley, W. E. (1994). Reflections of the mirror effect for item and associative recognition. *Memory & Cognition, 22,* 713-722. doi:10.3758/BF03209256

Hockley, W. E. (2008). The effects of environmental context on recognition memory and claims of remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1412-1429. doi:10.1037/a0013016

Hockley, W. E., Bancroft, T. D., & Bryant, E. (2012). Associative and familiarity-based effects of environmental context on memory. *Canadian Journal of Experimental Psychology, 66,* 81-89. doi:10.1037/a0027136

Hockley, W. E., & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory & Cognition, 35,* 679-688. doi:10.3758/BF03193306

Howell, W. C. (1973). Storage of events and event frequencies: A comparison of two paradigms in memory. *Journal of Experimental Psychology, 98,* 260-263. doi:10.1037/h0034380

Hu, X., Bergström, Z. M., Gagnepain, P., & Anderson, M. C. (2017). Suppressing unwanted memories reduces their unintended influences. *Current Directions in Psychological Science, 26,* 197-206. doi:10.1177/0963721417689881

Huber, D. E., Tomlinson, T. D., Jang, Y., & Hopper, W. J. (2015). The search of associative memory with recovery interference (SAM-RI) memory model and its application to retrieval practice paradigms. In J. G. W. Raaijmakers, A. H. Criss, R. L. Goldstone, R. M. Nosofsky, and M. Steyvers (Eds.), Cognitive modeling in perception and memory: A festschrift for Richard M. Shiffrin (pp. 81-98). New York, NY: Psychology Press.

Hulbert, J. C., Shivde, G., & Anderson, M. C. (2012). Evidence against associative blocking as a cause of cue-independent retrieval-induced forgetting. *Experimental Psychology, 59,* 11-21. doi:10.1027/1618-3169/a000120

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96,* 208-233. doi:10.1037/0033-295X.96.2.208

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology, 33,* 36-67. doi:10.1016/0022-2496(89)90003-5

Isarida, T., Isarida, T. K., Kubota, T., Higuma, M., & Matsuda, Y. (2018). Influences of context load and sensibleness of background photographs on local environmental context-dependent recognition. *Journal of Memory and Language, 101,* 114-123. doi:10.1016/j.jml.2018.04.006

Isarida, T., Sakai, T., Kubota, T., Koga, M., Katayama, Y., & Isarida, T. K. (2014). Odor-context effects in free recall after a short retention interval: A new methodology for controlling adaptation. *Memory & Cognition, 42,* 421-433. doi:10.3758/s13421-013-0370-1

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

Jahnke, J. C. (1969). Output interference and the Ranschburg effect. *Journal of Verbal Learning & Verbal Behavior, 8,* 614-621. doi:10.1016/S0022-5371(69)80113-1

James, W. (1890). The principles of psychology. New York, NY: Henry Holt. [Reprinted as W. James (1983). The principles of psychology. Cambridge, MA: Harvard University Press.]

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research, 30,* 138-149. doi:10.1086/374692

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *American Journal of Psychology, 35,* 19-24. doi:10.2307/1414040

Jin, C. Y., Borst, J. P., & van Vugt, M. K. (2019). Predicting task-general mind-wandering with EEG. *Cognitive, Affective & Behavioral Neuroscience.* Advance online publication. doi:10.3758/s13415-019-00707-1

Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review, 120,* 852-872. doi:10.1037/a0034246

Jonker, T. R., Seli, P., & MacLeod, C. M. (2015). Retrieval-induced forgetting in context. *Current Directions in Psychological Science, 24,* 273-278. doi:10.1177/0963721415573203

Kahana, M. J. (1996). Associate retrieval processes in free recall. *Memory & Cognition, 24,* 103-109. doi:10.3758/BF03197276

Kahana, M. J., & Greene, R. L. (1993). Effects of spacing on memory for homogeneous lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 159-162. doi:10.1037/0278-7393.19.1.159

Kahana, M. J., & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin & Review, 12,* 159-164. doi:10.3758/BF03196362

Kane, M. J., & McVay, J. C. (2012). What mind wandering reveals about executive-control abilities and failures. *Current Directions in Psychological Science, 21,* 348-354. doi:10.1177/0963721412454875

Kang, M.-S., & Oh, B.-I. (2016). Grouping influences output interference in short-term memory: A mixture modeling study. *Frontiers in Psychology, 7,* 585. doi:10.3389/fpsyg.2016.00585

Karlsen, P. J., & Snodgrass, J. G. (2004). The word-frequency paradox for recall/recognition occurs for pictures. *Psychological Research, 68,* 271-276. doi:10.1007/s00426-003-0138-5

Kelley, M. R., Parasiuk, Y., Salgado-Benz, J., & Crocco, M. (2016). Spatial part-set cuing facilitation. *Memory, 24,* 737-745. doi:10.1080/09658211.2015.1046382

Kim, A. S. N., Wong-Kee-You, A. M. B., Wiseheart, M., & Rosenbaum, R. S. (2019). The spacing effect stands up to big data. *Behavior Research Methods.* Advance online publication. doi:10.3758/s13428-018-1184-7

Kimball, D. R., Smith, T. A., & Kahana, M. J. (2007). The fSAM model of false recall. *Psychological Review, 114,* 954-993. doi:10.1037/0033-295X.114.4.954

Kihlstrom, J. F. (2002). No need for repression. *Trends in Cognitive Sciences, 6,* 502. doi:10.1016/S1364-6613(02)02006-5

Kılıç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology, 92,* 65-86. doi:10.1016/j.cogpsych.2016.11.005

Kılıç, A., & Öztekin, I. (2014). Retrieval dynamics of the strength based mirror effect in recognition memory. *Journal of Memory and Language, 76,* 158-173. doi:10.1016/j.jml.2014.06.009

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition, 39,* 348-363. doi:10.3758/s13421-010-0007-6

Kinnell, A., & Dennis, S. (2012). The role of stimulus type in list length effects in recognition memory. *Memory & Cognition, 40,* 311-325. doi:10.3758/s13421-011-0164-2

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

Konishi, M., Brown, K., Battaglini, L., & Smallwood, J. (2017). When attention wanders: Pupillometric signatures of fluctuations in external attention. *Cognition, 168,* 16-26. doi:10.1016/j.cognition.2017.06.006

Koop, G. J., Criss, A. H., & Malmberg, K. J. (2015). The role of mnemonic processes in pure-target and pure-foil recognition memory. *Psychonomic Bulletin & Review, 22,* 509-516. doi:10.3758/s13423-014-0703-5

Koop, G. J., Criss, A. H., & Pardini, A. M. (2019). A strength-based mirror effect persists even when criterion shifts are unlikely. *Memory & Cognition.* Advance online publication. doi:10.3758/s13421-019-00906-8

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

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning & Verbal Behavior, 12,* 119-131. doi:10.1016/S0022-5371(73)80001-5

Lawrence, M. A. (2016). ez: Easy analysis and visualization of factorial experiments [R package]. https://CRAN.R-project.org/package=ez

Lehman, M., & Malmberg, K. J. (2011). Overcoming the effects of intentional forgetting. *Memory & Cognition, 39,* 335-347. doi:10.3758/s13421-010-0025-4

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review, 120,* 155-189. doi:10.1037/a0030851

Levy, B. J., & Anderson, M. C. (2011). On the relationship between interference and inhibition in cognition. In A. S. Benjamin (Ed.), Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork (pp. 107-132). New York, NY: Psychology Press.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning & Verbal Behavior, 13,* 585-589. doi:10.1016/S0022-5371(74)80011-3

Lyon, D. O. (1914a). The relation of length of material to time taken for learning, and the optimum distribution of time. Part I. *Journal of Educational Psychology, 5,* 1-9. doi:10.1037/h0075090

Lyon, D. O. (1914b). The relation of length of material to time taken for learning, and the optimum distribution of time. Part II. *Journal of Educational Psychology, 5,* 85-91. doi:10.1037/h0070943

Lyon, D. O. (1914c). The relation of length of material to time taken for learning, and the optimum distribution of time. Part III. *Journal of Educational Psychology, 5,* 155-163. doi:10.1037/h0070925

Mack, C. C., Cinel, C., Davies, N., Harding, M., & Ward, G. (2017). Serial position, output order, and list length effects for words presented on smartphones over very long intervals. *Journal of Memory and Language, 97,* 61-80. doi:10.1016/j.jml.2017.07.009

MacLeod, C. M. (2007a). Inhibition: Elusive or illusion? In H. L. Roediger, Y. Dudai, & S. M. Fitzpatrick (Eds.), Science of memory: Concepts (pp. 301-306). Oxford, England: Oxford University Press.

MacLeod, C. M. (2007b). The concept of inhibition in cognition. In D. S. Gorfein & C. M. MacLeod (Eds.), Inhibition in cognition (pp. 3-23). Washington, DC: American Psychological Association. doi:10.1037/11587-001

MacLeod, C. M., Dodd, M. D., Sheard, E. D., Wilson, D. E., & Bibi, U. (2003). In opposition to inhibition. In B. H. Ross (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 43, pp. 163-214). New York, NY: Elsevier Science.

Maddox, G. B., Pyc, M. A., Kauffman, Z. S., Gatewood, J. D., & Schonhoff, A. M. (2018). Examining the contributions of desirable difficulty and reminding to the spacing effect. *Memory & Cognition, 46,* 1376-1388. doi:10.3758/s13421-018-0843-3

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning & Verbal Behavior, 8,* 828-835. doi:10.1016/S0022-5371(69)80050-2

Maguire, A. M., Humphreys, M. S., Dennis, S., & Lee, M. D. (2010). Global similarity accounts of embedded-category designs: Tests of the global matching models. *Journal of Memory and Language, 63,* 131-148. doi:10.1016/j.jml.2010.03.007

Maillet, D., & Rajah, M. N. (2013). Age-related changes in frequency of mind-wandering and task-related interferences during memory encoding and their impact on retrieval. *Memory, 21,* 818-831. doi:10.1080/09658211.2012.761714

Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and

judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 319-331. doi:10.1037/0278-7393.30.2.319

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 616-630. doi:10.1037/0278-7393.28.4.616

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 322-336. doi:10.1037/0278-7393.31.2.322

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30,* 607-613. doi:10.3758/BF03194962

Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by midazolam. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 540-549. doi:10.1037/0278-7393.30.2.540

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252-271. doi:10.1037/0033-295X.87.3.252

Mandler, G. (2008). Familiarity breeds attempts: A critical review of dual-process theories of recognition. *Perspectives on Psychological Science, 3,* 390-399. doi:10.1111/j.1745-6924.2008.00087.x

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724-760. doi:10.1037/0033-295X.105.4.734-760

McGeoch, J. A. (1929). The influence of degree of learning upon retroactive inhibition. *American Journal of Psychology, 41,* 252-262. doi:10.2307/1415236

McGeoch, J. A. (1931). The influence of four different interpolated activities upon retention. *Journal of Experimental Psychology, 14,* 400-413. doi:10.1037/h0074435

McGeoch, J. A. (1932a). Forgetting and the law of disuse. *Psychological Review, 39,* 352-370. doi:10.1037/h0069819

McGeoch, J. A. (1932b). The influence of degree of interpolated learning upon retroactive inhibition. *American Journal of Psychology, 44,* 695-708. doi:10.2307/1414532

McGeoch, J. A. (1933a). Studies in retroactive inhibition: I. The temporal course of the inhibitory effects of interpolated learning. *Journal of General Psychology, 9,* 24-43. doi:10.1080/00221309.1933.9920911

McGeoch, J. A. (1933b). Studies in retroactive inhibition: II. Relationships between temporal point of interpolation, length of interval, and amount of retroactive inhibition. *Journal of General Psychology, 9,* 44-57. doi:10.1080/00221309.1933.9920912

McGeoch, J. A. (1936). Studies in retroactive inhibition: VII. Retroactive inhibition as a function of the length and frequency of presentation of the interpolated lists. *Journal of Experimental Psychology, 19,* 674-693. doi:10.1037/h0060533

McGeoch, J. A. (1942). The psychology of human learning. Oxford, England: Longmans, Green.

