# ON THE RELATION BETWEEN MEMORY AND METAMEMORY IN FREE RECALL: THE EFFECTS OF LIST LENGTH AND WORD FREQUENCY ON DUAL PROCESSES

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Arts

by

Carlos Falcao de Azevedo Gomes

August 2013

## Abstract

Subjective judgments about retrieval phenomenologies (remember/know) and memory strength (confidence) are often used in dual-process research to separate recollective from nonrecollective retrieval. Although such methods provide process-level explanations of the effects of list length and word frequency on recognition, whether the same applies to free recall is unknown. I compared subjective and objective methods of measuring dual processes and investigated their process-level explanations of the effects of list length and word frequency on free recall. Ninety-five undergraduates received multiple study-test trials and made retrospective judgments about items recalled on the last trial. A dual-retrieval Markov chain was used to quantify recollective and nonrecollective retrieval from subjects' recall performance rather than their metacognitive judgments to recalled items. In free recall, list length affected both recollective and nonrecollective retrieval, whereas word frequency only affected recollective retrieval. In addition, although remember judgments and correct source identification correlated with an objective measure of recollective retrieval (direct access), high confidence did not and it was unrelated to memory strength in free recall.

## Biographical Sketch

Carlos Gomes is a graduate student in the Human Development department at Cornell University, United States, since August/2011, and has a bachelor's degree in Psychology from the Pontifical Catholic University of Rio Grande do Sul, Brazil. He has been involved with psychological research since the beginning of his undergraduate education and has published journal articles and book chapters on memory research. As an undergraduate, he was awarded scholarships from the Brazilian National Council for Scientific and Technological Development for his work in false memory research and emotion. In 2012, he received a fellowship from the CAPES Foundation, a funding agency linked to the Brazilian Ministry of Education, to conduct research on the markers of impairment during aging. He is interested in four areas of psychological research, namely memory development, false memory, metacognition, and emotion. Of late, his work has focused on two main lines of investigation. The first regards the investigation of retrieval processes that predict transitions between healthy and non-healthy aging, while the second focuses on the relation between memory and metamemory. The content of this thesis is the result of the latter.

**Acknowledgments**

I would like to thank my current advisor and chairperson of my graduate committee, Dr Brainerd, for his continuing guidance during the making of this thesis and for all helpful commentaries on previous drafts. It has been both an honor and a pleasure to work with him for the last two years. I am also extremely thankful to the members of my graduate committee for taking the time to read and to talk about my thesis, and to the Cornell University and the Human Development department for supporting my research and for providing the facilities and materials I used in this study. They have all provided direct contribution to the elaboration of my thesis and made the research possible. For this reason, I am very thankful to them.

I have met and worked with many people in the past who indirectly contributed to this master's thesis. I offer my sincere and earnest gratitude to all of them. In particular, I am forever grateful to my first science mentor in psychology, Dr Stein, for instigating my curiosity and for providing invaluable advices throughout my undergraduate education. I thank all funding agencies who have sponsored me in the past, especially the Brazilian National Council for Scientific and Technological Development, for supporting research projects who gave me enough experience to conduct this study, and to the members of all research labs I have worked so far. Finally, I would like to thank both my parents for their unconditional love, for believing in me and my work, and for supporting my decisions.

## Table of Contents

## List of Figures

## List of Tables

Metacognitive judgments about memory can be separated into prospective judgments (e.g., ease of learning, judgments of learning, and feelings of knowing) and retrospective judgments (e.g., confidence judgments, and judgments of remembering and knowing). This study addressed the latter form of judgment. Specifically, the aim of the present study was to investigate the relationship between memory and retrospective metamemory judgments, which is motivated by applied as well as theoretical considerations. On the applied side, retrospective judgments about memory are often used in eyewitness identification procedures in the United States (Wells et al., 1998; Keast, Brewer, & Wells, 2007), predicated on the idea that such judgments can provide reliable information about memory accuracy. Eyewitnesses to crime scenes, for instance, are often asked by law enforcement officers to identify a culprit in a lineup and to make a confidence judgment about their decision. In the same vein, assessments of the credibility of an eyewitness' testimony are often influenced by changes in the degree of certainty about the information reported (Brewer & Burke, 2002; Whitley & Greenberg, 1986). On the theoretical side, retrospective judgments about memory play a key role in many theories and measurement models of memory. In the recognition memory literature, for instance, confidence judgments are often assumed to be a proxy of memory strength (e.g., Wixted, 2007; Yonelinas, 1999). Similarly, remember/know judgments and source judgments are often used on the assumption that they can distinguish between different types of retrieval processes, such as recollection and familiarity (Yonelinas & Jacoby, 1995), or between different memory systems, such as episodic and semantic (Tulving, 1985).

However, the extent to which retrospective metamemory judgments tap the assumed theoretical operations is controversial (Strack & Förster, 1995; Wells, Lindsay, & Ferguson, 1979). Confidence judgments about recall accuracy have been shown to be influenced by

several factors other than memory strength, such as the amount of information retrieved (Koriat, Lichtenstein, & Fischnoff, 1980), the vividness of the information recalled (Robinson & Johnson, 1996), the degree of familiarity with an item's theme (Chandler, 1994), subjects' chronological age (Koriat & Ackerman, 2011), and the amount of time to provide a response to a question (Kelley & Lindsay, 1993; Nelson et al., 1990). In connection with that, Wells, Rydell, and Seelau (1993) showed that confidence in identification accuracy can be influenced by the degree to which members of a lineup fit the culprit's description, regardless of actual accuracy in identification. Subjects in Wells et al.'s experiment witnessed a staged theft and were then asked to give a description of the culprit. Next, subjects were presented with one of three photospread conditions in which they were asked to say whether the culprit was present in the photospread or not and to rate their confidence in their judgment. In the mismatch-description condition, innocent members of the lineup violated at least one major feature of the culprit's description reported by the subject. In the resemble-culprit condition, innocent members of the lineup resembled the culprit in several features. In the match-description condition, none of the innocent members of the lineup violated the culprit's description reported by the subject. Regardless of whether the identification was true or false, confidence in identification accuracy decreased from the mismatch-description condition to both the resemble-culprit condition and the match-description condition.

In fact, the notion that people have direct access to the strength of memory traces, even of memories not consciously available (Hart, 1967), has received little support in the metacognition literature (for a review, see Schwartz, Benjamin, & Bjork, 1997). Benjamin, Bjork, and Schwartz (1998), for example, showed that subjects' prediction about future memory performance can be inversely related to memory strength under certain conditions. In their

experiment, subjects were instructed to answer twenty general-knowledge questions such as "what is the name of the horse-like animal with black and white stripes?". After each answer, subjects were instructed to give the probability that they would be able to free recall the answer again in twenty minutes. Ten minutes later, subjects performed a free recall test for the answers provided in the first part of the experiment (e.g., zebra), which revealed two main results, both shown in Figure 1. First, answers to general knowledge questions that took a long time to output in the beginning of the experiment (hard answers) were better recalled than the ones that were promptly output (easy answers), a finding that was interpreted as the result of deeper encoding of hard answers relative to easy answers. More specifically, the increased time searching in semantic memory for answers to hard questions enhanced the strength of the episodic traces of those answers. Consequently, if the direct access hypothesis is correct, predicted recall performance should follow the same direction as actual recall performance. However, the second result showed the exact opposite pattern, namely predicted recall decreased from easy to hard answers.

## Retrospective Judgments about Memory

Despite the massive evidence that metamemory judgments are influenced by factors other than memory itself (Koriat, 2002; Nelson & Narens, 1990), recognition and recall are often supplemented by retrospective judgments about memory as a means of quantifying latent memory operations (Tulving, 1985; Wixted, 2007; Yonelinas, 1994, 1999). Next, we describe three such retrospective judgments about memory: remember/know, confidence, and source judgments.

### Judgments of remembering and knowing

Tulving's (1985) remember/know procedure was an early attempt to characterize and to measure distinct phenomenologies that are often induced by retrieval, such as when remembering the name of a wine drank yesterday is followed by a vivid recollection of its label and bottle, as opposed to remembering the name of it without such experience. In this procedure, subjects study a list of items and then perform a recall or recognition test. In an old/new recognition test, old decisions to test probes are supplemented by judgments of remembering and knowing, while in a recall test, remember/know judgments are made as items are output or after recall of all items. "Remember" judgments are associated with an autonoetic state of conscious awareness about an item's previous occurrence, in which its features are consciously re-experienced during retrieval. In the previous example, a remember judgment would characterize the retrieval of a wine's name accompanied by recollection of its label and bottle. "Know" judgments, on the other hand, are associated with a noetic state of conscious awareness about an item's previous occurrence, in which subjects have knowledge about its occurrence but do not consciously re-experience it during retrieval (e.g., retrieval of the wine's name without recollecting any distinctive feature of it or the context in which it was seen). Therefore, subjects' responses on both recognition and recall tests can be partitioned into "remembered" and "known" responses, or recollective and nonrecollective retrieval, respectively. In two experiments, Tulving found that (a) free recall of a list of 27 names was more often "remembered" than "known" (88% of the items recalled received a remember judgment), (b) remember judgments decreased after long as opposed to short delays between study and test (the bias-corrected rate of remember judgments for items recognized as old was 38% immediately after study and 15% after a seven-day delay), and (c) confidence in recognition response was

higher for "remembered" items than for "known" items (the average confidence rating on a 3-point confidence scale for "remembered" items was 2.7, while for "known" items it was 2.1).

In its original conception, the two states of consciousness that support judgments of remembering and knowing were assumed to be closely connected to episodic and semantic memory, respectively. Of late, however, remember/know judgments have been assumed to underlie different memory processes, such as recollection and familiarity (Yonelinas, 1994, 2001). The validity of the latter mapping, however, has been controversial (Donaldson, 1996; Dunn, 2004, 2008). In the recognition memory literature, Donaldson (1996) argued that remember/know data can be accommodated by a one-process signal detection model that allows participants to set different decision criteria for remember and know responses, as opposed to using different retrieval processes, thus casting doubt on the ability of remember/know judgments to separate recollective from nonrecollective retrieval.

Interestingly, although the recollective experience that characterizes remember judgments seems to be indicative of an event's past occurrence, this idea has not received support. Multiple experiments have shown that recollective experiences can occur even during retrieval of memories about events that have never happened (Brainerd, Payne, Wright, & Reyna, 2003; Geraci & McCabe, 2006; McCabe, Roediger, McDaniel, & Balota, 2009; Roediger & McDermott, 1995). Subjects in Roediger and McDermott's (1995) experiment studied lists of semantically associated words (e.g., *note, sound, piano, sing*) and then performed an old/new recognition test, which was composed of targets (*sound*), new related words (*music*), and new unrelated words (*apple*), supplemented by judgments of remembering and knowing. The results showed that remember judgments occurred as often for targets (the bias-corrected rate was equal to 39%) as for new related words (35%).

Therefore, although remember judgments capture a particular retrieval phenomenology, such experience seems to be neither diagnostic of memory accuracy nor of the type of underlying memory representation. The generality of this conclusion, however, is weakened by the fact that it relies largely upon findings from recognition memory studies. Even though the remember/know procedure was also initially used in paradigms other than recognition (e.g., free and cued recall) (Tulving, 1985), its subsequent use has been almost exclusively with recognition—the few exceptions include the studies conducted by Hamilton and Rajaram (2003) and McDermott (2006). From a theoretical perspective, this represents a major restriction as the retrieval operations that take place during recognition are not necessarily the same as the ones that operate in recall (Crowder, 1976; Brainerd & Reyna, 2010). The extent to which judgments of remembering and knowing tap recollective and nonrecollective processes in recall, respectively, is largely unknown because previous studies have not addressed this issue.

**Confidence judgments**

Retrospective confidence judgment is another type of metamemory judgment that has figured in the memory literature. In contrast to remember/know judgments, confidence judgments have been used in psychological research for more than a century (Bernbach, 1967; Egan, 1959; Fullerton & Cattell, 1892; Henmon, 1911; Hollingworth, 1913). But similar to remember/know judgments, confidence was initially used as a method of quantifying subjects' introspection about their mental processes (Metcalf, 1917).

In a standard old/new recognition test supplemented by confidence judgments, subjects are presented with a set of items (e.g., *tea, book, clock,* and *table*) and then are asked to rate old items (*book, table*) and new ones (*green, monkey*) on a old/new confidence scale (e.g., on a 6-point scale, 1 = sure new, ..., 6 = sure old). In a free recall experiment, subjects usually rate their

confidence about their response (i.e., whether the response is correct or incorrect) (e.g., on a 6-point scale, 1 = sure incorrect, ..., 6 = sure correct) rather than their confidence about the episodic state of an item (i.e., whether it is old or new), as in the recognition example. However, although retrospective confidence judgments about subjects' response can also be made in old/new recognition experiments (e.g., How confident you are about your old/new decision?, 1 = not confident at all, 2 = moderately confident, 3 = very confident), such distinction has not always been acknowledged (Banks, 1970; Baranski & Petrusic, 1998; Fullerton & Cattell, 1892; Rabin & Cain, 1984; but see Higham, Perfect, & Bruno, 2009) and has led to confusion in the past (Healy & Jones, 1973; Lockhart & Murdock, 1970).

**Confidence as a proxy for memory strength.** In theories that draw upon signal detection analogies (Mickes, Hwe, Wais, & Wixted, 2011; Yonelinas, 1994; Wixted, 2007), confidence judgments have been regarded as a method of obtaining information about subjects' response criterion placement across one or more memory strength dimensions (Macmillan & Creelman, 2005). Although these theories differ in many respects (e.g., whether memory strength reflects the contribution of one or two memory processes, or a pure or aggregated signal), they rely on the common assumption that confidence is a proxy of strength. For example, recognition of an item with a high degree of certainty is assumed to reflect a strong memory for the item, whereas recognition of an item with a low degree of certainty is assumed to reflect a weak memory for the item. This assumption, however, seems disconnected from experiments showing that retrospective confidence judgments are not sampled directly from memory strength (Chandler, 1994; Nelson et al., 1990; Van Zandt, 2000; Wells et al., 1979, 1993), the evidence of violations of such assumption being particularly compelling when it comes to confidence in

recall (Kelley & Lindsay, 1993; Koriat et al., 1980; Koriat & Ackerman, 2011; Robinson & Johnson, 1996).

In a standard free recall experiment, subjects are presented with a set of stimuli and are asked to recall the studied items in any order. Although simple, such paradigm offers ground for testing a counterintuitive prediction about the relation between memory strength and retrospective confidence in recall that has not been previously investigated, namely the idea that subjects can assign higher confidence judgments to weak relative to strong episodic memories. This idea is predicted by fuzzy-trace theory (Brainerd & Reyna, 1990) based on cognitive triage effects (Brainerd, Reyna, Howe, & Kevershan, 1991) and findings about the relation among confidence, retrieval latency, and output position in free recall (Kelley & Lindsay, 1993; Jou, 2008; Rohrer & Wixted, 1994).

The cognitive triage effect is a recall output pattern in which the output position of recalled targets is organized in a weak → strong → weak item strength fashion, forming an inverted-U relation between output position and memory strength, as illustrated in Figure 2. This effect is usually investigated by plotting a behavioral index of memory strength, namely the mean total number of errors on previous tests (MTE), as a function of output position (e.g., Marche, Howe, Lance, Owre, & Briere, 2009). The triage process maximizes recall, which is the goal of free recall tasks, by minimizing output interference and maximizing episodic activation (for a review, see Brainerd et al., 1991). Similar to MTE, confidence is often regarded as a proxy for memory strength (Mickes et al., 2011; Yonelinas, 1994; Wixted, 2007), but in contrast to MTE, it is a metacognitive marker of memory strength that has been shown to be highly influenced by retrieval latency (items that can be quickly retrieved are judged to be highly accurate; Kelley & Lindsay, 1993), which increases as a function of output position in free recall

tests (fast → slow latencies) (Rohrer & Wixted, 1994). In connection with that, Jou (2008) found that confidence in recall is a monotonic decreasing function of output position. Of course, such data do not elucidate whether confidence judgments are assigned based on retrieval latency or output position, or both, as the latter two variables are also correlated with each other. Nevertheless, the implication of such a finding is that subjects will make high confidence judgments to weak memories relative to strong memories during the beginning and middle of the free recall protocol whenever cognitive triage effects are observed, as indicated by the MTE statistic. This prediction, which is illustrated in Panel A of Figure 3, follows from the notion that subjects *do not have* access to the strength of memory traces (Schwartz et al., 1997) and, therefore, are not aware of the effects of cognitive triage on output position. An alternative hypothesis, illustrated in Panel B of Figure 3, is that subjects *have* access to the strength of memory traces and, therefore, can monitor cognitive triage. As a result of the second prediction, retrospective confidence judgments will also show a triage pattern when cognitive triage effects are observed in free recall.

**Confidence as a method of measuring dual processes.** As in memory research using the remember/know procedure, confidence ratings have also been used to estimate dual processes in recognition, such as recollection and familiarity (Yonelinas, 1994; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). One frequently used procedure consists of constructing receiver operating characteristic (ROC) curves in order to estimate the intercept (recollection) and deflection of the curve (familiarity, $d'$) that best fits the observed values of the hit rate (acceptance of targets), $P(H)$, and the false alarm rate (acceptance of new items), $P(FA)$, across confidence ratings. Figure 4, for example, shows a hypothetical ROC curve in which + signs indicate levels of confidence that an item was studied (old) and – signs indicate levels of

confidence that an item was not studied (new). Notice that when the false alarm rate = 0, the hit rate = .27, which is the estimate of recollection according to Yonelinas' (1994) signal detection dual-process model. Indeed, estimates of recollective retrieval in this model are highly influenced by extreme confidence ratings that an item was presented. In particular, notice that recollection = $P(H \wedge +++)$ as $P(FA \wedge +++) \rightarrow 0$, and as long as the ability to discriminate old from new items is not below chance ($d' \geq 0$), recollection < $P(H \wedge +++)$ as $P(FA \wedge +++) \rightarrow P(H \wedge +++)$ but recollection > $P(H \wedge +++)$ can never occur according to the model. This indicates that correct recognition with extreme high confidence is a good proxy of recollection when the false alarm rate with extreme high confidence is low, but it tends to over-estimate it as the same false alarm rate increases and it can never underestimate recollection. Familiarity, on the other hand, reflects subjects' ability to discriminate old from new items, as measured in $d'$ (*SD* units) (Macmillan & Creelman, 2005). If subjects cannot discriminate old from new items, then familiarity ($d'$) = 0 and the ROC curve would be a straight line intercepting the hit-axis at the level of recollection. As subjects begin to better discriminate old from new items, the ROC curve starts bending towards the (hits, false alarms) = (1, 0) coordinate and, therefore, familiarity > 0. In Figure 4, for example, familiarity = 2.

Prior studies have shown that markers of recollective retrieval, such as remember judgments and correct source identification, are much higher when subjects are extremely confident that a test probe is a target than when subjects are not as confident. Yonelinas (2001), for example, showed that 94% of the remembered items in a standard single-trial recognition experiment were recognized with the highest level of confidence ("sure [the test probe] was old"). In the same vein, subjects in Yonelinas' (1999) study were presented with two lists and then asked to make confidence judgments about test probes (1 = "sure it was new", ..., 6 = "sure it

was old") that were either targets from either list or new items. In addition, for items recognized as old, subjects were asked to make source judgments regarding whether the items were presented in list 1 or list 2. Source accuracy was only reliably higher than chance when subjects were very confident that the test probe was a target—that is, when they gave a 6 confidence rating.

**Source judgments**

In everyday life, information that we are asked to recall or recognize (e.g., "Did you watch the Oscars on Sunday?") might sometimes have occurred in different contexts ("Yes, I was at my home" or "Yes, I was at my girlfriend's home"). Similarly, memory researchers often present a focal list of items with different sources, such as color (e.g., presented in red or blue), gender (said by a male or a female), and font (presented in Arial or Times New Roman), and later ask subjects to make retrospective judgments about the source of recalled or recognized items ("Was it presented in red or blue?"). After test, subjects' responses are compared to the studied items' actual source, in which subjects' ability to discriminate between correct and incorrect sources is referred to as source accuracy.

Over the last two decades, psychological research on the relationship between item and source memory, constructs in which the latter is usually regarded as dependent on the former, has proceeded in two ways. In one, theories and measurement models have been developed to explain, and more rarely to predict, the relationship between item and source memory (Brainerd, Reyna, Holliday, & Nakamura, 2012; DeCarlo, 2003; Klauer & Kellen, 2010; Starns, Hicks, Brown, & Martin, 2008). Batchelder and Riefer's (1990) source memory model is a prominent example of this approach. Let source $j \in \Theta$, in which $\Theta$ is the set of all sources in the experiment, then in a standard old/new recognition experiment in which subjects' item responses

are supplemented by judgments about its source, the model posits that correct item and source *j* recognition based on *memories of them* occurs with probability $D_j d_j$. However, the model assumes that source *j* recognition from memory ($d_j$) cannot occur without item memory ($D_j$), although it can be *guessed* in two ways, namely when source rather than item memory fails with probability $D_j(1 - d_j)a$ and when item memory fails with probability $(1 - D_j)bg$. Item memory is then a pre-requisite to source memory in this model. However, recent evidence suggests that, although very intuitive, this assumption does not always hold. For example, Starns et al. (2008) have shown that subjects can recognize sources above chance performance even when they are unable to correctly recognize targets. In addition, Brainerd et al. (2012) demonstrated that Batchelder and Riefer's source memory model can neither fit over-distribution data in source memory experiments—non-zero probabilities of an item occupying mutually exclusive states, such as presented in List 1 *and* 2 when items are never presented in both lists—nor provide a theoretical account as to why manipulations that are known to affect memory, such as concreteness, list order, and word frequency, also affect over-distribution.

In another vein of research, source judgments are assumed to rely on recollective retrieval (Yonelinas, 1999) or used as direct measures of it (Wais, Mickes, & Wixted, 2008; Wais, Squire, & Wixted, 2010; Wilding & Rugg, 1996). In the latter case, source accuracy becomes an alternative to metacognitive judgments (e.g., judgments of remembering and very high confidence). However, contrary to remember/know and confidence judgments, source accuracy is an objective measure of retrieval of contextual information, because subjects' responses can be compared to items' *actual* source—whether a studied item is presented in red or blue, for example, is known and manipulated by the experimenter. The items' actual state of remember/know and confidence, on the other hand, cannot be objectively assessed with

retrospective metacognitive judgments because of the intrinsic subjective nature of them. To the best of my knowledge, however, there is not a single recall study that compared estimates of recollective retrieval using source judgments against estimates obtained with metacognitive judgments. Therefore, whether the two produce similar estimates in recall is an empirical question.

**Summary**

Remember/know, confidence, and source judgments have been widely used as methods of measuring latent memory variables. Remember judgments are used as direct proxies of recollective retrieval. Confidence ratings are assumed to have a direct mapping with memory strength and used to construct ROC curves, in which the intercept of the latter is often thought to be a measure of recollective retrieval. Correct source judgments are regarded as objective measures of recollective retrieval because, contrary to remember/know and confidence judgments, subjects' source responses can be compared with the targets' actual source (contextual information). Despite differences in definition, in the type of instructions provided to subjects, and in the method of collecting each type of retrospective judgment, the reviewed judgments about memory are frequently used interchangeably in dual-process research predicated on the assumption that they tap similar concepts. This is surprising because the evidence supporting this assumption is at best scarce, is restricted to recognition memory experiments, and has not been consistent across studies (Martin et al., 2011; Yonelinas, 2001; Yonelinas et al., 1996). In the case of remember/know judgments, which is by far the most widely used method of measuring dual processes, Migo, Mayes, and Montaldi (2012) have argued that correct source and remember judgments should not be used interchangeably because remember judgments are often supported by noncriterial recollection (e.g., emotions and

impressions) that do not support correct source judgments. In addition, Naveh-Benjamin and Kilb (2012) have indicated that the very use of remember/know judgments can interfere with the memory task and produce changes in subjects' performance.

## The Present Study

There is compelling evidence that retrospective metacognitive judgments can be affected by factors other than memory itself. This poses obvious challenges to measurement models that make strong assumptions about the mapping between metacognitive judgments and memory (Tulving, 1985; Wixted, 2007; Yonelinas, 1999; Yonelinas & Jacoby, 1995) and it leaves a theoretical gap in our understanding about the relationship between memory and metamemory, and how to measure memory processes as well. In the present study, I investigated this relationship by implementing a free recall paradigm that allowed the quantification of memory processes and strength in two independent ways, namely from retrospective judgments, such as remember/know, confidence, and source judgments, and from subjects' history of recall across multiple trials. The latter method requires further explanation, which is described next.

Subjects receive a minimum of three study-test cycles on a focal list. By the end of all cycles, a target's history of recall across tests can be used as an objective measure of its memory strength, by assuming that the number of correct recalls in a target's history of recall is a monotonically increasing function of its strength (or monotonically decreasing function in terms of errors). For example, a target recalled on all tests is assumed to have a stronger memory trace than a target recalled only once. In addition, targets generate sequences of correct and incorrect recall patterns across tests, which can be analyzed with a dual-recall Markov chain that separates and quantifies the retrieval processes that control recall (Brainerd, Aydin, & Reyna, 2012; Brainerd & Reyna, 2010; Gomes, Brainerd, & Stein, in press).

**Dual-recall: Theory and model**

The notion that retrieval is supported by two fundamentally distinct types of processes is not by any means new (Strong, 1913) and has been target of much investigation and theoretical developments over the last three decades (Brainerd & Reyna, 2010; Jacoby, 1991; Mandler, 1980; Yonelinas, 2002). In the fuzzy-trace theory (Reyna & Brainerd, 1995), this notion takes the form of a distinction between two types of mental representations, namely verbatim and gist. Verbatim traces are realistic and detailed representations of an item, such as its surface features (e.g., color, font, and position), whose retrieval produces recollective phenomenology—vivid mental reinstatement of an item's prior occurrence, as if flashing in the mind's eyes. Gist traces, on the other hand, are impressionistic and fuzzy representations of an item, such as the bottom-line meaning of it (e.g., apple is an edible fruit), and thus reflect individuals' *understanding* of the target event rather than the actual event.

**The dual-recall theory.** In the dual-recall model, dual-process distinctions are implemented in the form of a two-stage Markov model (Brainerd, Reyna, & Howe, 2009; Gomes, Brainerd, Nakamura, & Reyna, 2013), which posits that, over trials, studied items (targets) transition through performance states whose entries are controlled by either a recollective process, direct access of targets' verbatim traces, or a nonrecollective one, reconstruction of targets from gist traces. More specifically, individual items are assumed to transition through three discrete performance states, called states U, P, and L. State U is a transient unlearned state, in which a target can be recalled with probability 0. State P is a transient partially learned state, in which a target can be recalled with probability equal to some value $0 < p < 1$. State L is an absorbing learned state, in which a target can be recalled with probability 1. Before the first trial, targets are assumed to begin in state U, as nothing has been learned about them and, therefore,

subjects cannot yet recall any target. After the first trial, however, subjects learn something about the focal list, and then targets can transition from state U to either states P or L. After subsequent trials, targets can also transition from state P to L, but once a target enters state L, it cannot leave this state as long as the study-test trials continue—hence, state L is an absorbing state.

Retrieval of a target in state L is controlled by *direct access*. This recall operation retrieves a target's verbatim trace without comparing or searching through the traces of other items and, consequently, it is the faster type of retrieval operation. In addition, direct access to verbatim traces supports errorless recall because it allows subjects to simply read targets out of consciousness as their surface forms are mentally restored. Nonetheless, verbatim traces are more susceptible to sources of interference than are gist traces. In a free recall test, for instance, output interference makes direct access more likely to operate during the initial part of the free recall test than later on, thus constraining subjects' capacity to rely exclusively on direct access to recall list items (Barnhardt, Choi, Gerkens, & Smith, 2006; Brainerd & Reyna, 1993). As subjects undergo additional trials, however, verbatim traces should become progressively less susceptible to interference, until extremes in which an entire list can be read out of consciousness—notice that, in the dual-recall model, this property is preserved by the absorbing feature of the recollective state L.

Retrieval of a target in state P is controlled by *reconstruction* plus a slave operation, *familiarity judgment*. Reconstruction controls entry into state P and is responsible for regenerating targets from partially identifying information. The target *apple,* for instance, might be reconstructed from a gist representation of *fruit*, its bottom-line meaning. Gist traces, however, provide a basis for reconstructing candidate items rather than identifying specific ones

(e.g., *fruit* generates candidates such as *orange*, *apple*, *banana*, and *lemon*) and, therefore, it is necessary a slave operation that performs familiarity checks on reconstructed items. Similar to signal detection theory (Macmillan & Creelman, 2005), the model posits that subjects have an internal response criterion that is used to evaluate whether a reconstructed item should be output or withheld. Therefore, the nonrecollective form of recall (reconstruction + familiarity judgment) is an error-prone operation because it will at times generate and authorize output of new items. Nonetheless, because it relies on episodic traces that are less susceptible to interference than verbatim traces, reconstruction avoids the interference obstacle faced by direct access. Consequently, the two forms of recall, recollective and nonrecollective, complement each other—whereas the limitations of direct access (susceptibility to interference) are repaired by reconstruction (interference resistant), the limitations of reconstruction (error-prone recall) are repaired by direct access (errorless recall)—and thus maximize correct recall.

**The measurement model.** The model posits that the probability of recalling a target is a function of both recollective and nonrecollective recall parameters, namely direct access ($D$), reconstruction ($R$), and familiarity judgment ($J$). After an opportunity to study a focal list, and immediately prior to recall, the model posits that a target will be recalled if it occupies either the recollective state L, with probability $D$, or the nonrecollective state $P_C$, with probability $(1 - D)RJ$. Conversely, a target will not be recall if it occupies either state $P_E$, with probability $(1 - D)R(1 - J)$, or the state U, with probability $(1 - D)(1 - R)$. States U, $P_E$, $P_C$, and L are then mutually exclusive and exhaustive, as they describe all possible episodic states of a target immediately prior to recall.

After a single study-test cycle, however, there will be only one empirical degree of freedom to estimate three free parameters, which makes the model's parameters unidentifiable in single-

trial designs. One solution to this problem that has been advocated in prior studies (Brainerd et al., 2009, 2010) consists of defining the model over multiple- rather than single-trial designs. In any recall paradigm (e.g., free or associative) in which subjects receive multiple study-test trials, as in this study, correct recall of a target on each test either occurs (1) or not (0). After $k$ successive trials, targets generate a frequency distribution over $2^k$ possible error-success patterns across trials. For $k = 3$, for instance, a target will generate one out of the 8 error-success patterns, namely 111 (recalled on all tests), 110 (recalled on al but the last test), …, 000 (never recalled). Such changes in recall over trials can be conceptualized as transitions through a discrete and finite state space, in which finite Markov chains (Kemeny & Snell, 1960) provide a natural formalism by assuming the following three properties. First, changes in recall over trials consists of making transitions through a finite set of discrete episodic states $\psi_1, ..., \psi_s \in \Psi$. Second, the state a target occupies on trial $n$ (for $n = 1, ..., k$) depends only on the state it occupied immediately prior to the current state, $n - 1$. And third, at the level of individual targets, transitions through states between consecutive trials occur in an all-or-none fashion. A plethora of evidence that has accumulated since the 1960's has shown that all three assumptions hold in multi-trial recall designs, namely inter-trial transitions are all-or-none and Markovian at the item level and can be explained by models that have a small (more than two and less than five) set of exhaustive states (Bower & Theios, 1963; Estes & DaPolito, 1967; Greeno, 1968; Halff, 1977; Kintsch, 1963; Kintsch & Morris, 1965; Pagel, 1973) (for a review, see Brainerd, Howe, & Desrochers, 1982).

Markov chains can be represented in terms of a unit starting vector, whose entries give the starting unconditional probabilities of each state, and one or more transition matrices, whose entries give the conditional probability of transitioning from state $i$ on trial $n$ - 1 to state $j$ on trial

*n*. In the dual-recall model, there are four mutually exclusive episodic states, namely U, $P_E$, $P_C$, L$\in$ $\Psi$. Let $\mathbf{w}^{(1)} = [w_j^{(1)}]_{1x4}$ be a starting row vector and $\mathbf{M} = [m_{ij}]_{4x4}$ a transition matrix, then

$$\begin{aligned} \mathbf{w}^{(1)} &= \left[P(L(1)),\ P(P_E(1)),\ P(P_C(1)),\ P(U(1))\right] \\ &= \left[w_1^{(1)},\ w_2^{(1)},\ w_3^{(1)},\ w_4^{(1)}\right] \end{aligned}, \tag{1}$$

$$\begin{aligned} \mathbf{M} &= \begin{bmatrix} P(L(n)\,|\,L(n-1)) & P(P_E(n)\,|\,L(n-1)) & P(P_C(n)\,|\,L(n-1)) & P(U(n)\,|\,L(n-1)) \\ P(L(n)\,|\,P_E(n-1)) & P(P_E(n)\,|\,P_E(n-1)) & P(P_C(n)\,|\,P_E(n-1)) & P(U(n)\,|\,P_E(n-1)) \\ P(L(n)\,|\,P_C(n-1)) & P(P_E(n)\,|\,P_C(n-1)) & P(P_C(n)\,|\,P_C(n-1)) & P(U(n)\,|\,P_C(n-1)) \\ P(L(n)\,|\,U(n-1)) & P(P_E(n)\,|\,U(n-1)) & P(P_C(n)\,|\,U(n-1)) & P(U(n)\,|\,U(n-1)) \end{bmatrix} \\ &= \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{34} & m_{44} \end{bmatrix} \end{aligned}. \tag{2}$$

Remember that the model assumes that state L is an absorbing state as long as study-test trials continue. This assumption can then be formalized by changing the model's transition matrix in Equation 2 as follows:

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{34} & m_{44} \end{bmatrix} = \left[\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline \mathbf{a} & & \mathbf{T} & \end{array}\right], \tag{3}$$

in which $\mathbf{T}$ is a 3 x 3 sub-matrix of $\mathbf{M}$ whose entries give the probabilities of making transitions through transient states (U, $P_E$, and $P_C$) when a transition to the recollective and absorbing state L does not occur, and $\mathbf{a}$ is the 3 x 1 column sub-vector of $\mathbf{M}$ whose entries give the probability of transitioning from transient states on trial *n* - 1 to the state L on trial *n*. Although it is beyond the scope of this study to describe how all known properties of absorbing chains apply to the Markov chain of the dual-recall model, the canonical form of the transition matrix presented in Equation 3 is useful in deriving several statistics that are characteristic of absorbing chains via computation of the model's fundamental matrix $\mathbf{F} = (\mathbf{I} - \mathbf{T})^{-1}$, in which $\mathbf{I}$ is a 3 x 3 identity

matrix. For example, the fundamental matrix of the dual-recall model can be used to compute statistics such as the probabilities of being absorbed by the recollective state L from each transient state, $\mathbf{F} \times \mathbf{a}$, the probabilities of making a transition from each transient state to another transient state, $(\mathbf{F} - \mathbf{I}) \times \mathbf{F}_{diag}^{-1}$, and the expected number of trials before a target is absorbed by the recollective state L, $\mathbf{F} \times [1, \quad 1, \quad 1]^T$. More importantly, however, when Equations 1 and 3 are multiplied together, the entries of the resulting unit row vector $\mathbf{w}^{(n)} = [w_j^{(n)}]_{1x4}$ give the probability of a target occupying state *j* on trial *n*, as follows

$$\mathbf{w}^{(n)} = \mathbf{w}^{(1)} \times \mathbf{M}^{n-1}, \tag{4}$$

$$w_j^{(n)} = \begin{cases} w_j^{(1)}, & n = 1 \\ \sum_{i=1}^{4} w_i^{(n-1)} m_{ij}, & n > 1 \end{cases}. \tag{5}$$

Because recall of a target occurs when it occupies either states L (recollective recall) or $P_C$ (nonrecollective recall), Equations 4 and 5 provide a straightforward method for computing the probability of correct recall of a target on trial *n*, $P_n(Rc)$, namely as the inner product between $\mathbf{w}^{(n)}$ and the vector $\mathbf{c} = [1, \quad 0, \quad 1, \quad 0]$, as follows

$$\begin{aligned} P_n(Rc) &= \left\langle \mathbf{c}, \mathbf{w}^{(n)} \right\rangle = w_1^{(n)} + w_3^{(n)} \\ &= \begin{cases} w_1^{(1)} + w_3^{(1)}, & n = 1 \\ w_1^{(n-1)} + \sum_{i=2}^{4} w_i^{(n-1)} (m_{i1} + m_{i3}), & n > 1 \end{cases}. \end{aligned} \tag{6}$$

Even though many parameterizations of the entries of the transition matrix **M** are possible (e.g., Brainerd et al., 2009, 2010, 2012), Equation 3 shows that direct access ought to be entries of the $\mathbf{a}$ sub-vector, while reconstruction and familiarity judgment are the entries of the $\mathbf{T}$ sub-matrix. In this study, I used a model version with two direct access parameters ($D_1$, $D_2$), one reconstruction parameter ($R$), and three familiarity judgment parameters ($J_1$, $J_2$, $J_3$), defined over

a canonical study-test design of form $S_1T_1\ S_2T_2\ S_3T_3$, in which S is an opportunity to study a focal list and T to recall. The definition of each parameter is shown in Table 1. In this version of the dual-recall model, Equations 1 and 3 are re-written as

$$\mathbf{w}^{(1)} = \left[D_1, (1-D_1)R(1-J_1), (1-D_1)RJ_1, (1-D_1)(1-R)\right] , \tag{7}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ D_2 & (1-D_2)(1-J_n) & (1-D_2)J_n & 0 \\ 0 & (1-J_n) & J_n & 0 \\ D_2 & (1-D_2)R(1-J_n) & (1-D_2)RJ_n & (1-D_2)(1-R) \end{bmatrix} , \tag{8}$$

which produce the probabilities of each error-success pattern shown in Appendix 1.