McGeoch, J. A., & McDonald, W. T. (1931). Meaningful relation in retroactive inhibition. *American Journal of Psychology, 43,* 579-588. doi:10.2307/1415159

McGeoch, J. A., & McGeoch, G. O. (1937). Studies in retroactive inhibition: X. The influence of similarity of meaning between lists of paired associates. *Journal of Experimental Psychology, 21,* 320-329. doi:10.1037/h0062260

Melton, A. W. (1967). Repetition and retrieval from memory. *Science, 158,* 532. doi:10.1126/science.158.3800.532-b

Melton, A. W. (1970). The situation with respect to the spacing of repetitions in memory. *Journal of Verbal Learning & Verbal Behavior, 9,* 596-606. doi:10.1016/S0022-5371(70)80107-4

Mensink, G.-J. M., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review, 95,* 434-455. doi:10.1037/0033-295X.95.4.434

Mensink, G.-J. M., & Raaijmakers, J. G. W. (1989). A model for contextual fluctuation. *Journal of Mathematical Psychology, 33,* 172-186. doi:10.1016/0022-2496(89)90029-1

Menzel, R., Manz, G., Menzel, R., & Greggers, U. (2001). Massed and spaced learning in honeybees: The role of CS, US, the intertrial interval, and the test interval. *Learning & Memory, 8,* 198-208. doi:10.1101/lm.40001

Metcalfe, J. (1990). Composite Holographic Associative Recall Model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119,* 145-160. doi:10.1037/0096-3445.119.2.145

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review, 100,* 3-22. doi:10.1037/0033-295X.100.1.3

Mooneyham, B. W., & Schooler, J. W. (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology, 67,* 11-18. doi:10.1037/a0031569

Moran, R., & Goshen-Gottstein, Y. (2015). Old processes, new perspectives: Familiarity is correlated with (not independent of) recollection and is more (not equally) variable for targets than for lures. *Cognitive Psychology, 79,* 40-67. doi:10.1016/j.cogpsych.2015.01.005

Morehead, K., Dunlosky, J., Rawson, K. A., Bishop, M., & Pyc, M. A. (2018). Does mediator use contribute to the spacing effect for cued recall? Critical tests of the mediator hypothesis. *Memory, 26,* 535-546. doi:10.1080/09658211.2017.1381266

Mueller, J. H., & Brown, S. C. (1977). Output interference and intralist repetition in free recall. *American Journal of Psychology, 90,* 157-164. doi:10.2307/1421649

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609-626. doi:10.1037/0033-295X.89.6.609

Murdock, B. B. (1983). A distributed memory model for serial-order information. *Psychological Review, 90,* 316-338. doi:10.1037/0033-295X.90.4.316

Murdock, B. B. (1989). Learning in a distributed memory model. In C. Izawa (Ed.), Current issues in cognitive processes: The Tulane Flowerree Symposium on Cognition (pp. 69-106). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Murdock, B. B. (1992). Item and associative information in a distributed memory model. *Journal of Mathematical Psychology, 36,* 68-99. doi:10.1016/0022-2496(92)90053-A

Murdock, B. B. (1997). Context and mediators in a theory of distributed associated memory (TODAM2). *Psychological Review, 104,* 839-862. doi:10.1037/0033-295X.104.4.839

Murdock, B. B., & Kahana, M. J. (1993a). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 689-697. doi:10.1037/0278-7393.19.3.689

Murdock, B. B., & Kahana, M. J. (1993b). List-strength and list-length effects: Reply to Shiffrin, Ratcliff, Murnane, and Nobel (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1450-1453. doi:10.1037/0278-7393.19.6.1450

Murnane, K., & Phelps, M. P. (1993). A global activation approach to the effect of changes in environmental context on recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 882-894. doi:10.1037/0278-7393.19.4.882

Murnane, K., & Shiffrin, R. M. (1991a). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 855-874. doi:10.1037/0278-7393.17.5.855

Murnane, K., & Shiffrin, R. M. (1991b). Word repetitions in sentence recognition. *Memory & Cognition, 19,* 119-130. doi:10.3758/BF03197109

Myles-Worsley, M., Johnston, W. A., & Simons, M. A. (1988). The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 553-557. doi:10.1037/0278-7393.14.3.553

Nickerson, R. S. (1984). Inhibition from part-set cuing: A persisting enigma for memory research. *Memory & Cognition, 12,* 531-552. doi:10.3758/BF03213342

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 384-413. doi:10.1037/0278-7393.27.2.384

Noel, H., & Vallen, B. (2009). The spacing effect in marketing: A review of extant findings and directions for future research. *Psychology & Marketing, 26,* 951-969. doi:10.1002/mar.20307

Norman, D. A., & Waugh, N. C. (1968). Stimulus and response interference in recognition-memory experiments. *Journal of Experimental Psychology, 78,* 551-559. doi:10.1037/h0026637

Norman, K. A. (1999). Differential effects of list strength on recognition and familiarity (doctoral dissertation, Harvard University). Retrieved from ProQuest Dissertations & Theses. (1999-95024-412)

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 1083-1094. doi:10.1037/0278-7393.28.6.1083

Norman, K. A., Newman, E. L., & Detre, G. (2007). A neural network model of retrieval-induced forgetting. *Psychological Review, 114,* 887-953. doi:10.1037/0033-295X.114.4.887

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110,* 611-646. doi:10.1037/0033-295X.110.4.611

Norman, K. A., Tepe, K., Nyhus, E., & Curran, T. (2008). Event-related potential correlates of interference effects on recognition memory. *Psychonomic Bulletin & Review, 15,* 36-43. doi:10.3758/PBR.15.1.36

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87-108. doi:10.1037/0278-7393.13.1.87

Ohrt, D. D., & Gronlund, S. D. (1999). List-length effect and continuous memory: Confounds and solutions. In C. Izawa (Ed.), On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson-Shiffrin model (pp. 105-125). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8,* 434-447. doi:10.1037/1082-989X.8.4.434

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122,* 260-311. doi:10.1037/a0038692

Osth, A. F., Dennis, S., & Kinnell, A. (2014). Stimulus type and the list strength paradigm. *Quarterly Journal of Experimental Psychology, 67,* 1826-1841. doi:10.1080/17470218.2013.872824

Osth, A. F., Fox, J., McKague, M., Heathcote, A., & Dennis, S. (2018). The list strength effect in source memory: Data and a global matching model. *Journal of Memory and Language, 103,* 91-113. doi:10.1016/j.jml.2018.08.002

Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology, 104,* 106-142. doi:10.1016/j.cogpsych.2018.04.002

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review, 105*, 761–781. doi:10.1037/0033-295X.105.4.761-781

Pan, S. (1926). The influence of context upon learning and recall. *Journal of Experimental Psychology, 9,* 468-491. doi:10.1037/h0073472

Parr, W. V., Heatherbell, D., & White, K. G. (2002). Demystifying wine expertise: Olfactory threshold, perceptual skill and semantic memory in expert and novice wine judges. *Chemical Senses, 27,* 747-755. doi:10.1093/chemse/27.8.747

Paynter, C. A., Reder, L. M., & Kieffaber, P. D. (2009). Knowing we know before we know: ERP correlates of initial feeling-of-knowing. *Neuropsychologia, 47,* 796-803. doi:10.1016/j.neuropsychologia.2008.12.009

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8-13. doi:10.1016/j.jneumeth.2006.11.017

Peixotto, H. E. (1947). Proactive inhibition in the recognition of nonsense syllables. *Journal of Experimental Psychology, 37,* 81-91. doi:10.1037/h0060509

Penney, C. G. (1985). Elimination of the suffix effect on preterminal list items with unpredictable list length: Evidence for a dual model of suffix effects. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 11, 229-247. doi:10.1037/0278-7393.11.2.229

Perfect, T. J., Stark, L.-J., Tree, J. J., Moulin, C. J., Ahmed, L., & Hutter, R. (2004). Transfer appropriate forgetting: The cue-dependent nature of retrieval-induced forgetting. *Journal of Memory and Language, 51,* 399-417. doi:10.1016/j.jml.2004.06.003

Pessin, J. (1932). The effect of similar and dissimilar conditions upon learning and relearning. *Journal of Experimental Psychology, 15,* 427-435. doi:10.1037/h0075537

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58,* 193-198. doi:10.1037/h0049234

Peynırcıoğlu, Z. F. (1989). Part-set cuing effect with word-fragment cuing: Evidence against the strategy disruption and increased-list-length explanations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 147-152. doi:10.1037/0278-7393.15.1.147

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review, 91,* 281-294. doi:10.1037/0033-295X.91.3.281

Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory & Cognition, 1,* 19-40. doi:10.3758/BF03198064

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria.

Raaijmakers, J. G. W., & Jakab, E. (2013a). Is forgetting caused by inhibition? *Current Directions in Psychological Science, 22,* 205-209. doi:10.1177/0963721412473472

Raaijmakers, J. G. W., & Jakab, E. (2013b). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language, 68,* 98-122. doi:10.1016/j.jml.2012.10.002

Raaijmakers, J. G. W., & Phaf, R. H. (1999). Part-list cuing revisited: Testing the sampling-bias hypothesis. In C. Izawa (Ed.), On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson-Shiffrin model (pp. 87-104). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

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

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93-134. doi:10.1037/0033-295X.88.2.93

Rand, G., & Wapner, S. (1967). Postural status as a factor in memory. *Journal of Verbal Learning & Verbal Behavior, 6,* 268-271. doi:10.1016/S0022-5371(67)80107-5

Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin, 140,* 1411-1431. doi:10.1037/a0037428

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163-178. doi:10.1037/0278-7393.16.2.163

Ratcliff, R., & Hockley, W. E. (1980). Repeated negatives in item recognition: Nonmonotonic lag functions. In R. S. Nickerson (Ed.), Attention and Performance (vol. VIII, pp. 555-573). Hillsdale, NJ: Erlbaum.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763-785. doi:10.1037/0278-7393.20.4.763

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review, 83,* 190-214. doi:10.1037/0033-295X.83.3.190

Ratcliff, R., Sheu, C., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518-535. doi:10.1037/0033-295X.99.3.518

Reder, L. M., Donavos, D. K., & Erickson, M. A. (2002). Perceptual match effects in direct tests of memory: The role of contextual fan. *Memory & Cognition, 30,* 312-323. doi:10.3758/BF03195292

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 294-320. doi:10.1037/0278-7393.26.2.294

Reed, H. J. (1931). The influence of a change of conditions upon the amount recalled. *Journal of Experimental Psychology, 14,* 632-649. doi:10.1037/h0069294