## Experiment

Two variables that have been shown to affect retrieval processes in recognition, namely list length (Brainerd, Wright, Reyna, & Payne, 2002; Cary & Reder, 2003; Yonelinas, 1994) and word frequency (Gardiner & Java, 1990; Guttentag & Carroll, 1997), were factorially manipulated in a multi-trial free recall design, the aim being to investigate how such variables affected recollective and nonrecollective retrieval as measured by subjective methods (remember/know and confidence judgments) and objective ones (source judgments and the dual-recall model) in free recall. From previous investigations, I expected that recall accuracy would decrease as both list length increased and word frequency decreased.

At the process-level, one hypothesis is that subjective measures of dual processes tend to over-estimate recollective recall relative to objective ones, because subjective measures of recollection are boosted by factors that do not necessarily affect objective measures of it, namely noncriterial recollection. Nonetheless, one might expect that the parameters of the dual-recall model would be correlated with retrospective judgments about memory because, theoretically, their definition overlaps to some degree. More specifically, recollective recall, as measured by

retrospective judgments about memory, should correlate with direct access and the proportion of items in the recollective state L. Indeed, recall of targets whose source can be correctly identified or are judged remembered should be particularly related to direct access, as the very definition of direct access (retrieval of a target's verbatim trace) holds that subjects should be able to retrieve surface features of targets (e.g., color) in the recollective recall state L, which is critical to make accurate source judgments and to support remember judgments. Conversely, nonorecollective recall, as measured by retrospective judgments about memory, should then correlate with reconstruction, familiarity judgment, and the proportion of items in the nonrecollective state $P_C$. In both cases, however, correlations should be moderated by subjects' capacity to monitor the episodic states of recalled items.

For each list, subjects received three study-test cycles and performed retrospective judgments about the items recalled on the last test. The dual-recall model was fit to the data generated across test trials in order to estimate recollective and nonrecollective parameters. Model parameters were then used to provide process-level explanations of the effects of list length and word frequency on free recall and to separate recollective from nonrecollective retrieval of the same items that subjects made retrospective judgments about, thus allowing comparison of subjective and objective measures of dual processes. Correlational analyses between retrospective judgments and individualized model parameter estimates were also conducted. Finally, analysis of recall output position was conducted to investigate cognitive triage effects.

## Method

### Subjects

Ninety-five undergraduates (64 female), aged 20 years on average (*SD* = 2 years), participated in this experiment for extra credit. Written informed consent was obtained from all subjects prior to the beginning of the experiment.

**Experimental design**

A 3 (list length: short, medium, long) × 2 (word frequency: low, high) × 3 (type of retrospective judgment: remember/know/guess, confidence, source) full factorial mixed-design was used, in which list length and word frequency were manipulated within-subjects, and type of retrospective judgment was manipulated between-subjects.

**Materials**

Each subject studied three unrelated word lists across the experiment, a short list composed of 16 words (8 low and 8 high frequency), a medium list composed of 30 words (15 low and 15 high frequency), and a long list composed of 60 words (30 low and 30 high frequency). Low frequency (*M* = 2) and high frequency words (*M* = 197) (Kucera & Francis, 1967; Thorndike & Lorge, 1944) were randomly selected from Toglia and Battig's (1978) word norms. None of the conditions differed with respect to mean number of letters (*M* = 5.7), concreteness (*M* = 5.5), and imagery (*M* = 5.5) at the .05 significance level. Half of the words were presented in blue, while the other half in red. Each word list is presented in Appendix 2.

**Procedure**

At the beginning of the experiment, subjects were randomly assigned to one of three groups that either made remember/know/guess, confidence, or source judgments, and the order that the short, medium, and long lists were learned was also randomly assigned to each subject. For each focal list, subjects received a standard multiple-trial noncanonical study-test procedure that has been used in recent experiments to measure recollective and nonrecollective retrieval in

recall (Brainerd et al., 2012; Brainerd & Reyna, 2010; Gomes et al., in press). The overall procedure was $S_1 B_{1a} T_{1a} B_{1b} T_{1b}$ $S_2 B_2 T_2$ $S_3 B_3 T_3$ $R$, in which $S$ denotes a study phase, $B$ a buffer activity, $T$ a free recall test, and $R$ a retrospective judgment phase. During the study phases, the words were presented individually at a 2 sec rate on a computer screen, with 1 sec of inter-word interval, and subjects were told to pay close attention because their memory for the words would be tested later. The presentation order of all words was randomized for each subject and study phase. During the test phases, subjects performed a free recall test in which they were told to type as many studied words as possible and not to worry about spelling. Free recall was self-paced and terminated upon an input from the subject (the word "finish"). During the retrospective judgment phase, subjects were exposed to the words they recalled on the last test, in the same order they recalled them, and were instructed to make retrospective judgments about the recalled words according to the group to which they were assigned in the beginning of the experiment.

The instructions to make the three types of retrospective judgments were modeled on ones used in previous studies (e.g., Gardiner, 1988; Mickes et al., 2011; Rajaram, 1993; Wais et al., 2008). Subjects in the remember/know/guess group were instructed to make the following judgments regarding the words recalled on the last test: a "remember" judgment if they were able to bring back to mind a particular association, image, or something more personal from the time of study, or what the word looked like, such as its color, or sounded like when they read it to themselves during study; a "know" judgment if they had a feeling that they saw the word during study, but were unable to bring back to mind any qualitative information about it; and a "guess" judgment if they simply guessed the word during recall. Subjects in the confidence group were instructed to make the following judgments regarding the words recalled on the last test: (- - -) 1

= *Very confident* that the recalled word *was not* studied, ..., (+ + +) 6 = *Very confident* that the recalled word *was* studied. In addition, subjects were told to be cautious about using the end points of the scale, and to use them only if they could not be mistaken about their answer: "Use 1 or 6 only if you are so confident in your answer that you would be willing to testify in a court of law, even in a life-or-death situation." And finally, subjects in the source group were instructed to make judgments regarding the color (red or blue) of the words they recalled on the last test. The instructions are shown in Appendix 3.

## Results

The results are reported in five sections. In the first section, I report analyses of variance (ANOVAs) that evaluated whether there were reliable differences in recall accuracy among experimental conditions. In the second section, I present model-based analyses. Specifically, the dual-recall model was fit to the data in order to separate recollective from nonrecollective retrieval and to provide process explanations for differences in recall accuracy. In the third section, I present the results from analyses of the retrospective judgments (remember/know/guess, confidence, and source judgments) for items recalled on the last test. Because both the dual-recall model and the three retrospective judgments can be used as methods of separating memory processes in recall, in the fourth section I report comparisons among the various methods of decomposing recall performance into its recollective and nonrecollective components as well as correlational analyses between the dual-recall model and retrospective judgments. It will be seen that retrospective judgments tend to over-estimate recollective retrieval in free recall relative to the dual-recall model. Finally, in the fifth section, I report differences in performance measures and retrospective judgments as a function of the output position of items recalled on the last test. For all statistical tests, I adopted a .05 significance criterion. Because Type I error increases as a

function of the number of tests, multiple comparisons were only conducted when there was global statistical evidence of treatment effects, as indicated by omnibus tests, and multiple *t*-tests were adjusted using the Bonferroni correction.

**Recall accuracy**

Recall accuracy was evaluated with two statistics, the proportion of words correctly recalled on each trial and the mean total correct recalls per item across trials (MTR). In both cases, I computed the statistics for the $T_{1a}T_2T_3$ and $T_{1b}T_2T_3$ sequences and then averaged ($T_1T_2T_3$) because the results were identical. Summary statistics for each measure are presented in Table 2 as a function of word frequency and list length. The proportion of words correctly recalled was submitted to a 3 (test trial: $T_1$, $T_2$, $T_3$) x 3 (list length: short, medium, long) x 2 (word frequency: low, high) repeated measures ANOVA. There were main effects of test trial, $F(2,188) = 936.5$, $MSE = .02$, $\eta^2_p = .91$, and list length, $F(2,188) = 225.6$, $MSE = .05$, $\eta^2_p = .71$, and interactions between test trial and list length, $F(4,376) = 20.9$, $MSE = .01$, $\eta^2_p = .18$, between test trial and word frequency, $F(2,188) = 7.82$, $MSE = .01$, $\eta^2_p = .08$, and between word frequency and list length, $F(2,188) = 13.7$, $MSE = .03$, $\eta^2_p = .13$. Regardless of experimental condition, the mean proportion of words correctly recalled increased from trial 1 (.28) to trial 2 (.51) to trial 3 (.62). Regarding list length, the mean proportion of recalled words increased from the long list condition (.34) to the medium list (.46) to the short list (.61), and the interaction with test trial indicates that such difference was larger on later trials (trials 2 and 3) relative to the first trial. The effects of word frequency on recall, however, depended on both list length and test trial. More specifically, subjects recalled more high than low frequency words only in the long list condition, and the recall advantage of high frequency words over low frequency was only reliable on trial 3.

The results with the MTR were consistent with the analysis of the proportion of words correctly recalled. Because the MTR is a measure of overall performance, however, it cannot address trial-by-trial variability in recall. Nonetheless, in addition to list length and word frequency, it was used to investigate whether there were differences in performance among the groups of subjects assigned to each type of retrospective judgment[1]—there were not. The MTR was submitted to a 3 (list length: short, medium, long) x 2 (word frequency: low, high) x 3 (type of retrospective judgment: remember/know/guess, confidence, source) repeated measures ANOVA, which produced a main effect of list length, $F(2,184) = 217.4$, $MSE = .14$, $\eta^2_p = .70$, and an interaction between list length and word frequency, $F(2,184) = 13.1$, $MSE = .08$, $\eta^2_p = .13$. Inspection of Table 2 shows that, on average, the MTR increased from the long list condition (1.03) to medium list (1.39) to short list (1.84), regardless of word frequency. As before, however, the effects of word frequency depended on list length, namely it was reliable only in the long list condition. The difference in MTR for high and low frequency words was .22, which was more than half a *SD* unit (.58 *SD* units) and highly reliable.

**Dual-recall model analysis**

While the analysis of recall accuracy pinpointed treatment effects of list length and word frequency on recall, the purpose of this section is to pinpoint the process loci of such effects. As in the previous analysis, the dual-recall model was first fit to the data from both $T_{1a}T_2T_3$ and $T_{1b}T_2T_3$ sequences. Because the results of fit tests and the parameter estimates did not differ between the two, I report only the results with the pooled data ($T_1T_2T_3$). Equations A1-A8 were applied to the data of each experimental condition via an EM algorithm (Hu & Batchelder, 1994) that maximized Equation A12, which produced the maximum likelihood estimates of direct access, reconstruction, and familiarity judgment parameters shown in Table 3. First, however, I

address the question of whether the dual-recall model provides a close description of the data, which was investigated with goodness of fit tests. If the dual-recall model fails to fit the data by rejecting the null hypothesis of fit, then a more complex model is needed. Each test produces a $G^2$ statistic that is asymptotically distributed as $\chi^2$ (Riefer & Batchelder, 1988) with 1 degree of freedom. Because there are 3 (list length) x 2 (word frequency) = 6 conditions in the experiment, the critical value to reject the null hypothesis of fit for the experiment as a whole is 12.53. However, the observed value for the experiment as a whole was below the critical value, $G^2$ = 9.43, and thus the dual-recall model passed goodness of fit tests. In addition, simpler (one-stage) models in which items transition either through states U and L or through states P and L, did not pass goodness of fit tests. For the experiment as a whole, the one-stage model in which items transition through states U and L produces a $G^2$ statistic with 30 dfs (the critical value is 43.77), while the one-stage model in which items transition through states P and L produces a $G^2$ statistic with 12 dfs (the critical value is 21.03). The null hypothesis of fit for each model was rejected, as the observed fit statistics were well above the respective critical values (for the U and L model, $G^2$ = 47909.22, and for the P and L model, $G^2$ = 1564.32).

Turning to the parameter estimates shown in Table 3, we first conducted an experimentwise likelihood ratio test to address whether there were reliable process-level differences among all experimental conditions. This test produces a $G^2$ statistic with 30 degrees of freedom and a critical value of 43.77 to reject the null hypothesis that there are no process-level differences among experimental conditions. The observed value for this test was roughly 42 times higher than the critical value, $G^2$ = 1834.35, thus rejecting the null hypothesis. Next, I conducted a series of condition-wise tests, which are the analogue of ANOVAs' *F*-tests. This

analysis revealed reliable process-level differences for both list length, $G^2(12) = 1652.17$, and word frequency, $G^2(6) = 75.35$.

List length affected both recollective and nonrecollective processes. Regarding recollective retrieval, inspection of Table 3 suggests that mean direct access decreased from the short list condition (.21) to the medium list (.16) to the long list (.12). Likelihood ratio tests indicated that such decreases were reliable for $D_1$, $G^2(2) = 27.77$, but not for $D_2$, $G^2(2) = .70$, indicating that list length produces baseline differences in recollective retrieval rather than in the rate of transition to a recollective retrieval state after the first trial. Regarding nonrecollective retrieval, mean reconstruction also decreased from the short list condition (.44) to the medium list (.27) to the long list (.18). Such differences were supported by likelihood ratio tests, which revealed a reliable difference in $R$ among list length levels, $G^2(2) = 92.99$. Similarly, mean familiarity judgment decreased from the short list condition (.64) to the medium list (.58) to the long list (.52). However, close inspection of Table 3 suggests that such difference was primarily due to trial-by-trial invariance in the $J$ parameters in the long list condition, as opposed to the $J$ increases across trials in both the medium and short list conditions. Concerning between-condition differences, there was a significant difference in $J_2$ among list length levels, $G^2(2) = 11.93$, and in $J_3$, $G^2(2) = 10.64$, but not in $J_1$, $G^2(2) = 1.49$. Concerning within-condition differences, $J$ reliably increased across trials in both the short list, $G^2(2) = 51.41$, and medium list conditions, $G^2(2) = 12.89$, but not in the long list condition, $G^2(2) = 5.76$. Therefore, the effects of list length on learning (recall increases across trials) were due to nonrecollective retrieval.

Process-level differences in word frequency were only reliable in the long list condition, and affected primarily recollective retrieval. Visual inspection of Table 3 suggests that mean direct access decreased slightly from high to low frequency in the long list condition, and

likelihood ratio tests indicated that both differences in $D_1$, $G^2(1) = 20.59$, and in $D_2$, $G^2(1) = 16.61$, between high and low frequency words were small but highly reliable. This result indicates that word frequency not only affects baseline levels of recollective retrieval (more high than low frequency words being in that state), but it also affects the transition rate to the recollective state L across trials (more high than low frequency words transition to the recollective retrieval state L as subjects have additional opportunities to study and recall targets). Although reconstruction did not significantly differ between high and low frequency words, familiarity judgment did—specifically, $J_1$ was reliably higher for high frequency words than low frequency, $G^2(1) = 8.34$. There were no other reliable process-level differences.

**Retrospective judgments about memory**

There were three types of retrospective judgments for items recalled on the last trial (trial 3), namely remember/know/guess, confidence, and source judgments. I report the results for each of them separately.

**Remember/know judgments.** On average, subjects assigned to the remember/know/guess group made more "remember" judgments (.71) than both "know" (.27) and "guess" judgments (.02). Because "guess" judgments were not reliably different from 0, they were omitted from subsequent analyses. Summary statistics for the proportion of items recalled on the last test that received either a "remember" judgment or a "know" judgment are shown in Table 4 as a function of experimental conditions. The data in Table 4 were submitted to a 3 (list length: short, medium, long) x 2 (word frequency: low, high) x 2 (type of judgment: remember, know) repeated measures ANOVA, which revealed a main effect of type of judgment, $F(1, 26) = 54.6$, $MSE = .15$, $\eta^2_p = .68$, and list length, $F(2, 52) = 43.4$, $MSE = .01$, $\eta^2_p = .63$. The mean proportion of words recalled on the last test was higher for those that received a

"remember" judgment than for those that received a "know" judgment, and furthermore, both measures decreased significantly from the short list to the medium list to the long list conditions. There were neither reliable effects of word frequency nor reliable interactions.

**Confidence judgments.** Changes in confidence across conditions were investigated with two measures, the mean confidence rating and the proportion recalled per confidence rating. Summary statistics for both measures are shown in Table 5. Visual inspection of Table 5 suggests that mean confidence ratings were invariant across experimental conditions, which was confirmed by a 3 (list length: short, medium, long) x 2 (word frequency: low, high) repeated measures ANOVA that did not produce any reliable effects, $F$s $\leq 1.99$. However, the mean proportion of recalled words that received very high confidence ratings (+ + +, 6) seemed to decrease from the short list condition to the long. To investigate whether such declines were reliable, the mean proportion of recalled words that either received a very high confidence rating or lower ratings was submitted to a 3 (list length: short, medium, long) x 2 (word frequency: low, high) x 2 (confidence level: + + +, lower) repeated measures ANOVA. The analysis produced a main effect of confidence level, $F(1, 27) = 11.3$, $MSE = .57$, $\eta^2_p = .30$, list length, $F(2, 54) = 62.5$, $MSE = .01$, $\eta^2_p = .70$, and word frequency, $F(1, 27) = 11.2$, $MSE = .01$, $\eta^2_p = .29$. The mean proportion of recalled words that received a very high confidence rating was reliably higher (.46) than those that received lower confidence ratings (.18). In addition, regardless of confidence level, the mean proportion of recalled words decreased both as list length increased and as word frequency decreased. There were no reliable interactions.