Reitman, J. S. (1974). Without surreptitious rehearsal, information in short-term memory decays. *Journal of Verbal Learning & Verbal Behavior, 13,* 365-377. doi:10.1016/S0022-5371(74)80015-0

Robbins, D., Bray, J. F., & Irvin, J. R. (1974). Intralist contrast effects in cued recall. *Journal of Experimental Psychology, 103,* 150-155. doi:10.1037/h0036823

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology, 92,* 365-372. doi:10.1037/h0032278

Robinson, E. S., & Darrow, C. W. (1924). Effect of length of list upon memory for numbers. *American Journal of Psychology, 35,* 235-243. doi:10.2307/1413826

Robinson, E. S., & Heron, W. T. (1922). Results of variations in length of memorized material. *Journal of Experimental Psychology, 5,* 428-448. doi:10.1037/h0075024

Roediger, H. L. (1974). Inhibiting effects of recall. *Memory & Cognition, 2,* 261-269. doi:10.3758/BF03208993

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803-814. doi:10.1037/0278-7393.21.4.803

Roediger, H. L., & Neely, J. H. (1982). Retrieval blocks in episodic and semantic memory. *Canadian Journal of Psychology, 36,* 213-242. doi:10.1037/h0080640

Roediger, H. L., & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired-associate lists. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 91-105. doi:10.1037/0278-7393.6.1.91

Ross, B. H., & Landauer, T. K. (1978). Memory for at least one of two items: Test and failure of several theories of spacing effects. *Journal of Verbal Learning & Verbal Behavior, 17,* 669-680. doi:10.1016/S0022-5371(78)90403-6

Ruch, T. C. (1928). Factors influencing the relative economy of massed and distributed practice in learning. *Psychological Review, 35,* 19-45. doi:10.1037/h0074423

Rundus, D. (1973). Negative effects of using list items as recall cues. *Journal of Verbal Learning & Verbal Behavior, 12,* 43-50. doi:10.1016/S0022-5371(73)80059-3

Rundus, D., & Atkinson, R. C. (1970). Rehearsal processes in free recall. *Journal of Verbal Learning & Verbal Behavior, 9,* 99-105. doi:10.1016/S0022-5371(70)80015-9

Runquist, W. N., & Horton, K. D. (1977). Output interference and the effects of phonemic similarity among cue stimuli. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 467-476. doi:10.1037/0278-7393.3.4.467

Rupprecht, J., & Bäuml, K.-H. T. (2016). Retrieval-induced forgetting in item recognition: Retrieval specificity revisited. *Journal of Memory and Language, 86,* 97-118. doi:10.1016/j.jml.2015.09.003

Russo, R., Parkin, A. J., Taylor, S. R., & Wilks, J. (1998). Revising current two-process accounts of spacing effects in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 161-172. doi:10.1037/0278-7393.24.1.161

Sahakyan, L., Abushanab, B., Smith, J. R., & Gray, K. J. (2014). Individual differences in contextual storage: Evidence from the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 873-881. doi:10.1037/a0035222

Sahakyan, L., & Malmberg, K. J. (2018). Divided attention during encoding causes separate memory traces to be encoded for repeated events. *Journal of Memory and Language, 101,* 153-161. doi:10.1016/j.jml.2018.04.004

Saltz, E. (1963). Compound stimuli in verbal learning: Cognitive and sensory differentiation versus stimulus selection. *Journal of Experimental Psychology, 66,* 1-5. doi:10.1037/h0043170

Samuels, S. J. (1970). Interaction of list length and low stimulus similarity on the von Restorff effect. *Journal of Educational Psychology, 61,* 57-58. doi:10.1037/h0028691

Saunders, J., & MacLeod, M. D. (2006). Can inhibition resolve retrieval competition through the control of spreading activation? *Memory & Cognition, 34,* 307-322. doi:10.3758/BF03193409

Scarmeas, N., Zarahn, E., Anderson, K. E., Honig, L. S., Park, A., Hilton, J., … Stern, Y. (2004). Cognitive reserve-mediated modulation of positron emission tomographic activations during memory tasks in Alzheimer disease. *Archives of Neurology, 61,* 73-78. doi:10.1001/archneur.61.1.73

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review, 108,* 257-272. doi:10.1037/0033-295X.108.1.257

Schooler, J. W., Mrazek, M. B., Franklin, M. S., Baird, B., Mooneyham, B. W., Zedelius, C., & Broadway, J. M. (2014). The middle way: Finding the balance between mindfulness and mind-wandering. In B. H. Ross (Ed.), The psychology of learning and motivation (vol. 60, pp. 1-33). San Diego, CA: Elsevier Academic Press. doi:10.1016/B978-0-12-800090-8.00001-9

Schulman, A. I. (1974). The declining course of recognition memory. *Memory & Cognition, 2,* 14-18. doi:10.3758/BF03197485

Schumsky, D. A., Grasha, A. F., Trinder, J., & Richman, C. L. (1969). List length and single-trial short-term memory. *Journal of Experimental Psychology, 82,* 238-241. doi:10.1037/h0028150

Searston, R. A., & Tangen, J. M. (2017). The emergence of perceptual expertise with fingerprints over time. *Journal of Applied Research in Memory and Cognition, 6,* 442-451. doi:10.1016/j.jarmac.2017.08.006

Seli, P., Carriere, J. S. A., Thomson, D. R., Cheyne, J. A., Martens, K. A. E., & Smilek, D. (2014). Restless mind, restless body. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 660-668. doi:10.1037/a0035260

Seli, P., Jonker, T. R., Cheyne, J. A., Cortes, K., & Smilek, D. (2015). Can research participants comment authoritatively on the validity of their self-reports of mind wandering and task engagement? *Journal of Experimental Psychology: Human Perception and Performance, 41,* 703-709. doi:10.1037/xhp0000029

Seli, P., Risko, E. F., Smilek, D., & Schacter, D. L. (2016). Mind-wandering with and without intention. *Trends in Cognitive Sciences, 20,* 605-617. doi:10.1016/j.tics.2016.05.010

Serra, M., & Nairne, J. S. (2000). Part-set cuing of order information: Implications for associative theories of serial order memory. *Memory & Cognition, 28,* 847-855. doi:10.3758/BF03198420

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. (1972). Further evidence on the MP-DP effect in free-recall learning. *Journal of Verbal Learning & Verbal Behavior, 11,* 1-12. doi:10.1016/S0022-5371(72)80053-7

Sheridan, H., & Reingold, E. M. (2014). Expert vs. novice differences in the detection of relevant information during a chess game: Evidence from eye movements. *Frontiers in Psychology, 5.* doi:10.3389/fpsyg.2014.00941

Shiffrin, R. M. (1970). Forgetting: Trace erosion or retrieval failure? *Science, 168,* 1601-1603. doi:10.1126/science.168.3939.1601

Shiffrin, R. M. (1973). Visual free recall. *Science, 180,* 980-982. doi:10.1126/science.180.4089.980

Shiffrin, R. M., & Raaijmakers, J. G. W. (1992). The SAM retrieval model: A retrospective and prospective. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), Essays in honor of William K. Estes: From learning processes to cognitive processes (Vol. 2, pp. 69-86). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 179-195. doi:10.1037/0278-7393.16.2.179

Shiffrin, R. M., Ratcliff, R., Murnane, K., & Nobel, P. (1993). TODAM and the list-strength and list-length effects: Comment on Murdock and Kahana (1993a).

Journal of Experimental Psychology: Learning, Memory, and Cognition, 19, 1445-1449. doi:10.1037/0278-7393.19.6.1445

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145-166. doi:10.3758/BF03209391

Shivde, G., & Anderson, M. C. (2001). The role of inhibition in meaning selection: Insights from retrieval-induced forgetting. In D. S. Gorfein (Ed.), On the consequences of meaning selection: Perspectives on resolving lexical ambiguity (pp. 175-190). Washington, DC: American Psychological Association. doi:10.1037/10459-010

Siegel, L. L., & Kahana, M. J. (2014). A retrieved context account of spacing and repetition effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 755-764. doi:10.1037/a0035585

Sirotin, Y. B., Kimball, D. R., & Kahana, M. J. (2005). Going beyond a single list: Modeling the effects of prior experience on episodic free recall. *Psychonomic Bulletin & Review, 12,* 787-805. doi:10.3758/BF03196773

Slamecka, N. J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology, 76,* 504-513. doi:10.1037/h0025695

Sloman, S. A., Bower, G. H., & Rohrer, D. (1991). Congruency effects in part-list cuing inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 974-982. doi:10.1037/0278-7393.17.5.974

Smallwood, J., Beach, E., Schooler, J. W., & Handy, T. C. (2008). Going AWOL in the brain: Mind wandering reduces cortical analysis of external events. *Journal of Cognitive Neuroscience, 20,* 458-469. doi:10.1162/jocn.2008.20037

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology, 66,* 487-518. doi:10.1146/annurev-psych-010814-015331

Smith, A. D. (1971). Output interference and organized recall from long-term memory. *Journal of Verbal Learning & Verbal Behavior, 10,* 400-408. doi:10.1016/S0022-5371(71)80039-7

Smith, A. D. (1973). Input order and output interference in organized recall. *Journal of Experimental Psychology, 100,* 147-150. doi:10.1037/h0035513

Smith, A. D., D'Agostino, P. R., & Reid, L. S. (1970). Output interference in long-term memory. *Canadian Journal of Psychology, 24,* 85-89. doi:10.1037/h0082845

Smith, E. E., Adams, N. E., & Schorr, D. (1978). Fact retrieval and the paradox of interference. *Cognitive Psychology, 10,* 438-464. doi:10.1016/0010-0285(78)90007-5

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 460-471. doi:10.1037/0278-7393.5.5.460

Smith, S. M. (1985). Effects of number of study environments and learning instructions on free-recall clustering and accuracy. *Bulletin of the Psychonomic Society, 23,* 440-442. doi:10.3758/BF03329846

Smith, S. M. (1988). Environmental context effects on memory. In G. M. Davies & D. M. Thomson (Eds.), Memory in context: Context in memory (pp. 13-34). New York, NY: Wiley.

Smith, S. M. (1994). Theoretical principles of context-dependent memory. In P. Morris & M. Gruneberg (eds.), Theoretical Aspects of Memory (pp. 168-195). New York, NY: Routledge.

Smith, S. M., Glenberg, A. M., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6,* 342-353. doi:10.3758/BF03197465

Smith, S. M., & Manzano, I. (2010). Video context-dependent recall. *Behavior Research Methods, 42,* 292-301. doi:10.3758/BRM.42.1.292

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review, 8,* 203-220. doi:10.3758/BF03196157

Snodgrass, J. G., Wasser, B., Finkelstein, M., & Goldberg, L. B. (1974). On the fate of visual and verbal memory codes for pictures and words: Evidence for a dual coding mechanism in recognition memory. *Journal of Verbal Learning & Verbal Behavior, 13,* 27-37. doi:10.1016/S0022-5371(74)80027-7

Son, L. K., & Simon, D. A. (2012). Distributed learning: Data, metacognition, and educational implications. *Educational Psychology Review, 24,* 379-399. doi:10.1007/s10648-012-9206-y

Spurgeon, J., Ward, G., Matthews, W. J., & Farrell, S. (2015). Can the effects of temporal grouping explain the similarities and differences between free recall and serial recall? *Memory & Cognition, 43,* 469-488. doi:10.3758/s13421-014-0471-5

Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1137-1151. doi:10.1037/a0028151

Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language, 63,* 18-34. doi:10.1016/j.jml.2010.03.004

Starns, J. J., White, C. N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory & Cognition, 40,* 1189-1199. doi:10.3758/s13421-012-0225-1

Stillman, R. C., Weingarten, H., Wyatt, R. J., Gillin, J. C., & Eich, J. (1974). State dependent (dissociative) effects of marihuana on human memory. *Archives of General Psychiatry, 31,* 81-85. doi:10.1001/archpsyc.1974.01760130061010

Storm, B. C. (2011). Retrieval-induced forgetting and the resolution of competition. In A. S. Benjamin (Ed.), Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork (pp. 89-105). New York, NY: Psychology Press.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1379-1396. doi:10.1037/0278-7393.24.6.1379

Strong, E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review, 19,* 447-462. doi:10.1037/h0069812

Strong, E. K. (1916). The factors affecting a permanent impression developed through repetition. *Journal of Experimental Psychology, 1,* 319-338. doi:10.1037/h0074989

Surprenant, A. M., & Neath, I. (2009). Principles of memory. New York, NY: Psychology Press.

Szpunar, K. K. (2017). Directing the wandering mind. *Current Directions in Psychological Science, 26,* 40-44. doi:10.1177/0963721416670320

Tabachnik, B., & Brotsky, S. J. (1976). Free recall and complexity of pictorial stimuli. *Memory & Cognition, 4,* 466-470. doi:10.3758/BF03213205

Thomson, D. R., Smilek, D., & Besner, D. (2014). On the asymmetric effects of mind-wandering on levels of processing at encoding and retrieval. *Psychonomic Bulletin & Review, 21,* 728-733. doi:10.3758/s13423-013-0526-9

Thorndike, E. L. (1913). The psychology of learning. New York, NY: Teachers College. doi:10.1037/13051-000

Toga, A. W. (1975). Strategy in auditory recognition memory. *Bulletin of the Psychonomic Society, 6,* 517-519. doi:10.3758/BF03337555

Tomlinson, T. D., Huber, D. E., Rieth, C. A., & Davelaar, E. J. (2009). An interference account of cue-independent forgetting in the no-think paradigm. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 106,* 15588-15593. doi:10.1073/pnas.0813370106

Torchino, M. (2018). effsize: Efficient effect size computation [R package]. doi:10.5281/zenodo.1480624

Tulving, E. (1974). Cue-dependent forgetting. *American Scientist, 62,* 74-82.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26,* 1-12. doi:10.1037/h0080017

Tulving, E., & Arbuckle, T. Y. (1963). Sources of intratrial interference in immediate recall of paired associates. *Journal of Verbal Learning & Verbal Behavior, 1,* 321-334. doi:10.1016/S0022-5371(63)80012-2

Tulving, E., & Arbuckle, T. Y. (1966). Input and output interference in short-term associative memory. *Journal of Experimental Psychology, 72,* 145-150. doi:10.1037/h0023344

Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology, 92,* 297-304. doi:10.1037/h0032367

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review, 64,* 49-60. doi:10.1037/h0044616

Underwood, B. J. (1978). Recognition memory as a function of length of study list. *Bulletin of the Psychonomic Society, 12,* 89-91. doi:10.3758/BF03329636

Underwood, B. J., & Freund, J. S. (1970). Word frequency and short-term recognition memory. *American Journal of Psychology, 83,* 343-351. doi:10.2307/1420411

van Son, D., De Blasio, F. M., Fogarty, J. S., Angelidis, A., Barry, R. J., & Putman, P. (2019). Frontal EEG theta/beta ratio during mind wandering episodes. *Biological Psychology, 140,* 19-27. doi:10.1016/j.biopsycho.2018.11.003

Verde, M. F. (2009). The list-strength effect in recall: Relative-strength competition and retrieval inhibition may both contribute to forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 205-220. doi:10.1037/a0014275

Verde, M. F. (2012). Retrieval-induced forgetting and inhibition: A critical review. In B. H. Ross (Ed.), The Psychology of Learning and Motivation (Vol. 56, pp. 47-80). San Diego, CA: Elsevier Academic Press. doi:10.1016/B978-0-12-394393-4.00002-9

Verde, M. F. (2013). Retrieval-induced forgetting in recall: Competitor interference revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1433-1448. doi:10.1037/a0032975

Verde, M. F., & Rotello, C. M. (2004). Strong memories obscure weak memories in associative recognition. *Psychonomic Bulletin & Review, 11,* 1062-1066. doi:10.3758/BF03196737

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35,* 254-262. doi:10.3758/BF03193446

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Memory & Cognition, 30,* 885-892. doi:10.3758/BF03195774

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1207-1241. doi:10.1037/a0020122

Waters, A. J., & Underwood, G. (1998). Eye movements in a simple music reading task: A study of expert and novice musicians. *Psychology of Music, 26,* 46-60. doi:10.1177/0305735698261005

Watkins, M. J., & Watkins, O. C. (1976). Cue-overload theory and the method of interpolated attributes. *Bulletin of the Psychonomic Society, 7,* 289-291. doi:10.3758/BF03337192

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 442-452. doi:10.1037/0278-7393.1.4.442

Weiner, B. (1968). Motivated forgetting and the study of repression. *Journal of Personality, 36,* 213-234. doi:10.1111/j.1467-6494.1968.tb01470.x

Weller, P. D., Anderson, M. C., Gómez-Ariza, C. J., & Bajo, M. T. (2013). On the status of cue independence as a criterion for memory inhibition: Evidence against the covert blocking hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition, 39,* 1232-1245. doi:10.1037/a0030335

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning & Verbal Behavior, 2,* 440-445. doi:10.1016/S0022-5371(63)80045-6

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21(12),* 1-20. doi:10.18637/jss.v021.i12

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40(1),* 1-29. doi:10.18637/jss.v040.i01

Wike, S. S., & Wike, E. L. (1970). *A test of the spew hypothesis using intralist repetition and a free-recall task. Psychonomic Science, 19,* 349-350. doi:10.3758/BF03328854

Williams, C. C., & Zacks, R. T. (2001). Is retrieval-induced forgetting an inhibitory process? *American Journal of Psychology, 114,* 329-354. doi:10.2307/1423685

Willis, G. B., & Underwood, B. J. (1983). A lack of interference effects in recognition memory. *Bulletin of the Psychonomic Society, 21,* 427-430. doi:10.3758/BF03329999

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language, 95,* 78-88. doi:10.1016/j.jml.2017.01.006

Wixted, J. T., Ghadisha, H., & Vera, R. (1997). Recall latency following pure- and mixed-strength lists: A direct test of the relative strength model of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 523-538. doi:10.1037/0278-7393.23.3.523

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review, 107,* 368-376. doi:10.1037/0033-295X.107.2.368

Wolk, D. A., Gold, C. A., Signoff, E. D., & Budson, A. E. (2009). Discrimination and reliance on conceptual fluency cues are inversely related in patients with mild Alzheimer's disease. *Neuropsychologia, 47,* 1865-1872. doi:10.1016/j.neuropsychologia.2009.02.029

Xue, G., Mei, L., Chen, C., Lu, Z., Poldrack, R., & Dong, Q. (2011). Spaced learning enhances subsequent recognition memory by reducing neural repetition

suppression. *Journal of Cognitive Neuroscience, 23,* 1624-1633. doi:10.1162/jocn.2010.21532

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. doi:10.1037/0278-7393.20.6.1341

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. doi:10.1006/jmla.2002.2864

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 345-355. doi:10.1037/0278-7393.18.2.345

Zechmeister, E. B. (1969). Orthographic distinctiveness. *Journal of Verbal Learning & Verbal Behavior, 8,* 754-761. doi:10.1016/S0022-5371(69)80040-X

Zimmerman, J. (1975). Free recall after self-paced study: A test of the attention explanation of the spacing effect. *American Journal of Psychology, 88,* 277-291. doi:10.2307/1421597

Tables

Table 4.1

Ratio of ratio (R_r) values reported in previous experiments using the mixed-pure paradigm.

Paper	Experiment	Strengthening Manipulation	R_r
Ratcliff et al. (1990)	1	Massed	0.88
Ratcliff et al. (1990)	2	Massed	1.10
Ratcliff et al. (1990)	3	Massed	0.93
Ratcliff et al. (1990)	4a	Massed	0.77
Ratcliff et al. (1990)	4b	Massed	0.80
Ratcliff et al. (1990)	5a	Spaced	1.03
Ratcliff et al. (1990)	5b	Massed	0.89
Ratcliff et al. (1990)	6a	Spaced	1.02
Ratcliff et al. (1990)	6b	Spaced	1.08
Ratcliff et al. (1990)	6c	Spaced	0.97
Ratcliff et al. (1992)	1	Massed	1.04
Ratcliff et al. (1992)	2	Spaced	1.21

Table 5.1

Means and standard deviations for hit and false-alarm rates in Experiment 1.

Probe	List type	Strength	List half	Mean	Standard error
target	mixed	weak	first	0.642	0.034
target	mixed	strong	second	0.668	0.038
target	pure	weak	first	0.618	0.039
target	pure	weak	second	0.579	0.039
target	pure	strong	first	0.761	0.030
target	pure	strong	second	0.701	0.039
distractor	mixed	weak	first	0.323	0.027
distractor	mixed	strong	second	0.326	0.030
distractor	pure	weak	first	0.342	0.036
distractor	pure	weak	second	0.310	0.032
distractor	pure	strong	first	0.275	0.027
distractor	pure	strong	second	0.310	0.038

Table 5.2

Means and standard errors for *d'* for Experiment 1.

List type	Strength	Test half	Mean	Standard error
mixed	weak	first	0.915	0.111
mixed	strong	second	0.997	0.117
pure	weak	first	0.814	0.141
pure	weak	second	0.801	0.120
pure	strong	first	1.488	0.141
pure	strong	second	1.163	0.139

Table 5.3

Means and standard deviations for hit and false-alarm rates in Experiment 2.

Probe	List type	Strength	List half	Mean	Standard error
target	mixed	strong	first	0.826	0.023
target	mixed	weak	second	0.583	0.043
target	pure	strong	first	0.761	0.028
target	pure	strong	second	0.710	0.037
target	pure	weak	first	0.700	0.033
target	pure	weak	second	0.628	0.041
distractor	mixed	strong	first	0.366	0.044
distractor	mixed	weak	second	0.393	0.048
distractor	pure	strong	first	0.352	0.040
distractor	pure	strong	second	0.365	0.046
distractor	pure	weak	first	0.318	0.036
distractor	pure	weak	second	0.342	0.040

Table 5.4

Means and standard errors for *d'* for Experiment 2.

List type	Strength	List half	Mean	Standard error
mixed	strong	first	1.437	0.131
mixed	weak	second	0.628	0.113
pure	strong	first	1.277	0.163
pure	strong	second	1.106	0.137
pure	weak	first	1.166	0.158
pure	weak	second	0.871	0.159

Table 5.5

Means and standard deviations for hit and false-alarm rtes in Experiment 3.