**Source judgments.** Source judgments (color) were used to compute two statistics, source accuracy and the proportion of words recalled whose source was either correctly identified or not. The means and standard deviations of both types of measures are shown in

Table 6 as a function of experimental conditions. Across all conditions, mean source accuracy (.63) was reliably higher than chance (.50), $t(39) = 6.89$. Visual inspection of Table 6 suggests that mean source accuracy tends to decrease as list length increases. However, when the mean source accuracy data were submitted to a 3 (list length: short, medium, long) x 2 (word frequency: low, high) repeated measures ANOVA, there were neither reliable effects nor interactions, $F$s $\leq 2.04$. However, the proportions of recalled words as a function of source accuracy produced reliable effects. The data were submitted to a 3 (list length: short, medium, long) x 2 (word frequency: low, high) x 2 (source accuracy: correct, incorrect) repeated measures ANOVA, which produced main effects of list length, $F(2, 76) = 65.6$, $MSE = .01$, $\eta^2_p = .63$, and source accuracy, $F(1, 38) = 37.5$, $MSE = .08$, $\eta^2_p = .50$, and an interaction between list length and source accuracy, $F(2, 76) = 6.95$, $MSE = .31$, $\eta^2_p = .16$. Subjects recalled more words whose source was correctly identified (.38) than words whose source was not correctly identified (.22), and such difference increased as list length decreased. Even though the proportion of recalled words decreased as list length increased, regardless of source accuracy, such decreases were larger for recalled words whose source was correctly identified than for words whose source was incorrectly identified. There were no other reliable effects.

**Relationship between the dual-recall model and retrospective judgments**

In this section, I addressed questions regarding similarities and differences among the various methods of decomposing recall into its recollective and nonrecollective components, namely via the dual-recall model or the three retrospective judgments about memory (remember/know, confidence, and source judgments), and whether components of such separation methods correlate at the level of individuals.

**Recall decomposition.** The proportion of words recalled on the last test (see Table 1) was decomposed into its recollective and nonrecollective components via four separation methods, namely the dual-recall model, source accuracy, remember/know judgments, and confidence ratings. Because retrospective judgments were made only for items output on the last trial, for the dual-recall model, the parameter estimates in Table 3 were plugged into the model's starting vector and transition matrices in order to estimate the proportion of recalled items in state L (recollective recall state) and state $P_C$ (nonrecollective recall state) on trial 3 via Equation 5. For source accuracy, the proportion of recalled words whose source was correctly identified was used as a measure of recollective retrieval, while nonrecollective retrieval was the proportion of recalled words whose source was incorrectly identified.[2] For remember/know judgments, the proportion of recalled words that received a "remember" judgment was used as a measure of recollective retrieval, while nonrecollective retrieval was the proportion of recalled words that received either a "know" or "guess" judgment. For confidence ratings, the proportion of recalled words that received the maximum confidence rating (+ + +, 6) was used as a measure of recollective retrieval, while nonrecollective retrieval was the proportion of recalled words that received confidence ratings lower than the maximum (5, 4, 3, 2, or 1).[3] In all cases, the decomposition was performed on recall performance of the whole sample (see Table 1), rather than individual groups, and separated only by list length, as previous analyses have not shown reliable differences in recall accuracy among retrospective judgment groups and none of the retrospective judgments interacted with word frequency.

Three main findings emerged when recall was decomposed into recollective and nonrecollective recall (see Figure 5), two concerning qualitative similarities among the separation methods and one concerning quantitative differences among them. Regarding their

similarities, all four separation methods showed that recollective recall prevailed over nonrecollective recall on the last trial. Across all conditions and separation methods, recall on the last test was roughly 2 times more likely to be supported by recollective retrieval than nonrecollective retrieval—it accounted for 2/3 of the recalled words. In addition, regardless of the separation method, recollective recall decreased as list length increased. However, close inspection of Figure 5 indicates that all three retrospective judgments, particularly remember/know and confidence judgments, tended to over-estimate recollective retrieval relative to the dual-recall model.[4] Nonetheless, such a tendency depended on list length—specifically, overestimation was higher in the short list condition than in the long list condition. Relative to the dual-recall model, recollective retrieval estimated from retrospective judgments was 1.28, 1.23, and 1.00 times higher in the short, medium, and long list conditions, respectively.

**Correlational analysis.** Next, I investigated the *individual*-level relationship between parameters of the dual-recall model and the retrospective measures of dual processes. This analysis consisted of, first, estimating parameters of the dual-recall model for each subject across all conditions (i.e., computing experimentwise parameter estimates to maximize parameter reliability) and, second, running correlations between parameters and experimentwise statistics of the retrospective measures of dual processes. Regarding model fit, the critical value to reject the null hypothesis of fit of the dual-recall model is 3.84 for each subject. The observed $G^2(1)$ value for one subject was roughly 6 times higher than the critical value and, therefore, the subject was removed from subsequent analyses. For the remaining subjects, the mean $G^2(1)$ value was 2.12 and the null hypothesis of fit could not be rejected for 82% of the subjects.

The correlations between parameters of the dual-recall model and retrospective measures of dual processes are shown in Table 7. The definition of recollective and nonrecollective

retrieval for each separation method was the same as used in the recall decomposition analysis. Inspection of Table 7 indicates that, overall, recall followed by either correct source identification or remember judgment correlate with direct access of targets' verbatim traces in free recall. Recall followed by correct source identification showed reliable positive correlations with the dual-recall model's recollective retrieval statistics, namely $D_1$, $D_2$, mean $D$, and the proportion of items in the recollective state L on the last trial. Similarly, recall followed by remember judgment also showed reliable correlations with the dual-recall model's recollective retrieval statistics. Unlike correct source identification, however, recall followed by remember judgment correlated positively with reconstruction and negatively with familiarity judgment, indicating that remember judgments are influenced by recollective retrieval processes and, to a lesser degree, by nonrecollective retrieval processes. In contrast, recall followed by a maximum confidence rating did not show reliable correlations with any statistic from the dual-recall model, and thus confidence was completely independent of recall accuracy.

Whereas recall followed by correct source identification or remember judgment predicted all recollective retrieval statistics of the dual-recall model, recall followed by maximum confidence did not. This result is illustrated in Figure 6, in which the proportion of items recalled in the recollective state L on trial 3, measured by the dual-recall model, is plotted against the three retrospective measures of recollective retrieval, namely recall followed by remember judgment (Figure 6A), or followed by correct source identification (Figure 6B), or followed by maximum confidence (Figure 6C). As in the recall decomposition analysis, Figure 6 shows that retrospective measures of dual processes, particularly remember/know and confidence judgments, tend to over-estimate recollective retrieval in free recall, as most observations fell under the identity line (perfect calibration) in panels A, B, and C. For correct source judgments, the best

fitting linear function (solid line), y = .71x, $F(1, 38) = 7.10$, indicates that calibration is best closer to the origin rather than at higher values. Similarly, for remember judgments, the best fitting linear function, y = .85x, $F(1, 25) = 12.30$, indicates that calibration is also best closer to the origin and it decreases thereafter, although such decreases are lower than for source judgments. Regarding nonrecollective retrieval, none of the retrospective judgments showed reliable correlations with the dual-recall model's nonrecollective retrieval statistics, suggesting that subjects were better able to monitor items occupying a recollective state on the last test than items occupying a nonrecollective state.

**Output position**

The analysis of output position focused on items recalled on the last trial, as retrospective judgments were only made to such items. The results are presented in two parts. In the first, the results of the analysis of output variability in four measures of items recalled on trial 3 are presented—specifically, the mean total number of errors per item (MTE), the mean confidence rating, the proportion of "remember" responses, and source accuracy. In the second part, the results of the analysis of output dependencies for each of the four measures of items recalled on trial 3 are presented. Because the latter analysis was not central to the present study, it was presented in Appendix 5. For all such analyses, output position was first partitioned into vincentised quartiles (VO1, VO2, VO3, and VO4) (Levine & Burke, 1972). For example, if a subject recalled 12 words on the last test, then the first 3 words in the subject's free recall protocol are part of the 1$^{st}$ vincentised output position (VO1), the next 3 are part of the 2$^{nd}$ vincentised output position (VO2), and so on. However, if a subject recalled 16 words on the last test, then the first 4 words in the subject's free recall protocol are part of VO1, the next 4 are part of VO2, and so on.

**Output variability.** The MTE, the mean confidence rating, the mean proportion of remember responses, and the mean source accuracy measures are shown in Figures 7, 8, 9, and 10, respectively, as a function of both vincentised output position in the last trial and list length. Regarding MTE, visual inspection of Figure 7 suggests that cognitive triage (i.e., a U-shaped pattern in MTE as a function of output position) interacted with list length, namely it increased as list length decreased. To investigate whether this interaction was reliable, we submitted the MTE to a 3 (list length: short, medium, long) x 4 (output position: VO1, VO2, VO3, VO4) x 3 (retrospective judgment group: remember/know/guess, confidence, source) repeated measures ANOVA, which showed a small but reliable interaction between list length and output position, $F(6, 552) = 2.02$, $MSE = .54$, $\eta^2_p = .02$. In the short list condition, MTE was reliably higher in VO2 (1.56) than in VO3 (1.20). Notice, however, that this pattern was not observed with mean confidence ratings (Figure 8), which only decreased as a function of output position (according to the MTE analysis, it should increase) and it did not interact with list length. As before, we submitted the mean confidence ratings to a 3 (list length: short, medium, long) x 4 (output position: VO1, VO2, VO3, VO4) repeated measures ANOVA, which only revealed a main effect of output position, $F(3, 81) = 3.88$, $MSE = .15$, $\eta^2_p = .13$. Regardless of list length, mean confidence ratings reliably decreased from VO1 (5.68) to VO4 (5.49). Although the mean proportion of remember responses (Figure 9) seemed to behave more like confidence ratings than MTE across output positions, and source accuracy (Figure 10) seemed to behave more like MTE than confidence ratings across output positions, neither source accuracy nor remember responses produced statistically reliable effects.

## Discussion

This study had three main objectives. The first objective was to investigate the effects of list length and word frequency on both subjective and objective measures of recollective and nonrecollective processes in free recall. The second objective was to investigate the relationship between the dual-recall model and retrospective judgments about memory that are used as tools for measuring dual processes in episodic memory tasks. The third objective was to test the hypothesis that subjects can assign higher confidence to items associated with weaker memory traces relative to items associated with stronger memory traces. Subjects received multiple study-test trials and then performed one out of three types of retrospective judgments about targets recalled on the last test, namely remember/know, confidence, and source judgments. This procedure allowed me to compute subjective and objective measures of memory processes and strength.

As expected from prior investigations (Brainerd et al., 2002; Cary & Reder, 2003; Deese, 1960; Gregg, Montgomery, & Castaño, 1980; Tulving & Patkau, 1962; Yonelinas, 1994), the proportion of correctly recalled words decreased as list length increased, and high frequency words were better recalled than low frequency words in the long list condition. The four methods of separating recollective from nonrecollective recall produced process explanations for the effects of list length and word frequency in recall. In addition, the separation methods revealed qualitative similarities and quantitative differences among them. I discuss such findings in the next sections.

**List length in free recall**

List length produces similar effects in recognition and recall (Brainerd et al., 2002; Cary & Reder, 2003; Ward, 2002)—specifically, recognition and recall accuracy increase as list length decreases. Although dual-process research has indicated that list length affects recollective

retrieval and spares nonrecollective retrieval in recognition (Yonelinas, 1994; Yonelinas & Jacoby, 1994), whether the same explanation applies to free recall is unknown. Indeed, the findings reported in this study suggest that current dual-process explanations of such effect tell an incomplete story in free recall. As in recognition, list length affected recollective retrieval in free recall. However, the current dual-process explanation is incomplete because list length affected a nonrecollective process in free recall that is not required in recognition, namely reconstruction. On average, recollective recall decreased from .22 to .15 to .10 on the first trial in the short, medium, and long list conditions, respectively, and similarly, nonrecollective recall (reconstruction + familiarity judgment) decreased from .16 to .11 to .08 in the same conditions. On the first trial, decreases in nonrecollective retrieval across list lengths were due to reconstruction rather than familiarity judgment, suggesting that list length affects target reconstruction from partial information rather than subjects' inclination to output reconstructed items.

Because a multi- rather than a single-trial design was used in this study, it was also possible to address a question that has not figured in dual-process research before, namely are the process level effects of list length across trials the same as on the first trial? The answer is no. As subjects received new opportunities to study and recall the focal lists, two patterns emerged. First, list length affected the rate of learning. While the difference in recall between the short and long list conditions was .14 on the first trial, the same difference increased roughly twofold on the third trial, to .27. Second, at the process level, list length affected the rate of learning via nonrecollective processes. Targets' transition rate from the no-recall states U and $P_E$ to the nonrecollective recall state $P_C$ (measured with parameters $R$, $J_2$, and $J_3$) increased as the list length decreased, but the transition rate from the same no-recall states to the recollective recall

state L (measured with parameter $D_2$) did not change reliably across list lengths. In particular, subjects inclination to output reconstructed targets increased from the first to the third trials in the short and medium list length conditions, but not in the long list length condition. Consequently, differences in nonrecollective recall between the short and long list length conditions increased from the first to the last trials. On the first trial, nonrecollective recall decreased from .16 in short list condition to .08 in the long list condition, whereas on the last trial, nonrecollective recall decreased from .35 to .16 in the same conditions.

After the last test trial, subjects were asked to make retrospective judgments about items recalled on the last test, which were then used to decompose recall in terms of its recollective and nonrecollective components (see Figure 5). This analysis indicated that, regardless of measurement method, list length affected both recollective and nonrecollective retrieval in free recall. In other words, the process-level effects of list length indicated via model-based analysis were also supported by retrospective judgments about memory. Nonetheless, retrospective judgments about memory tended to over-estimate recollective recall relative to the dual-recall model in the short and medium list length conditions.

In sum, list length affected free recall in two ways. The first is an extension to free recall of the process explanation of the effects of list length on recognition. Specifically, as list length decreases, recollective retrieval (direct access of targets' verbatim traces) increases. The second way that list length affects free recall is different from recognition, because it involves a nonrecollective process that is not necessary in recognition, namely targets' reconstruction from partial information. In addition, changes in nonrecollective processes across trials accounted for the effects of list length on the rate of learning. More specifically, from the first to the last trials,

subjects became ever more prone to output reconstructed targets from shorter than from longer lists.

**Word frequency in free recall**

Contrary to list length, word frequency has opposite effects in recognition and recall (Gregg, 1976; Guttentag & Carroll, 1997; MacLeod & Kampe, 1996). Recognition accuracy is higher for low rather than high frequency words, whereas recall accuracy is higher for high rather than low frequency words. In typical old/new recognition experiments in which subjects study lists of low and high frequency words, word frequency produces a mirror effect (Glanzer & Adams, 1985). Although subjects are slightly more prone to accept high frequency targets during test than low frequency targets (i.e., the hit rate is slightly higher for high frequency words relative to low frequency words), they are also much less prone to accept low frequency distractors than high frequency distractors (i.e., the false alarm rate is much lower for low frequency words relative to high frequency words), and thus, recognition accuracy is usually better for low rather than high frequency words.

In a free recall experiment, on the other hand, the test is self-cued, which means that subjects need to retrieve targets by themselves rather than to compare test probes against stored information about studied items. The present study used the latter type of test and revealed that high frequency words were better recalled than low frequency words in the long list length condition. At the process-level, the recall advantage of high frequency words relative to low frequency ones was mainly recollective. In the long list length condition, 12% of the high frequency targets were recalled recollectively, whereas 8% of the low frequency targets were recalled recollectively. Such difference, although small, was highly reliable. In addition, list length also affected familiarity judgment on the first trial, a component of nonrecollective recall.

However, because nonrecollective recall was particularly small in the long list condition during the first trial (.08), familiarity judgment accounted for a difference of roughly 1% in nonrecollective recall between high frequency words (.09) and low frequency words (.08), and thus it should be interpreted with caution.

As in the case of list length, word frequency also affected the rate of learning, even though the interaction between word frequency and test trial was smaller ($\eta^2_p$ = .08) than the interaction between list length and test trial ($\eta^2_p$ = .18). However, contrary to list length, process-level analysis indicated that the effect of word frequency on learning was purely recollective. Specifically, $D_2$ was reliably higher for high frequency words relative to low frequency words in the long list condition, while none of the nonrecollective parameters showed significant differences (see Table 3). Despite that, none of the retrospective judgments about memory showed reliable effects of word frequency in free recall and, therefore, they do not provide a process explanation of differences in recall accuracy. This could, of course, be due to differences in statistical power among measurement methods (model-based analyses used the data from the entire sample, whereas retrospective judgments used the data from sub-samples) and do not, by any means, indicate that retrospective judgments might provide a different explanation in comparison to the dual-recall model. In fact, for recalled targets from the long list condition that either received remember judgments or whose source were correctly identified, the results go in the same direction as the findings from model-based analysis. In the case of remember/know judgments (see Table 4), recall followed by remember judgments was numerically higher for high frequency words than low frequency words (the difference was .07), whereas the difference in recall followed by know judgments between high and low frequency words was numerically lower (the difference was .02). Similarly, in the case of source

judgments (see Table 6), recall followed by correct source identification was numerically higher for high frequency words than low frequency words (the difference was also .07), whereas the difference in recall followed by incorrect source identification between high and low frequency words was numerically lower (the difference was also .02). Notice that this trend was not observed with confidence ratings and, as before, none of them were statistically reliable.

In free recall, Gregg, Montgomery, and Castaño (1980) have shown that recall accuracy is different between high and low frequency words in pure lists designs (i.e., when subjects study either a list composed of only low frequency words or a list composed of only high frequency words), but not in mixed lists designs (see also May &Tryk, 1970). This result is clearly at odds with the findings from the current study, as subjects studied mixed rather than pure lists and produced reliable word frequency effects. However, there are several important methodological differences between the two experiments that might elucidate why they produced different results. First, list length differed between the two studies. In Gregg et al.'s experiments, subjects studied lists composed of 12 words, and in mixed lists, half were high frequency words and the other half, low frequency. In comparison, I found reliable word frequency effects only with a list composed of 60 words (48 additional data points per subject and test trial), in which half of them were of high frequency and the other half of low frequency. Low statistical power could then explain the null result in Gregg et al.'s study, but this is a weak explanation because in their experiment, absolute values go in the opposite direction and recall accuracy was reliably better for low rather than high frequency words when subjects' attention was divided during study. Interestingly, in the present study, recall of low frequency targets from the short list was numerically higher than recall of high frequency targets, a pattern that reversed as (a) list length increased and (b) subjects had additional opportunities to study and recall targets. In mixed list

designs, Watkins, LeCompte, and Kim (2000) found that subjects do not show differences in recall between high and low frequency words when they are aware that they will perform a recall test after study, but recall accuracy is better for high than low frequency words when they are not aware that they will perform a recall test after study. In light of this finding, Watkins et al. argued that, in mixed list designs, null and reversed word frequency effects in free recall are due to strategic study of the focal list that favors low over high frequency words in order to optimize recall. In this same vein, one possible explanation for the findings reported in this study is that subjects rely ever less on such strategy as list length increases and as they have increasing numbers of opportunities to study and recall targets.

**Subjective and objective measures of dual processes in free recall**

Subjective reports of retrieval phenomenologies, such as remember/know judgments, and memory strength, such as confidence judgments, as well as objective measures of retrieval of contextual information, such as source judgments, have been widely used, at times interchangeably, to separate recollective from nonrecollective retrieval in memory experiments (Wais et al., 2008, 2010; Wilding & Rugg, 1996; Yonelinas, 1999; Yonelinas & Jacoby, 1995). In recognition, Yonelinas (2001) has shown that dual-process parameters estimated from remember/know judgments, the process dissociation procedure (Jacoby, 1991), and ROC curves are positively and highly correlated with each other *across experimental conditions*. In free recall, we compared remember/know, confidence, and source judgments against parameter estimates from the dual-recall model, which estimates recollective and nonrecollective recall from subjects' performance rather than introspective judgments. This is the first study to make such a comparison in free recall and, additionally, subjective and objective measures of dual

processes were compared at the individual level, which allowed for inter-subject variability in the efficiency of each retrieval process.

Both subjective and objective measures of dual processes showed that free recall of items on the last trial was mainly supported by recollective retrieval. Across measures and experimental conditions, recollective retrieval accounted for roughly 2/3 of the words recalled on the last trial. This finding is consistent with Tulving's (1985) seminal study, for instance, in which remember judgments accounted for the majority of the words freely recalled (see also Hamilton & Rajaram, 2003). Nonetheless, subjective measures of dual processes (remember/know and confidence judgments) tended to over-estimate recollective recall relative to objective measures (source judgments and the dual-recall model). On average, recollective retrieval measured with subjective methods accounted for 71% of the words recalled on the last test, while recollective retrieval measured with objective methods accounted for 61%. Such differences increased as list length decreased and were particularly higher between the dual-recall model and subjective measures of recollective retrieval.

The idea that subjective measures of dual processes over-estimate recollective retrieval in comparison to objective ones is predicted by the concept of noncriterial recollection (Yonelinas & Jacoby, 1996). During test, subjects at times recollect information associated with a target (e.g., "I remember thinking of Isaac Newton when I saw the words *apple* and *tree*"; "I felt disgusted after seeing the word *rotten*") that provide a basis for making remember judgments or changing one's confidence about the recalled item. Nevertheless, the same information might be neither diagnostic of the target's occupancy in the recollective state L nor relevant to making correct identification of the target's color. Conversely, all information that provides a basis for objective measures of recollective retrieval also support subjective measures of it and, therefore,

subjective measures of recollective retrieval will tend to over-estimate recollective retrieval, relative to objective ones. This pattern was indeed observed both at the group and at the individual levels when comparing the dual-recall model against retrospective judgments about memory (see Figures 5 and 6).

In Figure 6, the dashed line represents the identity line, or perfect calibration, whereas the solid line represents the linear function that best fits the data. Only recall followed by correct source identification and recall followed by remember judgment were reliable predictors of the proportion of targets in the recollective state L. In particular, for remember judgments, calibration was best in subjects whose proportion of items in the recollective state L was higher than .20 (all subjects whose proportion of items in the recollective state L was below .20 over-estimated recollective recall). Indeed, recall followed by remember judgments was a better predictor of recollective retrieval parameters than recall followed by correct source identification. Source identification, however, does not always require item memory (Brainerd et al., 2012; Starns et al., 2008), and although source accuracy was above chance, subjects could have made correct source identifications based on guessing, as they were required to choose one of the two possible sources even in the absence of a vivid recollection of the target's color. In fact, when the source data was corrected for random guessing (see Appendix 4), none of the source measures predicted the parameters of the dual-recall model. Nonetheless, contrary to correct source identification, the recollective phenomenology that characterizes remember judgments was not a distinctive feature of direct access of targets' verbatim traces, as subjects also attributed it to targets reconstructed from partial information (see Table 7). However, subjects' ability to identify nonrecollective retrieval when it occurs in free recall was noticeably weaker than their ability to identify recollective retrieval. Specifically, recall followed by either

incorrect source identification, or know/guess responses, or low confidence was not a reliable predictor of nonrecollective recall.

**Is confidence a proxy of memory strength in free recall?**

The idea that subjects can consciously monitor the memory strength of recalled items was not supported by the findings reported in this study. I compared the effects of output position on two proxies of memory strength, one objective (MTE) and another subjective (retrospective confidence). Prior studies have shown that MTE shows a non-monotonic relationship with output position in free recall tasks (a U-shaped curve across output positions). This relationship, called the cognitive triage effect (Brainerd et al., 1991), contradicts our intuition that items are output in a strong → weak fashion. One hypothesis is that subjects are aware of such triage process and, therefore, are able to assign confidence judgments accordingly whenever cognitive triage effects are observed. This hypothesis, however, was not supported by the results of this study. In fact, whereas the objective measure indicated that memory strength *increased* from earlier to later output positions in the short list condition (MTE decreased), the subjective measure indicated that memory strength *decreased* from earlier to later output positions in the same condition (confidence decreased). In other words, confidence judgments were consistent with our mistaken intuition about how memory strength is distributed across output positions in free recall. Indeed, the correlation between mean confidence rating and the MTE was low and unreliable. In the short, medium, and long list length conditions, the correlation between confidence rating and MTE was .11, .13, and .18, respectively. In addition, whereas subjective memory strength showed strong output dependencies (see Appendix 5 for more details), objective memory strength did not, which further suggests that confidence is not sampled

directly from memory strength in free recall but it is also influenced by prior responses and our intuitions about how targets are recalled in a free recall task.

## Conclusion

List length and word frequency effects have long been target of investigation (e.g., Strong, 1912) and have traditionally played a critical role in the development of memory theories and models. In free recall, the findings reported in this study indicate that both recollective and nonrecollective processes are influenced by list length. Specifically, as list length decreases, direct access of targets' verbatim traces increases as well as reconstruction from partial information. In addition, list length also affects learning via nonrecollective processes—as subjects received additional opportunities to study and recall the targets from a focal list, reconstructed targets from short and medium lists become ever more prone to be output, but not targets from long lists. When word frequency was manipulated within-subjects, it produced small but reliable effects in recollective recall that increased as a function of test trial. However, such effects were restricted to long lists, in which high frequency words facilitated direct access in comparison to low frequency words.

The comparison between the dual-recall model and retrospective judgments about memory revealed four main results. First, direct access of targets' verbatim traces correlated positively with the proportion of recalled targets that received remember rather than know/guess judgments, and additionally, it correlated with the proportion of recalled targets whose source (color) were correctly identified rather than incorrectly identified. Remember judgments also correlated with reconstruction, indicating that items reconstructed from partial information can at times produce recollective phenomenology, although to a lesser degree than items directly accessed. Second, as a method of separating recollective from nonrecollective retrieval,

retrospective confidence did not show any reliable correlations with statistics from the dual-recall model. Third, as a proxy for memory strength, retrospective confidence was unrelated to an objective measure of memory strength (items' history of recall across trials) and conformed to people's intuition about how memory strength is distributed across output position in free recall protocols (strong → weak), even though analysis of objective strength indicated that such intuition was mistaken (targets were output in a weak → strong fashion). Lastly, across output position, confidence ratings made to items output later in the free recall protocol were largely influenced by ratings made to items output earlier (output dependencies), and a similar pattern, although less prominent, was observed with the proportion of remember judgments made to recalled items (see Appendix 5 for details). The findings indicate that confidence ratings should not be regarded as direct measures of either memory strength or dual processes in free recall, and importantly, at least in free recall, subjects are better able to monitor the episodic state of targets in a recollective state rather than targets in a nonrecollective state—in fact, there was no evidence that subjects can do the latter at all.

## References

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*, 81-99.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 929-945.

Barnhardt, T. M., Choi, H., Gerkens, D. R., & Smith, S. M. (2006). Output position and word relatedness effects in a DRM paradigm: Support for dual-retrieval process theory of free recall and false memories. *Journal of Memory and Language*, *55*, 213-231.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548-564.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metacognitive index. *Journal of Experimental Psychology: General*, *127*, 55-68.

Bernbach, H. A. (1967). Decision processes in memory. *Psychological Review*, *74*, 462-480.

Bower, G. H., & Theios, J. (1963). A learning model for discrete performance levels. In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 1–31). Stanford, CA: Stanford University Press.

Brainerd, C. J., Aydin, C., & Reyna, V. F. (2012). Development of dual-retrieval processes in recall: Learning, forgetting, and reminiscence. *Journal of Memory and Language*, *66*, 763-788.

Brainerd, C. J., Howe, M. L., & Desrochers, A. (1982). The general theory of two-stage learning: A mathematical review with illustrations from memory development. *Psychological Bulletin*, *91*, 634–665.

Brainerd, C. J., Payne, D. G., Wright, R., & Reyna, V. F. (2003). Phantom recall. *Journal of Memory and Language*, *48*, 445-467.

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review*, *10*, 3–47.

Brainerd, C. J., & Reyna, V. F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review*, *100*, 42-67.

Brainerd, C. J., & Reyna, V. F. (2010). Recollective and nonrecollective recall. *Journal of Memory and Language*, *63*, 425-445.

Brainerd, C. J., Reyna, V. F., Holliday, R. E., & Nakamura, K. (2012). Overdistribution in source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 413-439.

Brainerd, C. J., Reyna, V. F., & Howe, M. L. (2009). Trichotomous processes in early memory development, aging, and neurocognitive impairment: A unified theory. *Psychological Review*, *116*, 783-832.

Brainerd, C. J., Reyna, V. F., Howe, M. L., & Kevershan, J. (1991). Fuzzy-trace theory and cognitive triage in memory development. *Developmental Psychology*, *27*, 351-369.

Brainerd, C. J., Wright, R., Reyna, V. F., & Payne, D. G. (2002). Dual-retrieval processes in free and associative recall. *Journal of Memory and Language*, *46*, 120-152.

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior*, *26*, 353–364.

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.

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, *22*, 273–280.

Crowder, R. G. (1976). *Principles of learning and memory*. Oxford, England: Lawrence Erlbaum.

DeCarlo, L. T. (2003). Source monitoring and multivariate signal detection theory, with a model for selection. *Journal of Mathematical Psychology*, *47*, 292-303.

Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports*, *7*, 337-344.

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

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

Dunn, J. C. (2008). The dimensionality of the remember–know task: A state-trace analysis. *Psychological Review, 115*, 426–446.

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America*, *31*, 768-773.

Estes, W. K., & DaPolito, F. (1967). Independent variation of information storage and retrieval processes in paired-associate learning. *Journal of Experimental Psychology*, *75*, 18–26.

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

Fullerton, G. S., & Cattell, J. (1982). On the perception of small differences. *University of Pensylvania Philosophical Series*, *2*.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*. *16*, 309-313.

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, *18*, 23-30.

Geraci, L., & McCabe, D. P. (2006). Examining the basis for illusory recollection: The role of remember/know instructions. *Psychonomic Bulletin & Review*, *13*, 466-473.

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

Gomes, C. F. A., & Brainerd, C. J., Nakamura, K., & Reyna, V. F. (2013). Dual memory processes under Markovian interpretations. Manuscript in preparation.

Gomes, C. F. A., & Brainerd, C. J., Stein, L. M. (in press). Effects of emotional valence and arousal on recollection and nonrecollective recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

Greeno, J. G. (1968). Identifiability and statistical properties of two-stage learning with no successes in the initial stage. *Psychometrika*, *33*, 173–215.

Gregg, V. H., Montgomery, D. C., & Castaño, D. (1980). Recall of common and uncommon words from pure and mixed lists. *Journal of Verbal Learning and Verbal Behavior*, *19*, 240-245.

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

Halff, H. M. (1977). The role of opportunities for recall in learning to retrieve. *American Journal of Psychology*, *90*, 383–406.

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

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning and Verbal Behavior*, *6*, 685-691.

Healy, A. F., & Jones, C. (1973). Criterion shifts in recall. *Psychological Bulletin*, *79*, 335-340.

Henmon, V.A.C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*, 186-201.

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 57-80.

Hollingworth, R.L. (1913). Experimental studies in judgment. *Archives of Psychology*, 29, 1-119.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21-47.

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

Jou, J. (2008). Recall latencies, confidence, and output positions of true and false memories: Implications for recall and metamemory theories. *Journal of Memory and Language*, *58*, 1049-1064.

Keast, A., Brewer, N., & Wells, G. L. (2007). Children's metacognitive judgments in an eyewitness identification task. *Journal of Experimental Child Psychology*, *97*, 286-314.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*, 1–24.

Kemeny, J. G., & Snell, J. L. (1960). *Finite Markov chains.* New Jersey: Van Nostrand.

Kintsch, W. (1963). All-or-none learning and the role of repetition in paired-associate learning. *Science*, *140*, 310–312.

Kintsch, W., & Morris, C. J. (1965). Application of a Markov model to free recall and recognition. *Journal of Experimental Psychology*, *69*, 200–206.

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review*, *17*, 465-478.

Koriat, A. (2002). Metacognition research: An interim report. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 261-286). Cambridge University Press.

Koriat, A., & Ackerman, R. (2011). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science*, *13*, 441-453.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.

Levine, G., & Burke, C. J. (1972). *Mathematical model techniques for learning theories*. New York: Academic Press.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*, 100-109.

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

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*, 132-142.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum.

Marche, T. A., Howe, M. L., Lane, D. G., Owre, K. P., & Briere, J. L. (2009). Invariance of cognitive triage in the development of recall in adulthood. *Memory*, *17*, 518-527.

May, R. B., & Tryk, H. E. (1970). Word sequence, word frequency, and free recall. *Canadian Journal of Psychology*, *24*, 299-304

McCabe, D. P., Roediger, H. L., III, McDaniel, M. A., & Balota, D. A. (2009). Aging reduces veridical remembering but increases false remembering: Neuropsychological test correlates of remember-know judgments. *Neuropsychologia*, *47*, 2164-2173.

McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, *34*, 261-267.

Metcalf, J. T. (1917). An experimental study of the conscious attitudes of certainty and uncertainty. *Psychological Monographs*, *23*, 181-240.

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140*, 239-257.

Migo, E. M., Mayes, A. R., & Montaldi, D. (2012). Measuring recollection and familiarity: Improving the remember/know procedure. *Consciousness and Cognition, 21*, 1435-1455.

Naveh-Benjamin, M., & Kilb, A. (2012). How the measurement of memory processes can affect memory performance: The case of remember/know judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 194–203.

Nelson, T. O., Dunlosky, J., White, D. M., Steinberg, J., Townes, B. D., & Anderson, D. (1990). Cognition and metacognition at extreme altitudes on Mount Everest. *Journal of Experimental Psychology: General*, *119*, 367-374.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The psychology of learning and motivation: Advances in research and theory, 26*, 125–173.

Pagel, J. C. (1973). Markov analysis of transfer in paired-associate learning with high intralist similarity. *Journal of Verbal Learning and Verbal Behavior*, *12*, 456–470.

Rabin, M. D., & Cain, W. S. (1984). Odor recognition: Familiarity, identifiability, and encoding consistency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 316-325.

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

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*, 1-75.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318-339.

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence–accuracy correlation. *Journal of Applied Psychology*, *81*, 587–594.

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

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, *22*, 511-524.

Schwartz, B. L., Benjamin, A. S., Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, *6*, 132-137.

Starns, J. J., Hicks, J. L., Brown, N. L., & Martin, B. A. (2008). Source memory for unrecognized items: Predictions from multivariate signal detection theory. *Memory & Cognition*, *36*, 1-8.

Strack, F., & Förster, J. (1995). Reporting recollective experiences: Direct access to memory systems? *Psychological Science*, *6*, 352-358.

Strong, E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review*, *19*, 447-462.

Strong, E. K. (1913). The effect of time-interval upon recognition memory. *Psychological Review*, *10*, 339-372.

Thorndike, E.L., & Lorge, I. (l944). *The Teacher's Word Book of 30,000 Words*. Teachers College, Columbia University.

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Lawrence Erlbaum.

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

Tulving, E., & Patkau, J. E. (1962). Concurrent effects of contextual constraint and word frequency on immediate recall and learning of verbal material. *Canadian Journal of Psychology*, *16*, 83-95.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582-600.

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

Wais, P. E., Squire, L. R., & Wixted, J. T. (2010). In search of recollection and familiarity in the hippocampus. *Journal of Cognitive Neuroscience*, *22*, 109-123.

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Memory & Cognition*, *30*, 885-892.

Watkins, M. J., LeCompte, D. C., & Kim, K. (2000). Role of study strategy in recall of mixed lists of common and rare words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 239-245.

Wells, G. L., Lindsay, R. C. L., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology*, *64*, 440-448.

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). On the selection of distractors for eyewitness lineups. *Journal of Applied Psychology*, *78*, 835-844.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*, 603-647.

Whitley, B.E., Jr., & Greenberg, M.S. (1986). The role of eyewitness confidence in juror perceptions of credibility. *Journal of Applied Social Psychology*, *16*, 387–409.

Wilding, E. L., & Rugg, M. D. (1996). An event-related potential study of recognition memory with and without retrieval of source. *Brain*, *119*, 889-905.

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

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

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source memory: A formal dual-process model and an analysis of receiver operating

characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1415-1434.