Probe	List type	Strength	Mean	Standard error
target	mixed	strong	0.679	0.034
target	mixed	weak	0.557	0.035
target	pure	strong	0.653	0.030
target	pure	weak	0.563	0.034
distractor	mixed		0.311	0.036
distractor	pure	strong	0.283	0.027
distractor	pure	weak	0.313	0.038

Table 5.6

Means and standard errors for *d'* for Experiment 3.

List type	Strength	Mean	Standard error
mixed	strong	1.108	0.139
mixed	weak	0.746	0.113
pure	strong	1.041	0.138
pure	weak	0.746	0.111

Table 6.1

Mean and median sampling attempts required for each output position. Simulated subjects completed 1000 sampling attempts regardless of outcome. If an item was not retrieved after 1000 sampling attempts, 1001 was recorded as its output position.

	Pure				Mixed			
Output	Weak		Strong		Weak		Strong	
Position	mean	median	mean	median	mean	median	mean	median
1	2.74	2.50	2.14	2.00	8.26	5.00	2.07	1.00
2	9.06	8.50	6.61	6.50	21.87	15.00	5.17	4.00
3	17.92	16.50	12.07	12.00	44.68	28.00	9.17	8.00
4	32.46	28.00	18.49	18.00	91.53	47.00	13.93	13.00
5	66.90	45.50	26.36	25.50	174.91	75.00	19.83	18.00
6	168.65	78.00	36.37	35.00	315.77	126.00	28.09	24.00
7	388.99	165.50	50.46	47.50	506.24	276.00	41.50	33.00
8	672.63	1001.00	73.66	64.00	698.56	1001.00	71.16	44.00
9	888.32	1001.00	128.65	90.00	851.67	1001.00	135.67	62.00
10	977.42	1001.00	270.31	138.50	945.16	1001.00	264.47	92.00
11	998.57	1001.00	539.28	566.50	987.40	1001.00	463.69	175.00
12	1000.83	1001.00	825.79	1001.00	998.35	1001.00	695.35	1001.00
13	1001.00	1001.00	971.78	1001.00	1000.84	1001.00	888.32	1001.00
14	1001.00	1001.00	999.36	1001.00	1001.00	1001.00	981.95	1001.00

Figures

Figure 4.1

The figure shows that REM predicts a word-frequency mirror effect (Glanzer & Adams, 1985, 1990). This is evident in the false-alarm rates: Simulated subjects had a higher false-alarm rate to high-frequency words than low-frequency words. The pattern is present in hit rates, although the effect is very small and difficult to see in the graph. It should be noted, however, that the hit-rate advantage for low-frequency words over high-frequency words is consistent across simulation runs. The results are based on 10000 simulated subjects completing recognition tests of 16 targets and 16 distractors, evenly divided between low- and high-frequency words. REM parameters were as follows: g_l = .325, g_h = .45, w = 20, u = .28, c = .7, g_{base} = .4, and *criterion* = 1.

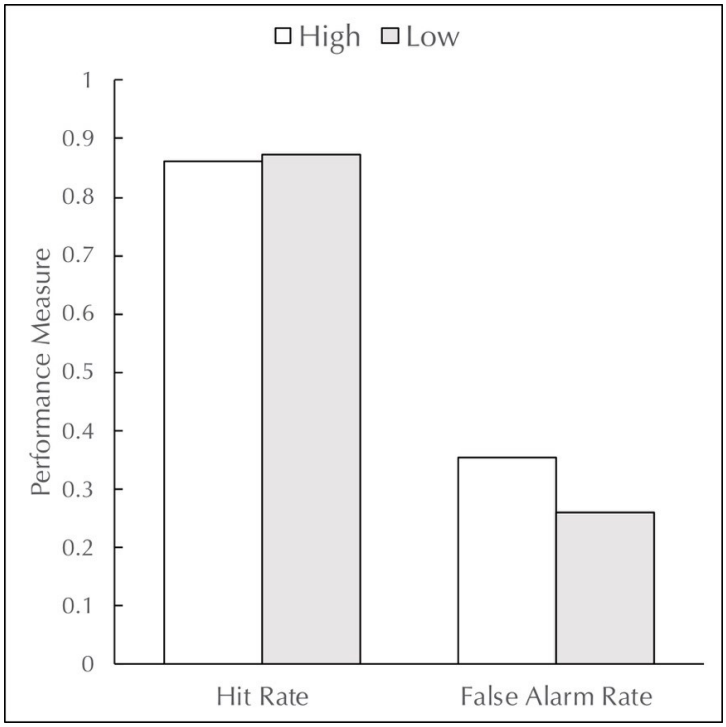

Figure 4.2

The figure shows that REM predicts a strength-based mirror effect (Stretch & Wixted, 1998). It is evident in the graph that the hit rate is higher for strong targets than weak targets, and that the false-alarm rate is higher for distractors on the weak test list than distractors on the strong test list. The results are based on 10000 simulated subjects, each of whom completed a weak list and a strong list (25 targets and 25 distractors per list). REM parameters were as follows: $u_{weak} = .28$, $u_{strong} = .4$, $w = 20$, $c = .7$, $g_{base} = .4$, $g_{draw} = .35$, and *criterion* = 1.

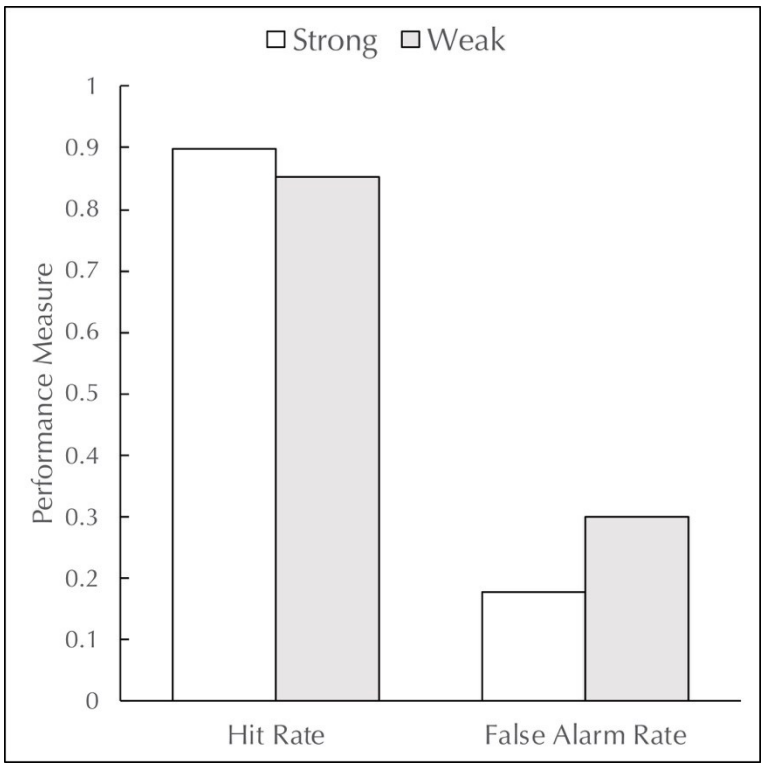

Figure 4.3

The figure shows that REM predicts a negative list-strength effect in recognition (Ratcliff et al., 1990). The graph plots *d′* as a function of item strength (weak vs. strong) and list type (pure vs. mixed). It is evident that discrimination is stronger for strong items than weak items, but that the magnitude of this difference is smaller in mixed lists than pure lists. It should be noted that, although the discrimination advantage for pure-strong items over mixed-strong items is small and difficult to see in the graph, it is consistent across simulation runs. The results are based on 10000 simulated subjects, each of whom completed a pure-weak list, a pure-strong list, and a mixed list (24 targets and 24 distractors per list). REM parameters were as follows: u_{weak} = .28, u_{strong} = .4, w = 20, g_{draw} = .35, g_{base} = .4, c = .7, and *criterion* = 1.

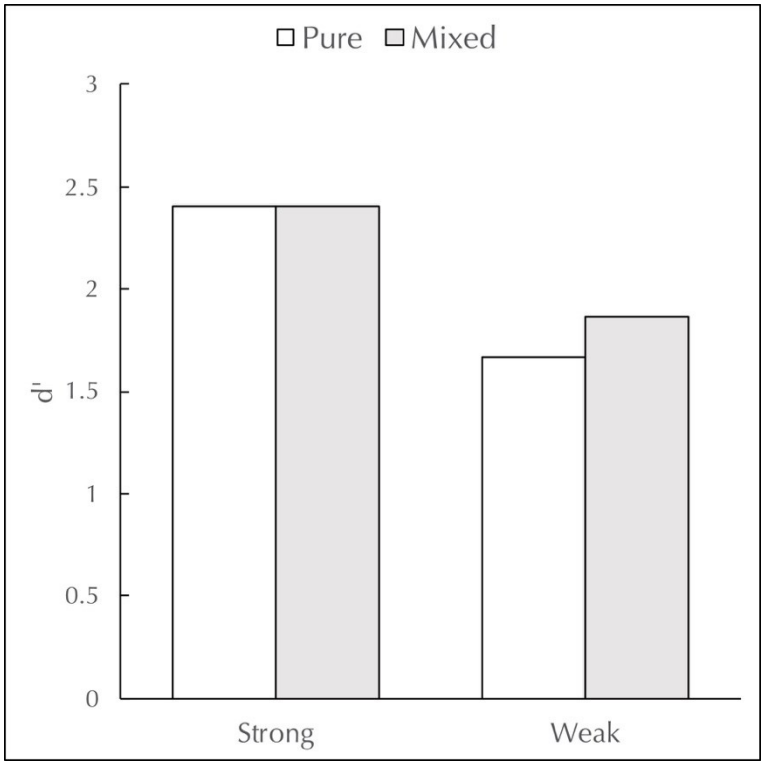

Figure 4.4

The figure shows that REM predicts a list-length effect in recognition (Gillund & Shiffrin, 1984; Strong, 1912). The graph plots *d'* as a function of list length (10, 20, 30, or 40 targets). It is evident that performance declines as list length increases. Results are based on 40000 simulated subjects, with 10000 in each list-length condition. REM parameters were as follows: $w = 20$, $g_{draw} = .35$, $g_{base} = .4$, $u = .28$, $c = .7$, and *criterion* = 1.

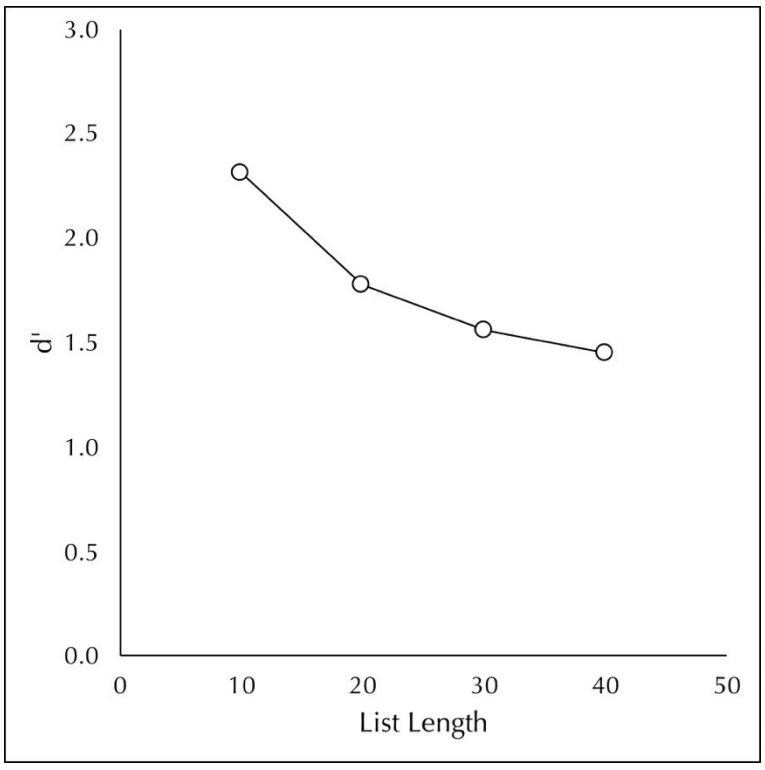

Figure 4.5

The graph plots the proportion of superimposition errors as a function of g_{draw} (.3, .35, .4, .45, and .5), list length (2, 4, 8, 16, 32, 64, 128, and 256), and $criterion_{study}$ (1, 2, and 3). It is evident that the probability of superimposition errors increases with list length and g_{draw}, and decreases as $criterion_{study}$ increases. Results are based on 120,000 simulated subjects, evenly divided among the cells of the between-subjects design. Other than g_{draw} and $criterion_{study}$, which were varied, REM parameters were as follows: $g_{base} = .4$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$.

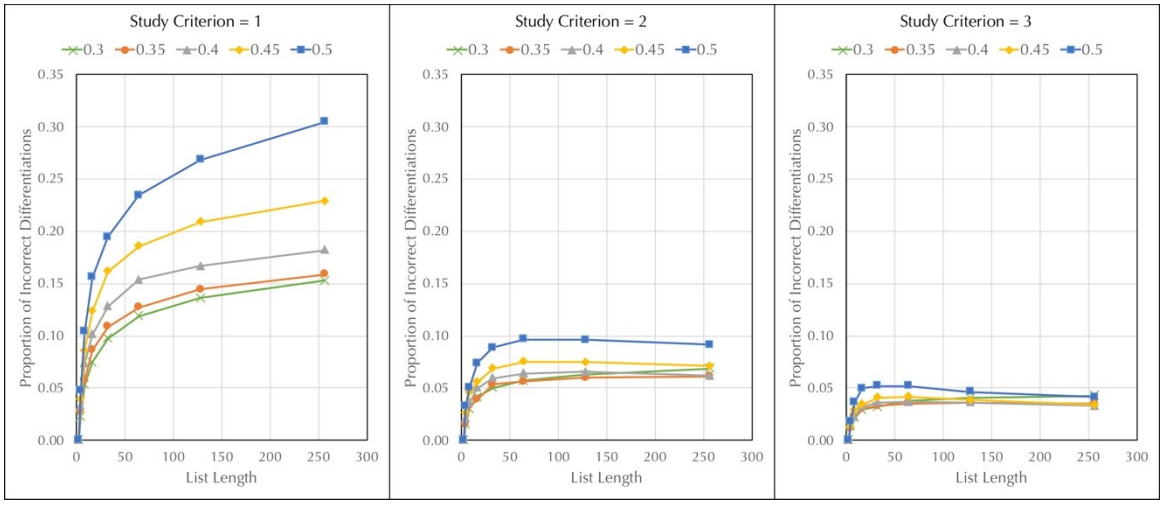

Figure 4.6

The figure shows that REM.3 predicts a discrimination advantage for spaced over massed items (Delaney et al., 2010). The graph plots *d′* as a function of strengthening method (massed vs. spaced), list length (16, 32, and 64), and $criterion_{study}$ (1, 2, and 3). It is evident that, regardless of list length, *d′* is higher for spaced items than massed items. Results are based on 30000 simulated subjects with strengthening method manipulated within list (i.e., massed and spaced repetitions on the same list) and list length between list. There were 10000 simulated subjects in each of the $criterion_{study}$ conditions. REM.3 parameters were fixed as follows: $g_{draw} = .35$, $g_{base} = .4$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$.

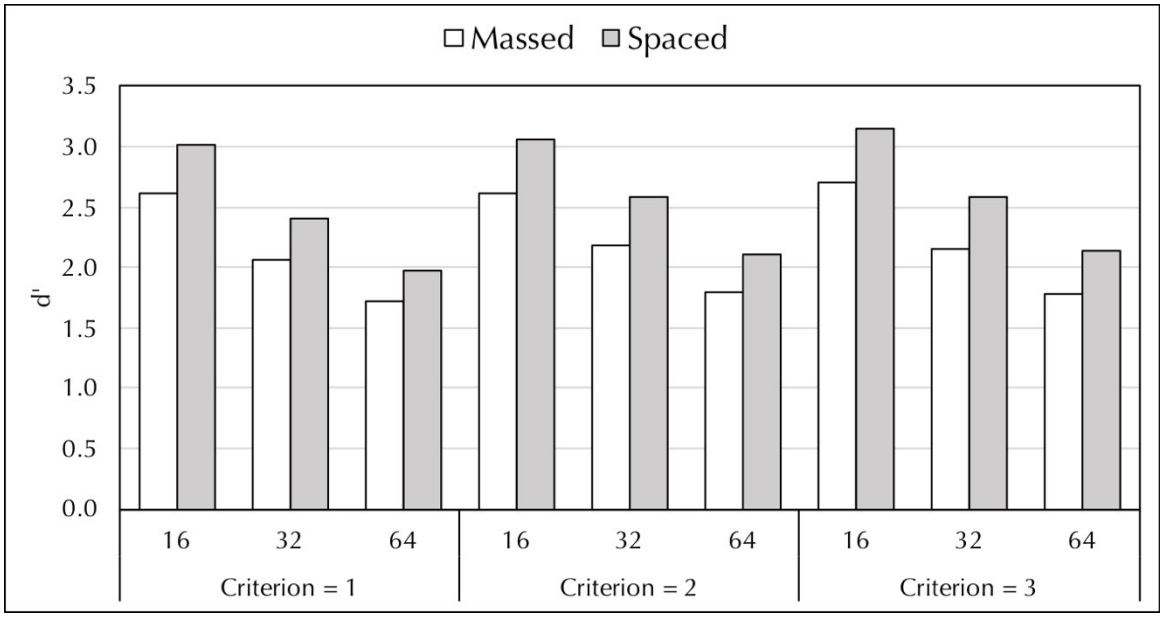

Figure 4.7

The graph plots d′ as a function of item strength (weak vs. strong), list type (pure vs. mixed), strengthening method (massed vs. spaced), strong item presentations (2, 3, 4, 5, and 6), list length (28 vs. 84), and $criterion_{study}$ (1, 2, and 3). It is clear that, across values of list length, degree of item strengthening, and $study_{criterion}$, REM.3 predicts a negative list-strength effect with massed strengthening and a positive list-strength effect with spaced strengthening. Results are based on 3000 simulated subjects, with simulated subjects evenly divided among cells of the design. Of the manipulated factors, list type, item strength, and strengthening technique were within-subject factors and the remaining were between-subjects factors. REM.3 parameters were set as follows: $g_{base} = .4$, $g_{draw} = .35$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$.

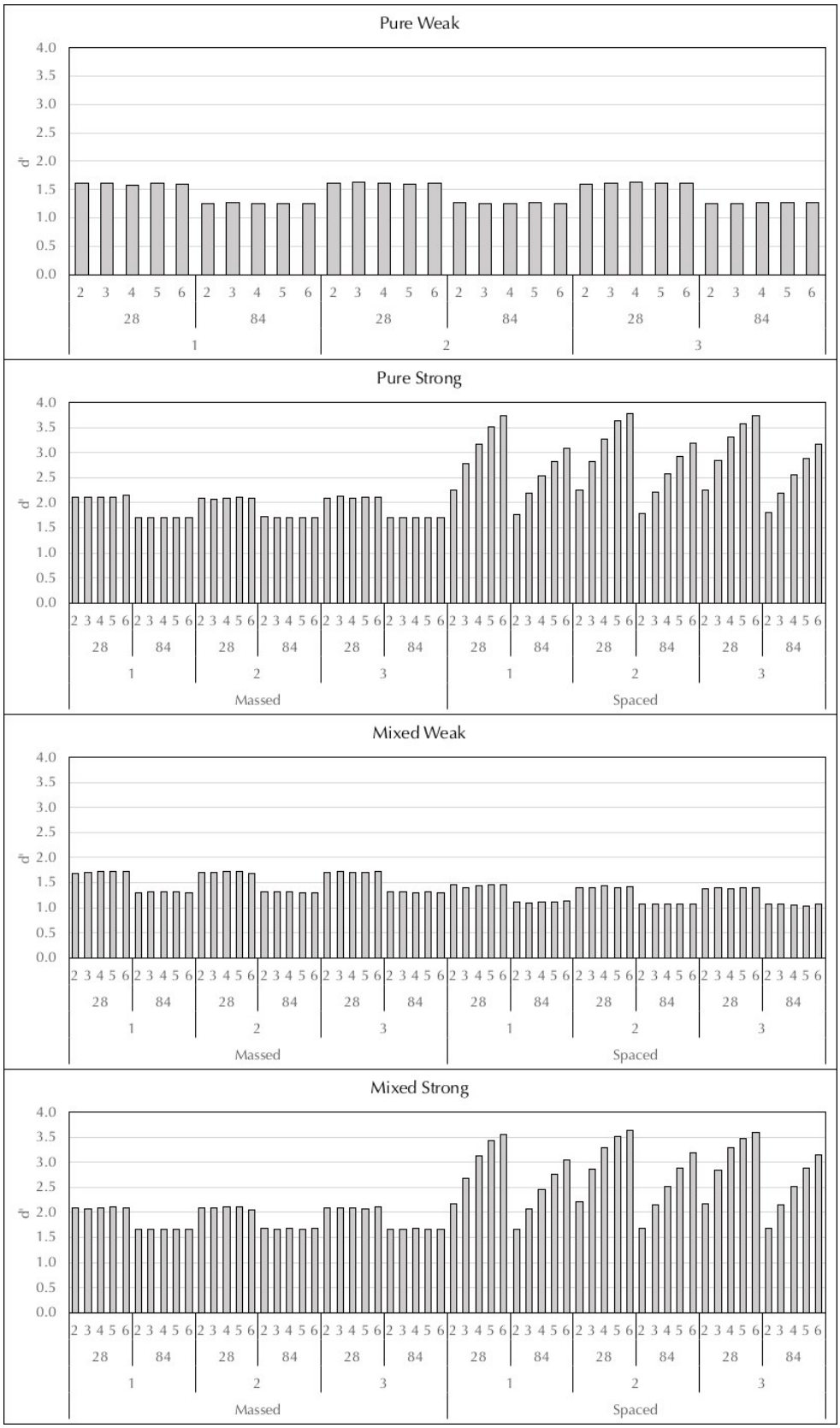
Pure Weak
Pure Strong
Mixed Weak
Mixed Strong
d'
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2 3 4 5 6
28
84
1
2
3
Massed
Spaced

Figure 4.8

The graph plots R_r as a function of item strength (weak vs. strong), list type (pure vs. mixed), strengthening method (massed vs. spaced), strong item presentations (2, 3, 4, 5, and 6), list length (28 vs. 84), and $criterion_{study}$ (1, 2, and 3). It is clear that, across values of list length, degree of item strengthening, and $study_{criterion}$, REM.3 predicts a negative list-strength effect with massed strengthening and a positive list-strength effect with spaced strengthening. Results are based on 30000 simulated subjects, with simulated subjects evenly divided among cells of the design. Of the manipulated factors, list type, item strength, and strengthening technique were within-subject factors and the remaining were between-subjects factors. REM.3 parameters were set as follows: $g_{base} = .4$, $g_{draw} = .35$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$.