Yonelinas, A.P. (2001). Components of episodic memory: the contribution of recollection and familiarity. *Philosophical Transcripts of the Royal Society of London, Biological Sciences*, *356*, 1363–1374.

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

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition*, *5*, 418-441.

Yonelinas, A. P., & Jacoby, L. L. (1994). Dissociations of processes in recognition memory: Effects of interference and of response speed. *Canadian Journal of Experimental Psychology*, *48*, 516-534.

Yonelinas, A. P., & Jacoby, L. L. (1995). The relation between remembering and knowing as bases for recognition: Effects of size congruency. *Journal of Memory and Language*, *34*, 622-643.

Yonelinas, A. P., & Jacoby, L. L. (1996). Noncriterial recollection: Familiarity as automatic, irrelevant information. *Consciousness and Cognition*, *5*, 131-141.

## Footnotes

1. When type of metacognitive judgment was included in the ANOVA with proportion of correctly recalled words as dependent variable, it neither produced a reliable main effect, $F(2, 92) = 1.76$, nor interacted with the other factors, $F\text{s} \leq 1.58$. For such reason, we omitted it from the previous ANOVA.

2. Correct source identification can be due to guessing as well and, therefore, the proportion of recalled words whose source was correctly identified is a biased measure of recollection. Alternatively, one might compute an unbiased measure of recollection with source data as follows. This procedure is describe in detail in Appendix 4, and summarized here. Let $P(\text{S})$ and $P(\text{G})$ refer to the probability of source retrieval and correctly guessing a target's source, respectively, then the unbiased measure of the proportion of recalled words whose source was correctly identified is the following joint probability: $P(\text{Recall and source retrieval}) = [P(\text{Recall and correct source}) - P(\text{Recall}) \times P(\text{G})] / [1 - P(\text{G})]$, in which both $P(\text{Recall})$ and $P(\text{Recall and correct source})$ can be estimated from subjects' data. For an experiment with $N$ different sources, and assuming that subjects guess randomly, $P(\text{G}) = 1 / N$, and then, $P(\text{Recall and source retrieval}) = [N \times P(\text{Recall and correct source}) - P(\text{Recall})] / (N - 1)$, and similarly, source retrieval can be estimated as $P(\text{S}) = [N \times P(\text{Recall and source retrieval}) - P(\text{Recall})] / [(N - 1) \times P(\text{Recall})]$. In the present study, $N = 2$, as subjects had to decide between either red or blue colors. Notice that this procedure for computing an unbiased estimate of recollection from source data needs to assume that source retrieval only occurs for recalled items. In other words, as in Batchelder & Riefer, 1990, source memory depends on item memory—if subjects are not able to recall a target, then the implication is that they will not be able to recall the target's source. Such assumption, although very intuitive, does not always hold (Brainerd et al., 2012; Starns et al., 2008). In

addition, it assumes that guessing is always aleatory—situations in which subjects guessed red or blue due to some strategy (e.g., "I think that most recalled words were red, so I will guess red if I cannot remember the actual color") were not taken into account. All statistical analyses with such measures are presented in Appendix 4.

3. Because the confidence scale was bipolar, one might measure recollective retrieval as the proportion of recalled words that received ratings 5 (+ +) or 4 (+). The analysis using such separation method was reported in Appendix 6.

4. However, the unbiased measure of recollection from source data, the joint probability *P*(Recall and source retrieval), *under*-estimated recollection in all conditions (see Appendix 4 and Table 8).

# Appendix 1

## Dual-Recall Markov Chain

In a multiple-trial experiment in which subjects receive 3 study-test trials of form $S_1T_1$ $S_2T_2$ $S_3T_3$, each target (studied item) generates one out of eight possible patterns of correct (C) and error (E) responses across tests, namely $C_1C_2C_3$, $C_1C_2E_3$, …, $E_1E_2E_3$. The 6 parameters presented in Table 1 can be estimated from the frequency of such error-success patterns by applying a dual-recall Markov chain that contains those parameters. The states of the model are *U* (an initial no-recall state), *P* (an intermediate partial-recall state), with a correct recall state $P_C$ and an incorrect recall state $P_E$, and *L* (a terminal and absorbing criterion-recall state). The Markov process for these states consists of the following starting vector **W** and transition matrices $\mathbf{M_1}$ and $\mathbf{M_2}$:

$$\mathbf{W} = [\mathrm{L}(1), \mathrm{P_E}(1), \mathrm{P_C}(1), \mathrm{U}(1)] = [D_1, (1\text{-}D_1)R(1\text{-}J_1), (1\text{-}D_1)RJ_1, (1\text{-}D_1)(1\text{-}R)], \qquad \text{(A1)}$$

$$\mathbf{M_1} = \begin{array}{c|cccc} & \mathrm{L}(2) & \mathrm{P_E}(2) & \mathrm{P_C}(2) & \mathrm{U}(2) \\ \hline \mathrm{L}(1) & 1 & 0 & 0 & 0 \\ \mathrm{P_E}(1) & D_2 & (1\text{-}D_2)(1\text{-}J_2) & (1\text{-}D_2)J_2 & 0 \\ \mathrm{P_C}(1) & 0 & (1\text{-}J_2) & J_2 & 0 \\ \mathrm{U}(1) & D_2 & (1\text{-}D_2)R(1\text{-}J_2) & (1\text{-}D_2)RJ_2 & (1\text{-}D_2)(1\text{-}R) \end{array}, \qquad \text{(A2)}$$

$$\mathbf{M_2} = \begin{array}{c|cccc} & \mathrm{L}(3) & \mathrm{P_E}(3) & \mathrm{P_C}(3) & \mathrm{U}(3) \\ \hline \mathrm{L}(2) & 1 & 0 & 0 & 0 \\ \mathrm{P_E}(2) & D_2 & (1\text{-}D_2)(1\text{-}J_3) & (1\text{-}D_2)J_3 & 0 \\ \mathrm{P_C}(2) & 0 & (1\text{-}J_3) & J_3 & 0 \\ \mathrm{U}(2) & D_2 & (1\text{-}D_2)R\,(1\text{-}J_3) & (1\text{-}D_2)RJ_3 & (1\text{-}D_2)(1\text{-}R) \end{array}. \qquad \text{(A3)}$$

The probabilities of the 8 individual error-success patterns are obtained by multiplying the starting vector and transition matrices together. Those expressions are:

$$P(C_1C_2C_3) = D_1 + (1-D_1)RJ_1J_2J_3; \quad \text{(A4)}$$

$$P(C_1C_2E_3) = (1-D_1)RJ_1J_2(1-J_3); \quad \text{(A5)}$$

$$P(C_1E_2C_3) = (1-D_1)RJ_1(1-J_2)D_2 + (1-D_1)RJ_1(1-J_2)(1-D_2)J_3; \quad \text{(A6)}$$

$$P(C_1E_2E_3) = (1-D_1)RJ_1(1-J_2)(1-D_2)(1-J_3)\ ; \quad \text{(A7)}$$

$$P(E_1C_2C_3) = (1-D_1)R(1-J_1)D_2 + (1-D_1)R(1-J_1)(1-D_2)J_2J_3 + (1-D_1)(1-R)D_2 + (1-D_1)(1-R)(1-D_2)RJ_2J_3; \quad \text{(A8)}$$

$$P(E_1C_2E_3) = (1-D_1)R(1-J_1)(1-D_2)J_2(1-J_3) + (1-D_1)(1-R)(1-D_2)RJ_2(1-J_3); \quad \text{(A9)}$$

$$P(E_1E_2C_3) = (1-D_1)R(1-J_1)(1-D_2)(1-J_2)D_2 + (1-D_1)R(1-J_1)(1-D_2)(1-J_2)(1-D_2)J_3 + (1-D_1)(1-R)(1-D_2)R(1-J_2)D_2 + (1-D_1)(1-R)(1-D_2)R(1-J_2)(1-D_2)J_3 + (1-D_1)(1-R)(1-D_2)(1-R)D_2 + (1-D_1)(1-R)(1-D_2)(1-R)(1-D_2)RJ_3; \quad \text{(A10)}$$

$$P(E_1E_2E_3) = (1-D_1)R(1-J_1)(1-D_2)(1-J_2)(1-D_2)(1-J_3) + (1-D_1)(1-R)(1-D_2)R(1-J_2)(1-D_2)(1-J_3) + (1-D_1)(1-R)(1-D_2)(1-R)(1-D_2)R(1-J_3) + (1-D_1)(1-R)(1-D_2)(1-R)(1-D_2)(1-R). \quad \text{(A11)}$$

Maximum likelihood estimates of the 6 parameters in Table 1 are then obtained by maximizing the following likelihood function using any optimization procedure:

$$L_6 = \Pi(p_i)^{N(i)}, \quad \text{(A12)}$$

in which the $p_i$ are the 8 expressions on the right sides of Equations A4–A11, and the $N(i)$ are empirical data counts of the corresponding error-success sequences. Because 6 free parameters are estimated, the likelihood value in A12 is computed with 1 degree of freedom. A goodness-

of-fit test that evaluates the null hypothesis that learning to recall involves two processes is then obtained by computing a likelihood ratio statistic that compares the likelihood in A12 to the likelihood of the same data when all 7 observable probabilities are free to vary. That test statistic, which is asymptotically distributed as $\chi^2(1)$, is

$$G^2 = -2\ln[L_6 / L_7], \qquad \text{(A13)}$$

where $L_7$ is the likelihood of the data when all 7 observable probabilities are free to vary.

# Appendix 2

## Word Lists

**Short list (16 words)**

**High-frequency words.** Club, hand, hospital, letter, lettuce, material, point, and race.

**Low-frequency words.** Apricot, cider, fable, heroin, jade, mutiny, pier, and tomahawk.

**Medium list (30 words)**

**High-frequency words.** Audience, black, children, court, figure, hall, hotel, mouth, note, picture, plant, pudding, school, town, and write.

**Low-frequency words.** Cavern, damsel, fawn, galaxy, glacier, hostage, hump, lice, monarch, napkin, noose, soot, spike, streamer, and walrus.

**Long list (60 words)**

**High-frequency words.** Building, case, color, degree, doctor, eight, family, floor, food, friend, human, market, money, paper, party, person, plane, president, pretty, room, secretary, sound, space, spring, staff, station, table, wall, white, and woman.

**Low-frequency words.** Blacksmith, boar, bristle, brook, cedar, clam, cloak, crook, crypt, dandelion, feast, filth, fowl, ginger, grocer, knob, loot, magnet, morphine, parcel, pliers, podium, rash, seaweed, shawl, shepherd, slipper, typhoon, wand, and yoke.

# Appendix 3

## Instructions to Make Retrospective Judgments

**Remember/know/guess judgments**

"All the words you recalled on the LAST test will be presented with you in the same order as you recalled them. For each word, you will be asked to make a "REMEMBER", "KNOW", or "GUESS" judgment. Make a REMEMBER judgment (press "r") if you are able to bring back to mind a particular association, image, or something more personal from the time of study, or what the word looked like, such as its color, or sounded like when you read it to yourself during study. Make a KNOW judgment (press "k") if you have a feeling that you saw the word during study, but you are unable to bring back to mind any qualitative information about it. For example, if someone asks for your name, you would typically respond in the "know" sense without becoming consciously aware of anything about a particular event or experience. However, when asked about the last movie you saw, you would typically respond in the "remember" sense, that is, becoming consciously aware again of some aspects of the experience (e.g., the room and who you were with). Make a GUESS judgment (press "g") if you simply guessed the word during recall."

**Confidence judgments**

"All the words you recalled on the LAST test will be presented with you in the same order as you recalled them. For each word, you will be asked to make a CONFIDENCE judgment on a scale ranging from 1 to 6 as follows:

1 = VERY confident that the recalled word WAS NOT studied;

2 = MODERATELY confident that the recalled word WAS NOT studied;

3 = A LITTLE BIT confident that the recalled word WAS NOT studied;

4 = A LITTLE BIT confident that the recalled word WAS studied;

5 = MODERATELY confident that the recalled word WAS studied;

6 = VERY confident that the recalled word WAS studied.

Please be extremely cautious about using the end points of 1 and 6 and use them only if you are 100% positive about your answer. If you use 1 or 6, that means you cannot possibly make a mistake. That is, you are so confident in your answer that you would be willing to testify in a court of law, even in a life-or-death situation."

**Source judgments**

"All the words you recalled on the LAST test will be presented with you in the same order as you recalled them. For each word, you will be asked to make a SOURCE judgment. During the study phases, the words were presented in either RED or BLUE. Make a RED judgment (press "r") if the recalled word was presented in red during the study phases. Make a BLUE judgment (press "b") if the recalled word was presented in blue during the study phases."

## Appendix 4

### Additional Analyses of the Source Judgments Data

Correct source identifications can be made via source retrieval and guessing. Consequently, the proportion of recalled words whose source was correctly identified, $P(\text{Rc} \wedge \text{Cs})$, is a biased measure of recollection. In this section I present alternative measures that separate source retrieval from guessing. Let $P(\text{S})$ and $P(\text{G})$ be the probability of source retrieval and correctly guessing a target's source, respectively, which are jointly independent events. Then, the unbiased measure of the proportion of recalled words whose source was correctly identified is the joint probability of correct recall and source retrieval, $P(\text{Rc} \wedge \text{S})$, which can be expressed in terms of $P(\text{S})$, $P(\text{G})$, and two known measures, namely the probability of correct recall, $P(\text{Rc})$, and the joint probability of correct recall and correct source identification, $P(\text{Rc} \wedge \text{Cs})$, as follows:

$$P(\text{Rc} \wedge \text{S}) = [P(\text{Rc} \wedge \text{Cs}) - P(\text{Rc}) \times P(\text{G})] / [1 - P(\text{G})]. \quad \text{(A14)}$$

In an experiment with $N$ distinct sources, one might assume that subjects randomly guess the source of recalled targets when source retrieval fails, and thus, $P(\text{G}) = 1 / N$. The revised Equation A14,

$$P(\text{Rc} \wedge \text{S}) = [N \times P(\text{Rc} \wedge \text{Cs}) - P(\text{Rc})] / (N - 1), \quad \text{(A15)}$$

provides unbiased estimates of recollective retrieval from source data. Similarly, source retrieval, $P(\text{S})$, can be estimated from known measures as follows,

$$\begin{aligned} P(\text{S}) &= [N \times P(\text{Rc} \wedge \text{S}) - P(\text{Rc})] / [(N - 1) \times P(\text{Rc})] \\ &= \{N \times [[N \times P(\text{Rc} \wedge \text{Cs}) - P(\text{Rc})] / (N - 1)] - P(\text{Rc})\} / [(N - 1) \times P(\text{Rc})] \\ &= [N^2 \times P(\text{Rc} \wedge \text{Cs}) - (2N - 1) \times P(\text{Rc})] / [(N - 1)^2 \times P(\text{Rc})]. \quad \text{(A16)} \end{aligned}$$

Correct identification of a target's source due to random guessing becomes an ever more infrequent event as $N \to \infty$ and, therefore, $\lim_{N \to \infty} P(\text{Rc} \wedge \text{S}) = P(\text{Rc} \wedge \text{Cs})$ and $\lim_{N \to \infty} P(\text{S}) =$

$P(\text{Rc} \wedge \text{Cs}) / P(\text{Rc})$. In the present study, however, $N = 2$, which leads to the following simplified versions of Equations A15 and A16, respectively,

$$P(\text{Rc} \wedge \text{S}) = 2P(\text{Rc} \wedge \text{Cs}) - P(\text{Rc}), \quad \text{(A17)}$$

$$P(\text{S}) = P(\text{Rc} \wedge \text{S}) / P(\text{Rc}). \quad \text{(A18)}$$

When these statistics were computed for the present experiment (see Table 8), they revealed a very similar pattern to the unadjusted source accuracy data (cf. Tables 6 and 8). $P(\text{S})$, as source accuracy, did not show reliable effects of list length and word frequency, $F\text{s} \leq 2.04$. $P(\text{Rc} \wedge \text{S})$, however, showed reliable effects of list length and word frequency when $P(\text{Rc} \wedge \text{S})$ and the joint probability of correct recall and incorrect source identification, $P(\text{Rc} \wedge \text{I})$, were submitted to a 3 (list length: short, medium, long) x 2 (word frequency: low, high) x 2 (source accuracy: $P(\text{Rc} \wedge \text{S})$, $P(\text{Rc} \wedge \text{I})$) repeated measures ANOVA. Specifically, the ANOVA produced a main effect of list length, $F(2, 76) = 41.6$, $MSE = .01$, $\eta^2_p = .52$, and an interaction between list length and word frequency, $F(2, 76) = 5.36$, $MSE = .01$, $\eta^2_p = .12$. The latter interaction indicated that, on average, both $P(\text{Rc} \wedge \text{S})$ and $P(\text{Rc} \wedge \text{I})$ were higher for high frequency words than low frequency words, but such effect was only reliable in the long list length condition.

However, contrary to the $P(\text{Rc} \wedge \text{Cs})$ measure of recollection shown in Figures 5 and 6, $P(\text{Rc} \wedge \text{S})$ greatly *under*-estimated recollective recall, relative to all other measurement methods (cf. Figure 5 and Table 8). In addition, at the level of individuals, the correlations shown in Table 9 indicate that none of the additional source measures, namely $P(\text{S})$, $P(\text{Rc} \wedge \text{S})$, and $P(\text{Rc} \wedge \text{I})$, were reliable predictors of the dual-recall model's statistics ($p\text{s} > .05$).

## Appendix 5

## Analysis of Output Dependency

Because vincentised output positions partition subjects' free recall protocol into four parts, they allow an investigation of whether measurements of items output earlier in the free recall protocol (e.g., VO1) are associated with measurements of items output later in the free recall protocol (e.g., VO4). Specifically, for each list length condition (short, medium, and long), I computed all 6 possible correlations between vincentised output positions, namely $r_{VO_i,VO_j}$, for $i \neq j$ and $i,j = \{1, .., 4\}$, using each of the following type of measure of items output on the last trial: MTE, mean confidence ratings, mean proportion of remember responses, and mean source accuracy. Reliable output dependencies (the mean of all *r*s) indicate that the measure of items output earlier in the free recall protocol can predict the measure of subsequent items. In the case of retrospective judgments, for example, reliable output dependencies might suggest that subjects rely on previous judgments, rather than introspection, to make new judgments. The results are shown in Figure 11, in which mean output dependencies are plotted as a function of both list length and type of measure. Retrospective metacognitive judgments (remember and confidence judgments) showed reliable output dependencies that interacted with list length. Specifically, for both remember and confidence judgments, output dependencies increased as list length increased. Nonetheless, on average, output dependencies were higher for confidence judgments (.69) than for remember judgments (.39), and both source accuracy (.21) and MTE (.02) did not show reliable output dependencies.