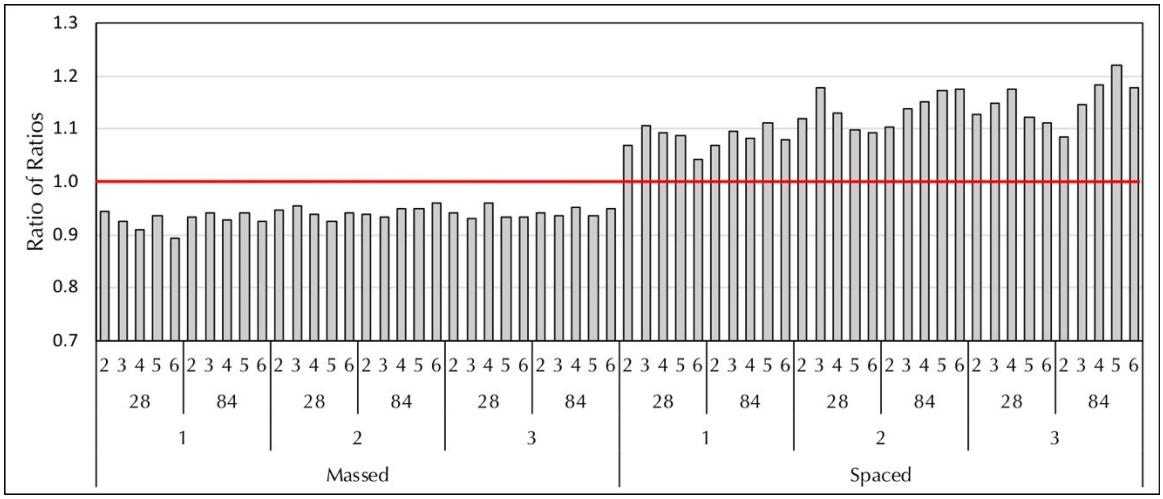

Figure 4.9

The graph plots hit rates, false-alarm rates, and *d′* as a function of u_{strong} (.4, .6, and .8) and list type (weak vs. strong interference). It is evident that, consistent with Norman (2002, Experiment 1), REM.1 predicts a concordant effect, with higher hit and false-alarm rates on the weak-interference list than the strong-interference list. However, REM.1 cannot predict the list-strength effect observed by Norman: Indeed, as u_{strong} increases, the list-strength effect becomes increasingly negative. Results are based on 3000 simulated subjects, evenly divided among the u_{strong} conditions. REM.1 parameters were set as follows: u_{weak} = .28, g_{draw} = .38, g_{base} = .4, w = 20, c = .7, and *criterion* = 1.

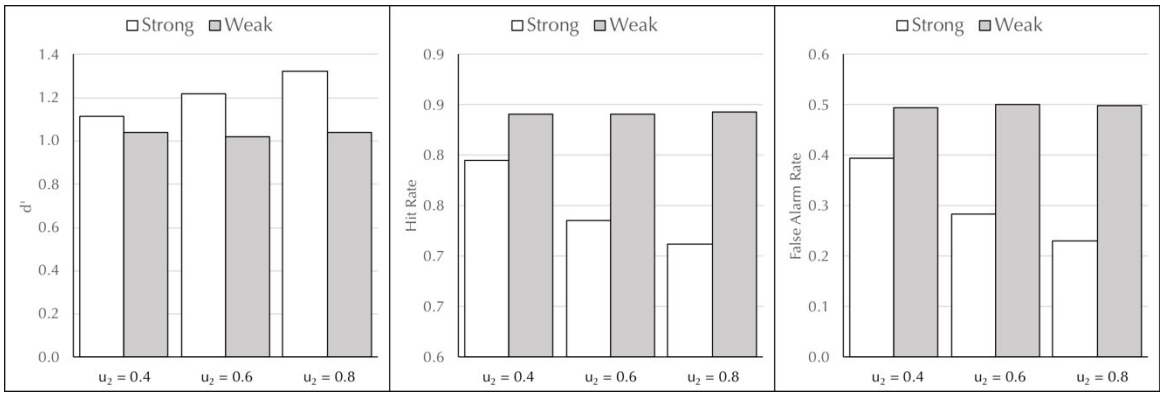

Figure 4.10

The graph plots hit rates, false-alarm rates, and *d′* as a function of $criterion_{study}$ (1, 2, 3, and 4). It is evident that, regardless of the value of $criterion_{study}$, REM.3 replicates the concordant effect documented by Norman (2002, Experiment 1). A positive list-strength effect is also predicted when $criterion_{study}$ is 2, 3, or 4. Results are based on 4000 simulated subjects, evenly divided among the $criterion_{study}$ conditions. REM.3 parameters were set as follows: $g_{draw} = .38$, $g_{base} = .4$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$.

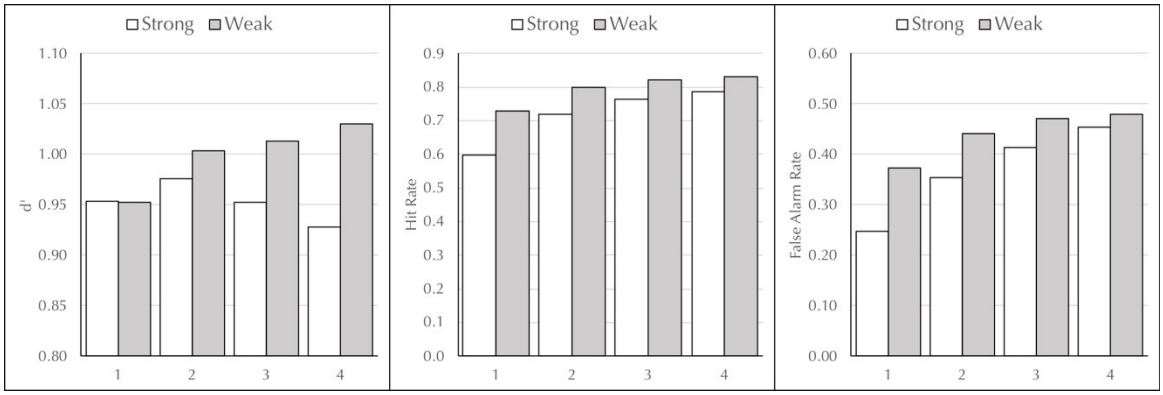

Figure 4.11

The graph plots hit rates, false-alarm rates, and *d'* as a function of $criterion_{study}$ (1, 2, 3, 4). Regardless of the value of $criterion_{study}$, there is a concordant effect. With $criterion_{study}$ values of 2, 3, or 4, Norman's (1999, Experiment 4) positive list-strength effect is replicated. When $criterion_{study}$ is 1, there is a null list-strength effect. Results are based on 4000 simulated subjects, evenly divided among $criterion_{study}$ conditions. REM.3 parameters were set as follows: $g_{draw} = .38$, $g_{base} = .4$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$.

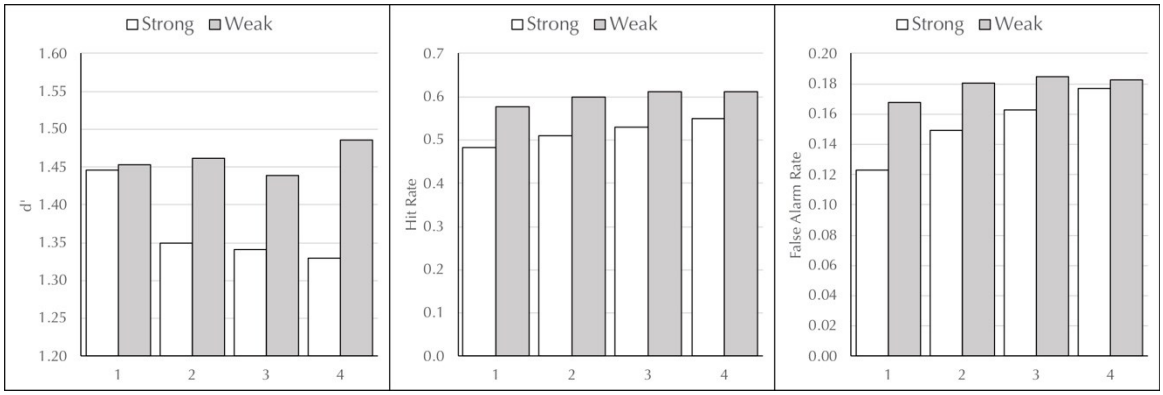

Figure 4.12

The graph plots hit rates, false-alarm rates, and *d'* as a function of $criterion_{study}$ (1, 2, 3, 4). Regardless of the value of $criterion_{study}$, there is a concordant effect. With $criterion_{study}$ values of 3 or 4, Norman's (1999, Experiment 4) positive list-strength effect is replicated. When $criterion_{study}$ is 1, there is a negative list-strength effect, and when $criterion_{study}$ is 2, there is a null list-strength effect. Results are based on 4000 simulated subjects, evenly divided among the $criterion_{study}$ conditions. To simulate semantic relatedness, distractors shared 8 features with a target. REM.3 parameters were set as follows: g_{draw} = .38, g_{base} = .4, w = 20, c = .7, u_1 = .28, u_2 = .12, and $criterion_{test}$ = 1.

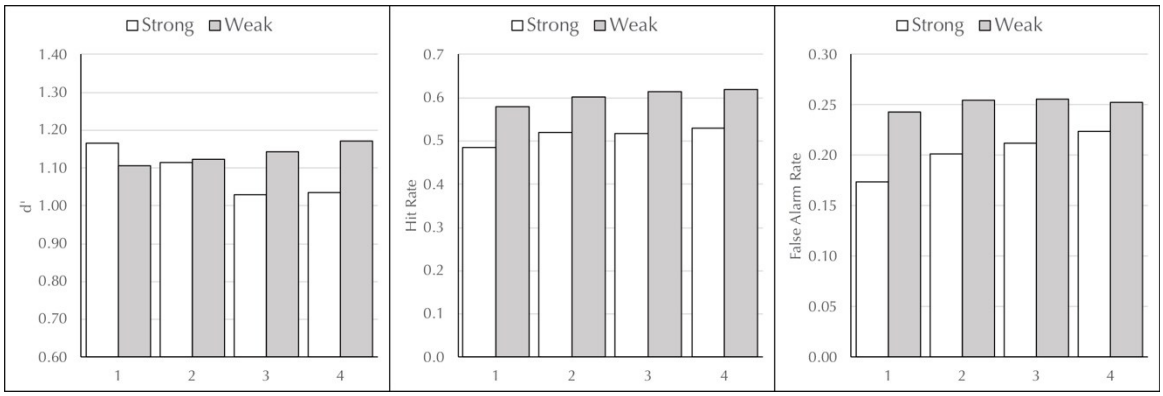

Figure 5.1

The graph shows hit and false-alarm rates as a function of item strength (weak vs. strong), list type (pure vs. mixed), list half (first vs. second), and $criterion_{study}$ (1, 2, and 3). It is evident from the figure that output interference was observed: The hit rate was higher and the false-alarm rate was lower in the first half of the pure test lists than the second half (it does not make sense to consider the mixed list here, since this is inherently confounded with target strength). There were 3000 simulated subjects, evenly divided among the $criterion_{study}$ conditions (the remaining factors were manipulated within subjects). REM parameters were set as follows: $g_{base} = .4$, $g_{draw} = .35$, $w = 20$, $u_1 = .28$, $u_2 = .12$, $c = .7$, and $criterion_{test} = 1$.