Therefore, the results of the analysis of output dependencies revealed differences among the three types of retrospective judgments. Subjects' confidence ratings, in particular, were highly correlated across output positions. In fact, confidence judgments made to items output

earlier in the free recall protocol accounted for roughly half (48%) of the confidence judgments made to items output later. This suggests, for instance, that if subjects are very confident about the items output in the beginning of the free recall protocol, they will tend to be very confident about the items output later on as well, regardless of underlying retrieval processes and whether items output later are associated to weak or strong memory traces. Although to a lesser degree, this results was also observed with the proportion of remember judgments made to recalled targets. Specifically, the proportion of remember judgments made to items output earlier in the free recall protocol accounted for 15% of the proportion of remember judgments made to items output later. In addition, for both confidence and remember judgments, output dependencies increased as the number of studied items increased. This suggests that, as the number of to-be-recalled items increases, subjects rely ever less on introspections, and more on prior judgments, to make both confidence and remember/know judgments to items recalled latter in free recall protocols, which represents an obvious challenge to the use of such judgments as methods of separating dual processes in free recall. Source accuracy, however, did not show reliable output dependencies.

## Appendix 6

### Additional Analyses of Confidence Judgments

The decomposition of recall into its recollective and nonrecollective components via confidence judgments can be made by assuming that the highest confidence level that the item was studied (6, or +++) signals the recollective form of retrieval, while lower confidence levels that the item was studied (5 and 4, or ++ and +) signal nonrecollective retrieval. This separation method differs from the one previously reported in that nonrecollective retrieval does not include confidence ratings lower than 4, namely judgments that the recalled item *was not* studied. Nonetheless, the results were very similar to the ones previously reported. A 3 (list length: short, medium, long) x 2 (word frequency: low, high) x 2 (type of process: recollective, nonrecollective) revealed a main effect of type of process, $F(1, 27) = 12.6$, $MSE = .20$, $\eta^2_p = .32$, list length, $F(2, 54) = 47.2$, $MSE = .01$, $\eta^2_p = .64$, word frequency, $F(1, 27) = 14.3$, $MSE = .01$, $\eta^2_p = .35$, and small but reliable interactions between type of process and list length, $F(2, 54) = 3.6$, $MSE = .03$, $\eta^2_p = .12$, and type of process and word frequency, $F(2, 54) = 4.6$, $MSE = .01$, $\eta^2_p = .14$. The interaction between type of process and list length indicated that the effects of list length on recollective retrieval (max confidence) were larger than on nonrecollective retrieval (lower confidence), whereas the interaction between type of process and word frequency indicated that only recollective retrieval differed between low- and high-frequency words. As before, however, this method of measuring dual processes in free recall also overestimated recollective retrieval, and thus to underestimated nonrecollective retrieval, relative to the dual-retrieval model. In addition, analysis of individual data did not reveal any reliable correlation between statistics from the dual-retrieval model and confidence judgments ($|r_s| \leq .35$), as the ones reported in Table 7.

Table 1

*Parameters of the Dual-Recall Model*

| Parameter | State | Definition |
|---|---|---|
| | | Recollective retrieval |
| Direct access | | |
| $D_1$ | L | The probability that an item's verbatim trace can be accessed after the first study trial. |
| $D_2$ | L | The probability that an item's verbatim trace can be accessed after the second or third study trials if it could not be accessed following the first study trial. |
| | | Nonrecollective retrieval |
| Reconstruction | | |
| $R$ | P | The probability that an item can be reconstructed from partially identifying information after any study trial if it can neither be directly accessed nor reconstructed following prior study trials. |
| Familiarity judgment | | |
| $J_1$ | $P_C$ | The probability that a reconstructed item is judged familiar to be output following the first study trial. |
| $J_2$ | $P_C$ | The probability that a reconstructed item is judged familiar to be output following the second |

| | | |
|---|---|---|
| | | study trial. |
| $J_3$ | $P_C$ | The probability that a reconstructed item is judged familiar to be output following the third study trial. |

Table 2

*Mean Recall Accuracy Measures as a Function of Word Frequency and List Length*

| Condition | Proportion recalled | | | MTR |
|---|---|---|---|---|
| | $T_1$ | $T_2$ | $T_3$ | |
| High frequency | | | | |
| Short list | .35 (.22) | .68 (.23) | .78 (.19) | 1.80 (.55) |
| Medium list | .28 (.18) | .48 (.20) | .63 (.21) | 1.37 (.52) |
| Long list | .21 (.11) | .40 (.15) | .51 (.18) | 1.13 (.40) |
| Low frequency | | | | |
| Short list | .41 (.17) | .69 (.19) | .76 (.17) | 1.86 (.44) |
| Medium list | .28 (.17) | .49 (.21) | .61 (.21) | 1.38 (.51) |
| Long list | .16 (.08) | .32 (.15) | .43 (.17) | .91 (.36) |

*Note.* MTR = Mean total correct recalls across trials per item. Standard deviation in parentheses.

Table 3

*Maximum Likelihood Estimates of the Parameters of the Dual-Recall Model as a Function of Experimental Conditions*

| | Recollective retrieval | | | Nonrecollective retrieval | | | | |
|---|---|---|---|---|---|---|---|---|
| Condition | $D_1$ | $D_2$ | Mean $D$ | $R$ | $J_1$ | $J_2$ | $J_3$ | Mean $J$ |
| High frequency | | | | | | | | |
| Short list | .23 | .22 | .23 | .39 | .38 | .75 | .73 | .62 |
| Medium list | .15 | .18 | .17 | .24 | .49 | .59 | .65 | .57 |
| Long list | .12 | .18 | .15 | .17 | .59 | .54 | .47 | .54 |
| Low frequency | | | | | | | | |
| Short list | .20 | .18 | .19 | .49 | .52 | .73 | .70 | .65 |
| Medium list | .15 | .14 | .15 | .29 | .51 | .64 | .63 | .59 |
| Long list | .08 | .11 | .09 | .19 | .45 | .53 | .52 | .50 |

Table 4

*Mean Proportion Recalled on the Last Trial that Received "Remember" or "Know" Judgments as a Function of Experimental Conditions*

| Condition | Remember | Know |
|---|---|---|
| High frequency | | |
| Short list | .57 (.23) | .21 (.20) |
| Medium list | .48 (.28) | .16 (.15) |
| Long list | .40 (.22) | .14 (.10) |
| Low frequency | | |
| Short list | .56 (.23) | .20 (.17) |
| Medium list | .50 (.23) | .13 (.13) |
| Long list | .33 (.16) | .12 (.10) |

*Note*. Standard deviation in parentheses.

Table 5

*Mean Confidence Measures as a Function of Experimental Conditions*

| Condition | Confidence rating | Proportion recalled per confidence rating | | | | | |
|---|---|---|---|---|---|---|---|
| | | + + + (6) | + + (5) | + (4) | - (3) | - - (2) | - - - (1) |
| High frequency | | | | | | | |
| Short list | 5.56 (.66) | .57 (.34) | .17 (.23) | .06 (.12) | .01 (.05) | .01 (.03) | .00 (.00) |
| Medium list | 5.54 (.54) | .46 (.29) | .15 (.20) | .04 (.10) | .01 (.02) | .00 (.00) | .00 (.03) |
| Long list | 5.54 (.59) | .34 (.24) | .11 (.16) | .03 (.05) | .01 (.02) | .00 (.01) | .00 (.01) |
| Low frequency | | | | | | | |
| Short list | 5.57 (.70) | .56 (.33) | .14 (.23) | .03 (.09) | .00 (.02) | .02 (.07) | .00 (.02) |
| Medium list | 5.70 (.50) | .49 (.30) | .10 (.18) | .02 (.05) | .00 (.01) | .00 (.00) | .00 (.03) |
| Long list | 5.54 (.69) | .31 (.22) | .09 (.14) | .02 (.04) | .00 (.01) | .00 (.01) | .00 (.01) |

*Note*. Standard deviation in parentheses.

Table 6

*Mean Source Accuracy Measures as a Function of Experimental Conditions*

| Condition | Source accuracy | Proportion recalled per source accuracy | |
|---|---|---|---|
| | | Correct source | Incorrect source |
| High frequency | | | |
| Short list | .64 (.25) | .46 (.22) | .28 (.21) |
| Medium list | .65 (.19) | .40 (.19) | .21 (.13) |
| Long list | .60 (.16) | .31 (.14) | .20 (.09) |
| Low frequency | | | |
| Short list | .67 (.25) | .52 (.22) | .24 (.16) |
| Medium list | .64 (.19) | .37 (.15) | .21 (.14) |
| Long list | .59 (.17) | .24 (.10) | .18 (.09) |

*Note*. Standard deviation in parentheses.

Table 7

*Correlations between Individualized Statistics of the Dual-Recall Model and Retrospective Measures of Dual Processes across all Experimental Conditions*

| | Remember/Know | | Confidence | | Source | |
|---|---|---|---|---|---|---|
| Dual-recall model | Recollective | Nonrecollective | Recollective | Nonrecollective | Recollective | Nonrecollective |
| Parameters | | | | | | |
| $D_1$ | .65* | .11 | .22 | .16 | .52* | .21 |
| $D_2$ | .56* | -.05 | .21 | .07 | .39* | .28 |
| Mean *D* | .62* | .01 | .25 | .12 | .49* | .27 |
| *R* | .49* | .18 | .23 | .33 | .30 | .30 |
| $J_1$ | -.54* | .13 | .13 | -.34 | -.08 | -.20 |
| $J_2$ | -.32 | .20 | .17 | -.09 | .18 | -.12 |
| $J_3$ | -.19 | .30 | .20 | .17 | .15 | -.02 |
| Mean *J* | -.42* | .26 | .20 | -.09 | .10 | -.14 |
| Recall on $T_3$ | | | | | | |
| Recollective | .57* | -.05 | .19 | .10 | .40* | .26 |
| Nonrecollective | -.08 | .09 | .28 | .06 | .11 | .05 |

*Note*. The definition of recollective and nonrecollective retrieval for each type of retrospective judgment was the same as the one used to generate Figure 5 (joint probabilities) and runs as follows. For remember/know judgments, recollective retrieval was the proportion of recalled words that received a "remember" judgment, and nonrecollective retrieval was the proportion of recalled words that received either a "know" or "guess" judgment. For confidence ratings, recollective retrieval was the proportion of recalled words that received the maximum confidence rating, while nonrecollective retrieval was the proportion of recalled words that received confidence ratings lower than the

maximum. For source accuracy, recollective retrieval was the proportion of recalled words whose source was correctly identified, while nonrecollective retrieval was the proportion of recalled words whose source was incorrectly identified.

* $p < .05$

Table 8

*Additional Mean Source Accuracy Measures as a Function of Experimental Conditions*

| Condition | P(S) | P(G)* | P(Rc and S) | P(Rc and G) = P(Rc and I) |
|---|---|---|---|---|
| High frequency | | | | |
| Short list | .28 (.49) | 0.5 | .19 (.37) | .28 (.21) |
| Medium list | .31 (.27) | 0.5 | .20 (.25) | .21 (.13) |
| Long list | .19 (.32) | 0.5 | .10 (.17) | .20 (.09) |
| Low frequency | | | | |
| Short list | .34 (.49) | 0.5 | .29 (.35) | .24 (.16) |
| Medium list | .27 (.38) | 0.5 | .15 (.22) | .21 (.14) |
| Long list | .17 (.34) | 0.5 | .06 (.13) | .18 (.09) |

*Note*. P(S) = Probability of source retrieval, P(G) = Probability of guessing the correct source when source retrieval fails, P(Rc and S) = Joint probability of recall and source retrieval, P(Rc and G) = Joint probability of recall and guessing the correct source when source retrieval fails, and P(Rc and I) = Joint probability of recall and incorrect source identification. P(Rc and G) = P(Rc and I) because G = 1/2. Standard deviation in parentheses. *Fixed values

Table 9

*Correlations between Individualized Statistics of the Dual-Recall Model and Additional Source Measures*

| | Source measures | | |
|---|---|---|---|
| Dual-recall model | P(S) | P(Rc and S) | P(Rc and G) = P(Rc and I) |
| Parameters | | | |
| $D_1$ | .15 | .30 | .21 |
| $D_2$ | .03 | .16 | .28 |
| $R$ | -.07 | .08 | .30 |
| $J_1$ | .10 | .04 | -.20 |
| $J_2$ | .22 | .21 | -.12 |
| $J_3$ | .10 | .13 | -.02 |
| Recall on $T_3$ | | | |
| Recollective | .04 | .18 | .26 |
| Nonrecollective | .03 | .07 | .05 |

*Note*. P(S) = Probability of source retrieval, P(G) = Probability of guessing the correct source when source retrieval fails, P(Rc and S) = Joint probability of recall and source retrieval, P(Rc and G) = Joint probability of recall and guessing the correct source when source retrieval fails, and P(Rc and I) = Joint probability of recall and incorrect source identification. P(Rc and G) = P(Rc and I) because G = 1/2.

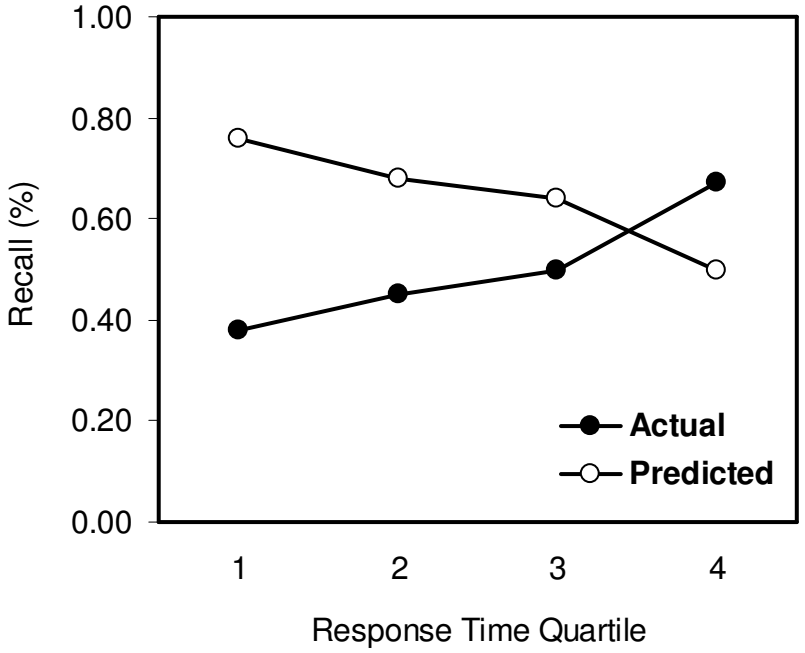


*Figure 1*. Actual and predicted recall as a function of time (in quartiles) to output answers to general knowledge questions in an experiment reported by Benjamin, Bjork, and Schwartz (1998).

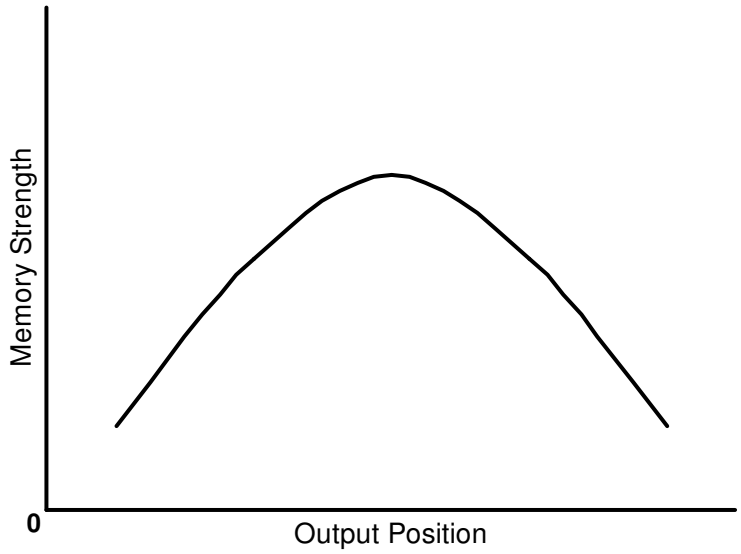


*Figure 2*. Hypothetical relationship between an item's output position and its memory strength. The figure illustrates the cognitive triage effect that items are output in a weak → strong → weak fashion during free recall.