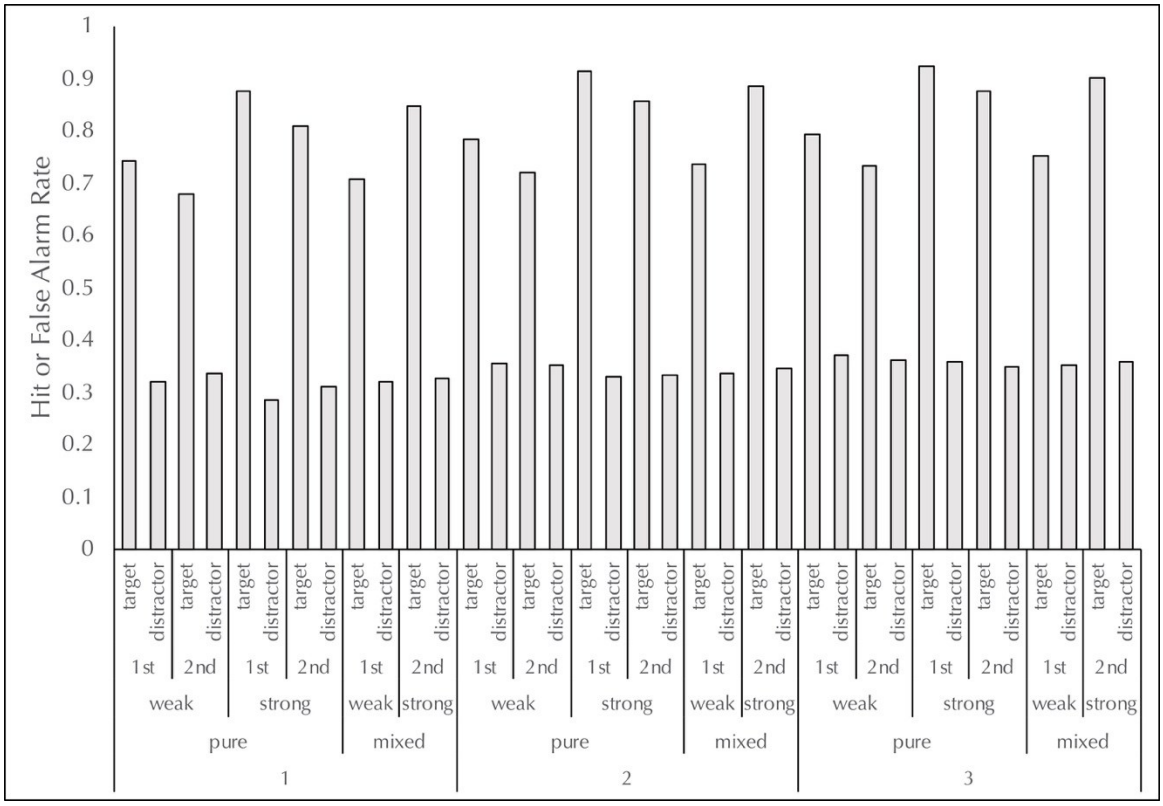

Figure 5.2

The graph shows *d'* as a function of item strength (weak vs. strong), list type (pure vs. mixed), list half (first vs. second), and $criterion_{study}$ (1, 2, and 3). It is evident from the figure that output interference was observed: *d'* was higher in the first half of the pure test lists than the second half (it does not make sense to consider the mixed list here, since this is inherently confounded with target strength). It is also clear that there is a positive list-strength effect when controlling list half (i.e., mixed-strong performance vs. second-half pure-strong performance and mixed-weak performance vs. first-half pure-weak performance), but not when considering pure lists without regard to list half. There were 3000 simulated subjects, evenly divided among the $criterion_{study}$ conditions (the remaining factors were manipulated within subjects). REM parameters were set as follows: $g_{base} = .4$, $g_{draw} = .35$, $w = 20$, $u_1 = .28$, $u_2 = .12$, $c = .7$, and $criterion_{test} = 1$.

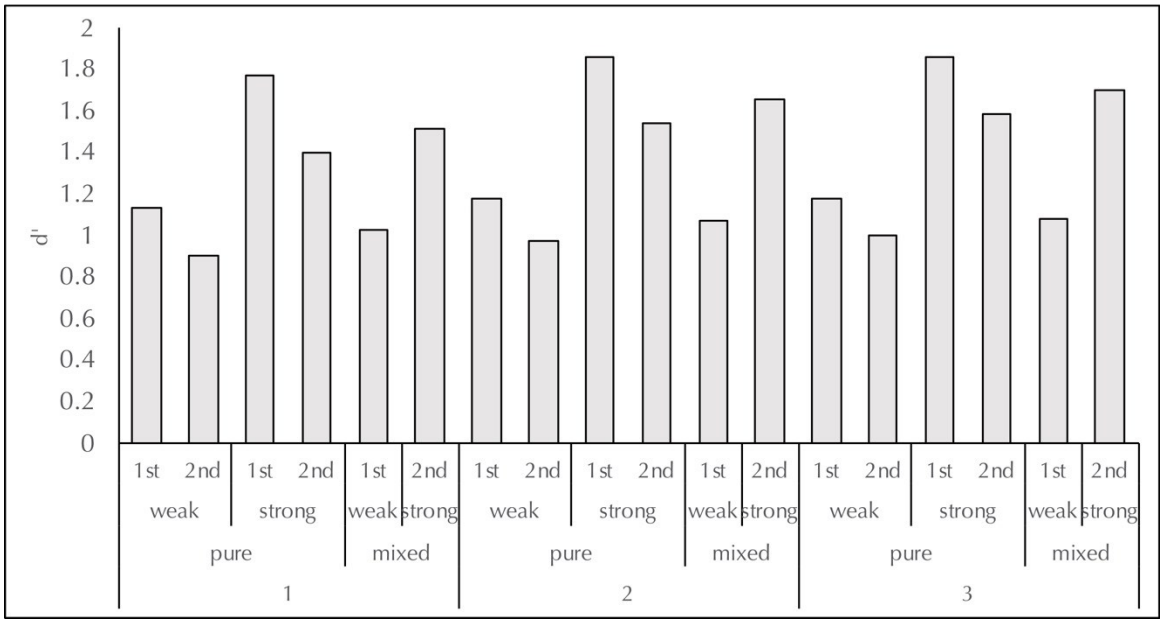

Figure 5.3

The graph shows hit and false-alarm rates as a function of item strength (weak vs. strong), list type (pure vs. mixed), list half (first vs. second), and $criterion_{study}$ (1, 2, and 3). It is evident that output interference was quite modest (see the main text for discussion). There were 3000 simulated subjects, evenly divided among the $criterion_{study}$ conditions (the remaining factors were manipulated within subjects). REM parameters were set as follows: $g_{base} = .4$, $g_{draw} = .35$, $w = 20$, $u_1 = .28$, $u_2 = .12$, $c = .7$, and $criterion_{test} = 1$.

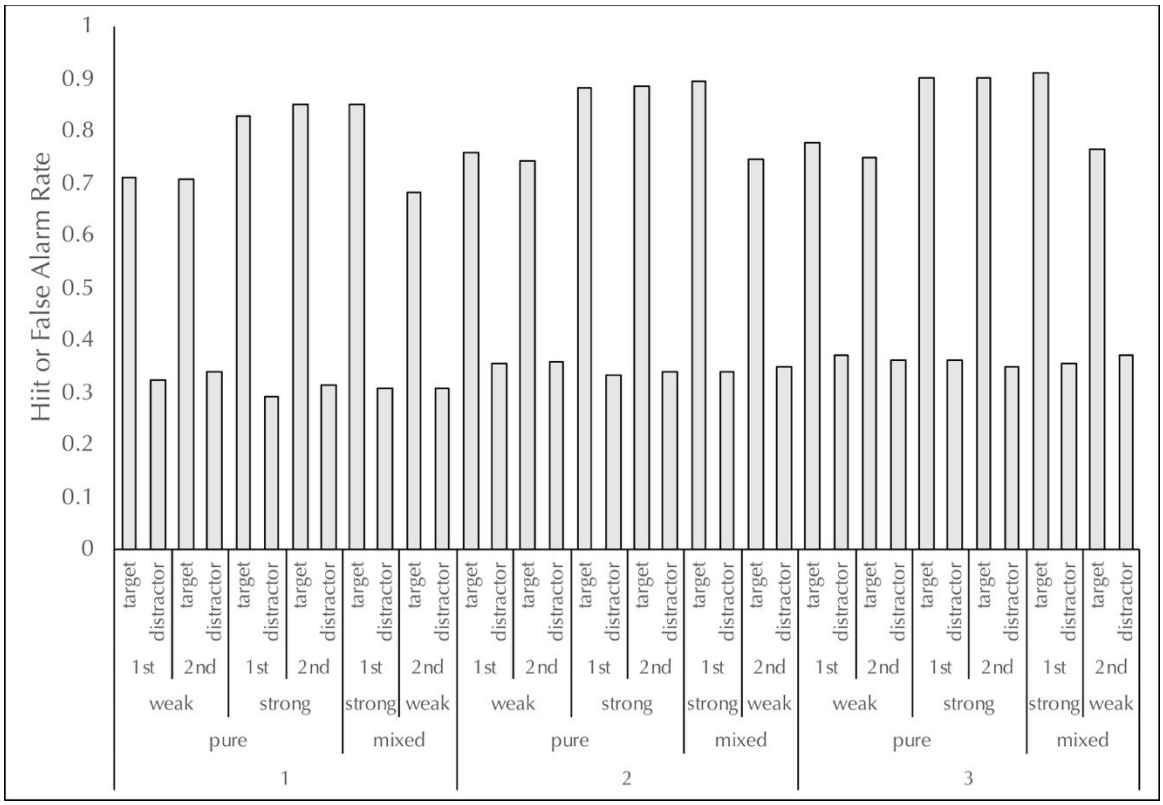

Figure 5.4

The graph shows *d'* as a function of item strength (weak vs. strong), list type (pure vs. mixed), list half (first vs. second), and $criterion_{study}$ (1, 2, and 3). It is clear that, when ignoring test position, there is a null list-strength effect. There is also a null list-strength effect when controlling test position (i.e., mixed-strong performance vs. first-half pure-strong performance and mixed-weak performance vs. second-half pure-weak performance). There were 3000 simulated subjects, evenly divided among the $criterion_{study}$ conditions (the remaining factors were manipulated within subjects). REM parameters were set as follows: $g_{base} = .4$, $g_{draw} = .35$, $w = 20$, $u_1 = .28$, $u_2 = .12$, $c = .7$, and $criterion_{test} = 1$.