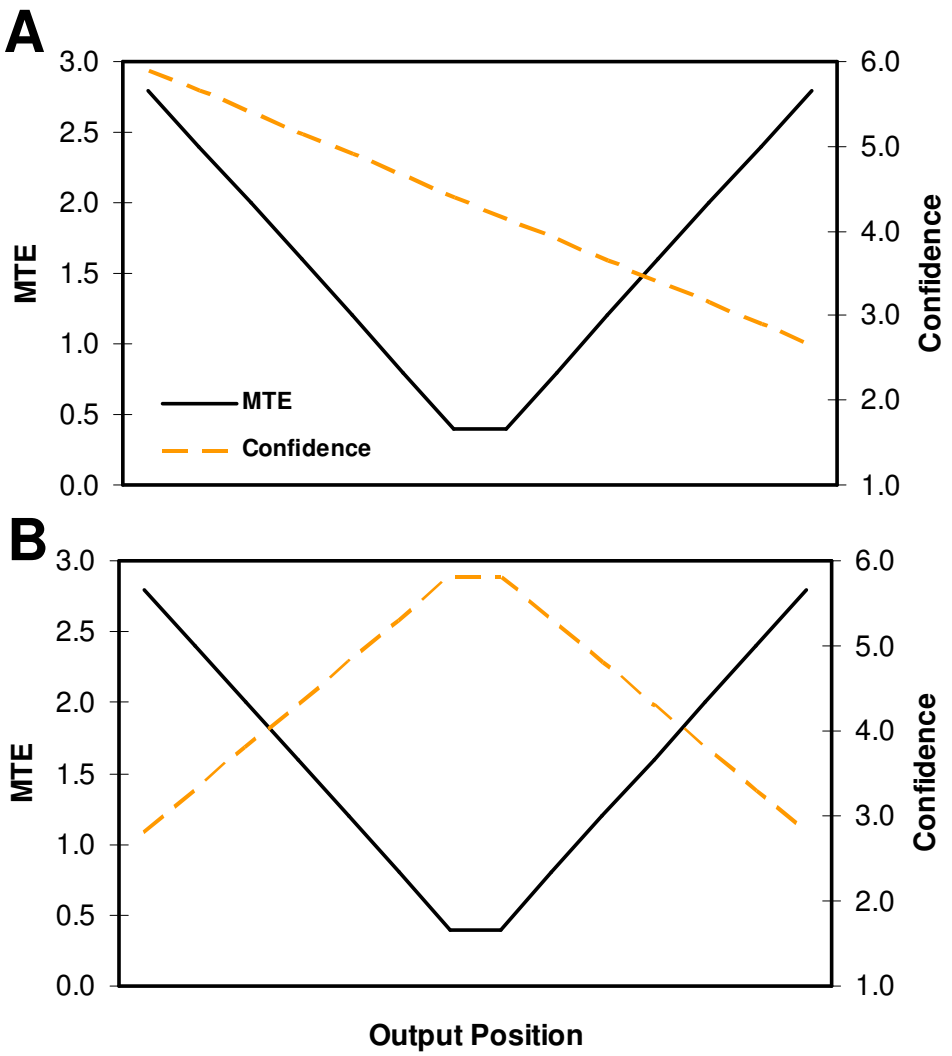


*Figure 3*. Hypothetical relationship between confidence (6 = *Very* confident that the item *was* studied, ..., 4 = *A little bit* confident that the item *was* studied, 3 = *A little bit* confident that the item *was not* studied, ..., 1 = *Very* confident that the item *was not* studied) and the mean total number of errors on previous tests (MTE) for items recalled after 3 recall tests. Panel A illustrates the hypothesis that subjects cannot monitor the distribution of the strength of memory traces across output position (weak → strong → weak), measured via the MTE statistic, and rely on beliefs about how memory strength is distributed across output position (strong → weak) to make confidence judgments. Panel B illustrates the hypothesis that subjects can monitor the distribution of the strength of memory traces across output

position (weak → strong → weak) and, therefore, they make confidence judgments across output position accordingly (weak → strong → weak).

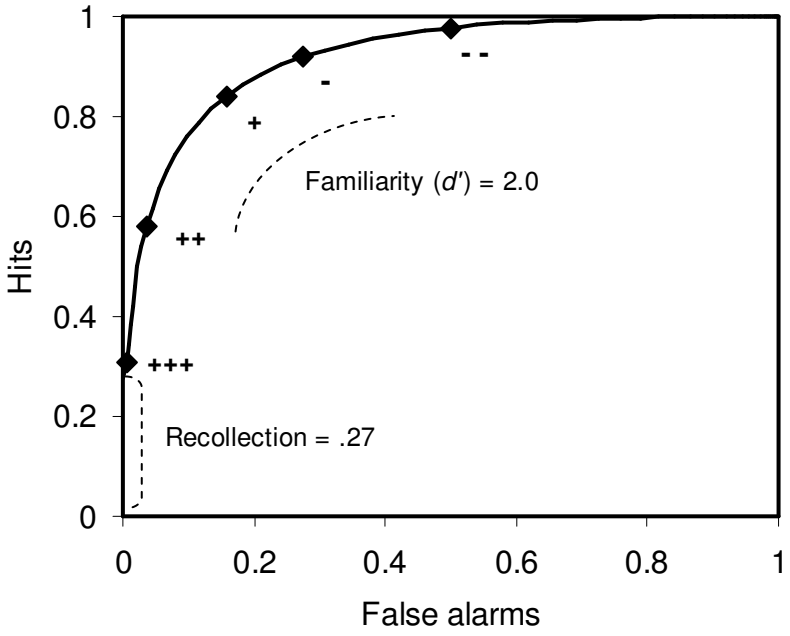


*Figure 4*. Hypothetical receiver operating characteristic (ROC) curve. + signs indicate confidence levels that a test probe was studied and – signs indicate confidence levels that a test probe was not studied. The estimates of recollection (intercept of the ROC curve) and familiarity (deflection of the ROC curve) were computed according to Yonelinas'(1994) signal detection dual-process model.

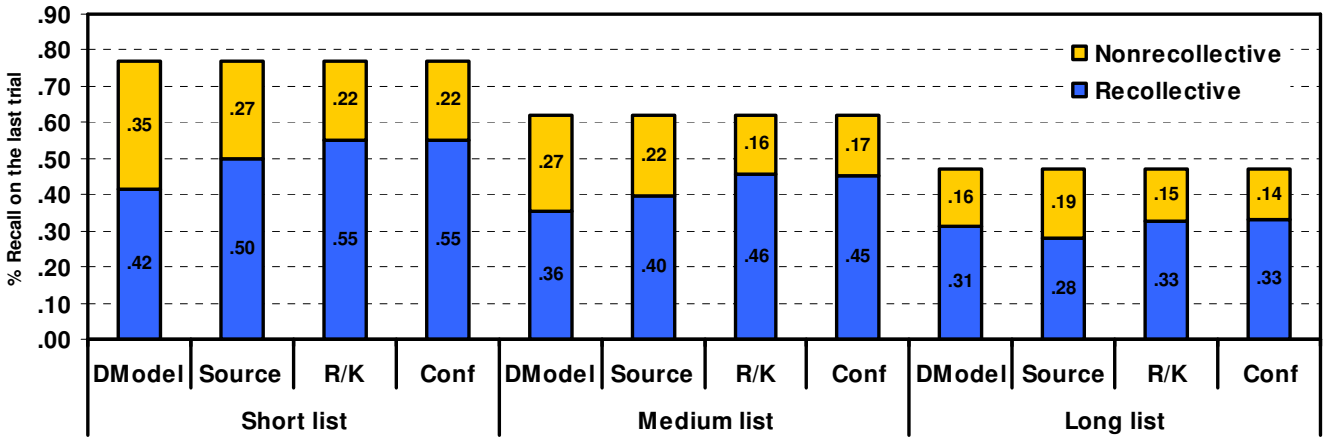


*Figure 5*. The proportion of correctly recalled items on the last trial as a function of list length and four methods of separating recollective retrieval from nonrecollective retrieval, namely the dual-recall model (DModel), source accuracy (Source), remember/know judgments (R/K), and confidence ratings (Conf).

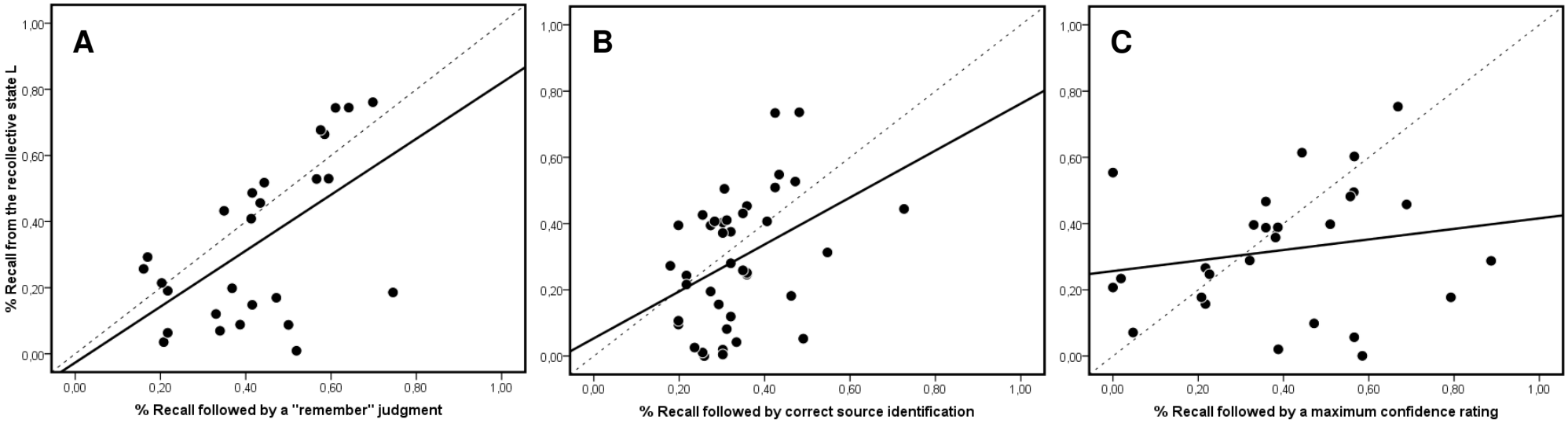


*Figure 6.* Correlations between recall in the recollective state L on the last trial, as measured by the dual-recall model, and the following three retrospective measures of recollective retrieval: recall followed by either remember judgments (Panel A), or correct source identification (Panel B), or maximum confidence (Panel C). The dashed line is the identity (perfect calibration). The solid line is the linear function that best describes the relationship between the two variables. In Panel A, the linear function explains 33% of the variability. In Panel B, the linear function explains 16% of the variability. In Panel C, the linear function explains 4% of the variability.

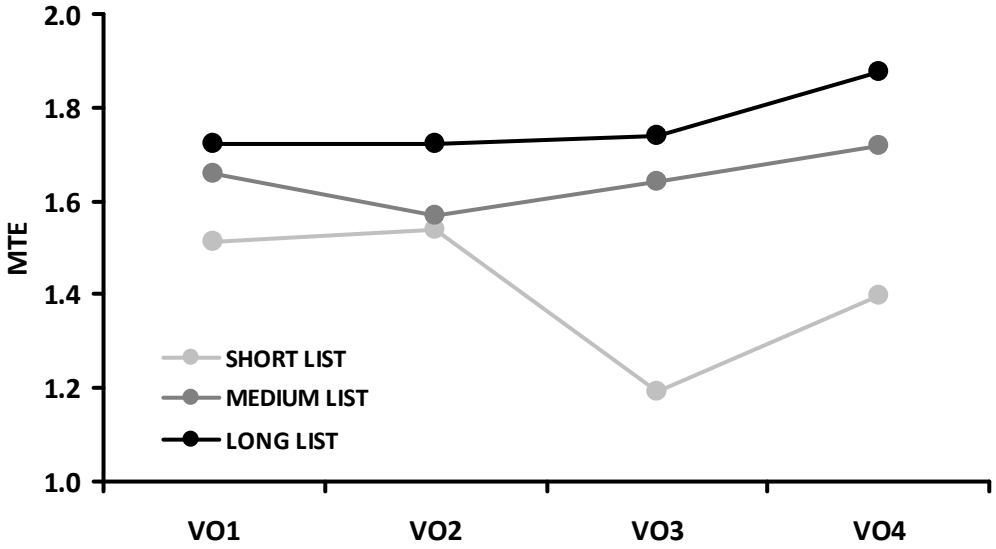


*Figure 7.* Mean total number of errors (MTE) of items recalled on trial 3 as a function of list length and vincentised output position.

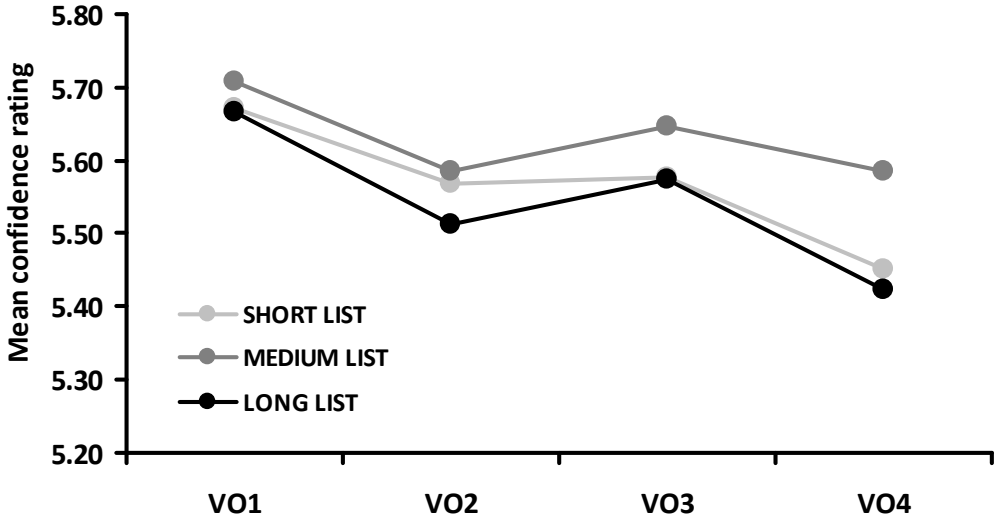


*Figure 8.* Mean confidence ratings for items recalled on trial 3 as a function of list length and vincentised output position.

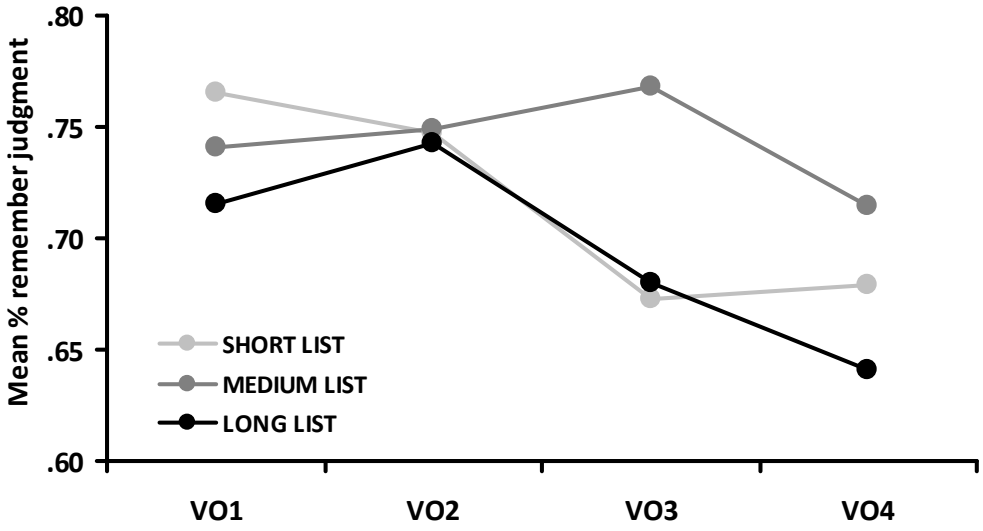


*Figure 9.* Mean proportion of remember judgments for items recalled on trial 3 as a function of list length and vincentised output position.

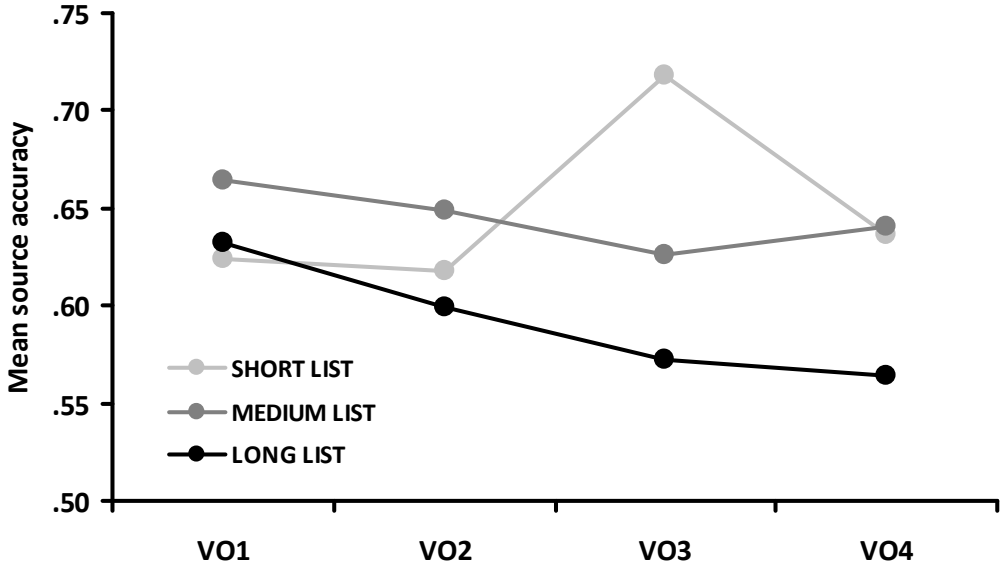


*Figure 10.* Mean source accuracy for items recalled on trial 3 as a function of list length and vincentised output position.

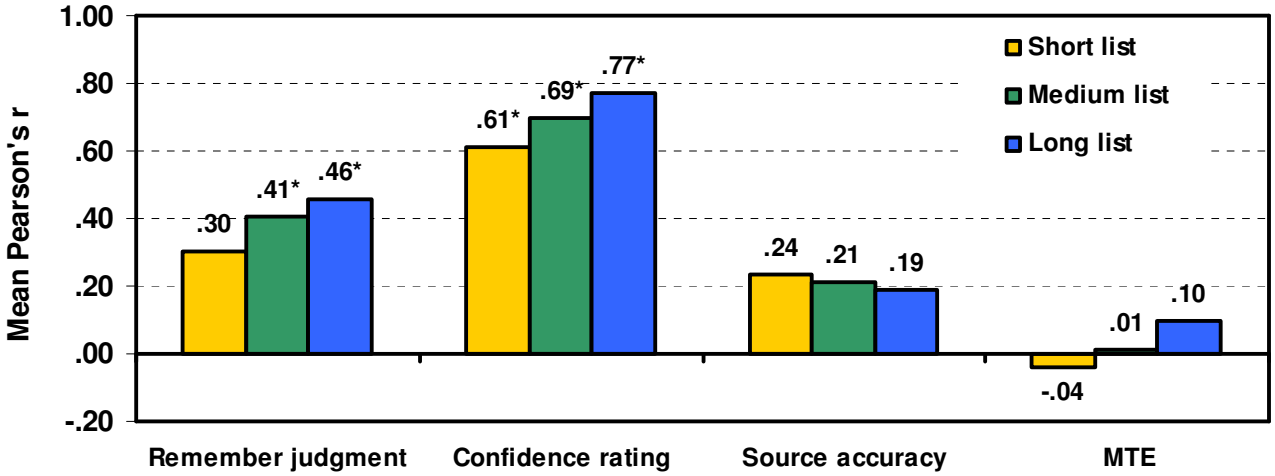


*Figure 11.* Output dependencies as a function of list length and type of measure for items recalled on trial 3. Asterisks indicate reliable output dependencies.