The role of phantom recollection in false recall

Tammy A. Marche · C. J. Brainerd

Abstract Although high levels of phantom recollection (illusory vivid experience of the prior "presentation" of unpresented items) have been found for false recognition, little is known about phantom recollection in recall. We examined this issue with Deese/Roediger–McDermott lists using two paradigms: repeated recall and conjoint recall. High levels of phantom recollection were observed with both standard behavioral measures and the parameters of fuzzy-trace theory's dual-recall model. In addition, phantom recollection and the true recollection that accompanies presented items appear to involve different retrieval processes, because they were dissociated by manipulations such as number of recall tests and list strength.

Keywords False memory · Phantom recollection · Dual-retrieval models · Fuzzy-trace theory

*Phantom recollectio*n is the illusory counterpart of recollective phenomenology—specifically, the realistic conscious experience of the prior presentation of nonpresented material (Brainerd, Wright, Reyna, & Mojardin, 2001). Phantom recollection has been detected at high levels in certain types of false recognition (e.g., Brainerd et al., 2001; Roediger & McDermott, 1995), but little is known about phantom recollection in recall. For instance, with Deese/Roediger–McDermott

(DRM; Deese, 1959; Roediger & McDermott, 1995) lists, phantom recollection is associated with 60%–70% of false alarms to critical distractors. That is, when *doctor* is falsely accepted after studying the list *nurse*, *sick*, *ill*, *hospital*, . . . , measurements of the accompanying phenomenology (e.g., remember/know or conjoint recognition) show that participants recollect *doctor*'s "presentation" on this list 60%–70% of the time. However, it is quite conceivable that phantom recollection of an item requires that the item be physically presented on a memory test, which occurs in recognition but not recall.

To study phantom recollection in recall, we used DRM lists, because those materials have already been found to produce high levels of illusory vivid phenomenology in recognition. We implemented a converging-operations strategy in which phantom recollection was measured in different ways, using different recall paradigms—namely, repeated (Experiment 1) and conjoint (Experiment 2) recall—to determine (a) the extent to which phantom recollection occurs in false recall and (b) whether true recollection and phantom recollection can be dissociated by experimental manipulations, as they have been for recognition. Before reporting our research, we will consider two topics that have influenced the design and predictions of these experiments: fuzzy-trace theory's explanation of false recall, and dual-recall theory.

Fuzzy-trace theory (FTT) and false memory

FTT provides a dual-process account of false recall (Brainerd, Aydin, & Reyna, in press; Brainerd & Reyna, 2005; Lampinen, Neuschatz, & Payne, 1998; Payne, Elie, Blackwell, & Neuschatz, 1996). The primary assumption of dual-recall models is that items are recalled by means of two dissociated retrieval operations—direct access and reconstruction—along

T. A. Marche (✉)
Department of Psychology, St. Thomas More College,
1437 College Drive,
Saskatoon, Saskatchewan S7N 0W6, Canada
e-mail: tmarche@stmcollege.ca

C. J. Brainerd
Department of Human Development, Cornell University,
Ithaca, NY, USA

with a judgment operation that assesses the products of recon-struction (e.g., Brainerd, Payne, Wright, & Reyna, 2003; Brainerd & Reyna, 2010; Brainerd, Reyna, & Howe, 2009; Brainerd, Wright, Reyna, & Payne, 2002; Reyna & Mills, 2007; also see Barnhardt, Choi, Gerkens, & Smith, 2006). Although there are numerous dual-process conceptions of recognition (e.g., Atkinson & Juola, 1973, 1974; Atkinson & Westcourt, 1975; Gillund & Shiffrin, 1984; Horton, Pavlick, & Moulin-Julian, 1993), this is not the case with recall. Dual-process theories of recall were initially developed to explain effects such as cognitive triage (Brainerd & Reyna, 2001; Marche, Howe, Lane, Owre, & Briere, 2009) and the output positions of intrusions in false recall (Barnhardt et al., 2006; Lampinen et al., 1998; Payne et al., 1996).

FTT's dual-process account of false recall (Brainerd & Reyna, 2005; Lampinen et al., 1998; Payne et al., 1996; Brainerd & Reyna, 2005) relies on the following principles: (a) parallel storage of verbatim and gist traces, (b) dissoci-ated retrieval of both types of traces, (c) different time courses of verbatim and gist memory, and (d) opponent judgments about false memory items. According to FTT, the processing and storage of verbatim traces occurs simul-taneously with the processing and storage of the gist of experience. Here, research has indicated that statistical and experimental dissociations occur between true recall/recog-nition of presented material and false recall/recognition of material that preserves its meaning (e.g., Brainerd & Reyna, 2005; Reyna & Kiernan, 1994, 1995). Also, many studies have indicated that verbatim memories become inaccessible more rapidly with time than do gist memories (e.g., Murphy & Shapiro, 1994), which means that as time goes by, reliance on gist retrieval increases, and so does false memory. This also means that the statistical independence and experimen-tal dissociation that are initially observed between true- and false-memory responses will be replaced with depen-dency and association as time passes, findings that have been reported in a number of studies (e.g., Brainerd & Reyna, 1996).

As for opposing judgments about false memory items, FTT assumes that true memory is initially due to verbatim traces and then shifts toward gist over time. Retrieval of verbatim traces induces vivid, item-specific phenomenology—the rec-ollective form of retrieval. However, targets may induce gist retrieval instead, which leads to the nonrecollective form of retrieval. Experimental evidence that episodic memories are sometimes accompanied by mental reinstatement of realistic details of prior experience (recollective retrieval) and some-times are not (nonrecollective retrieval) has been available for some time (Strong, 1913). On a recall test, both retrieval operations are available to support true memory; that is, targets can be recalled via direct access to verbatim traces or via reconstruction from gist (Brainerd et al., 2002). With false memory, however, direct access and reconstruction are

opponent processes, with reconstruction inducing false mem-ory and direct access suppressing it, through a process known as *recollection rejection*.

According to FTT, targets are typically recalled via direct access, at least initially, and produce recollective phenome-nology, whereas false-but-gist-consistent items are recon-structed and typically produce familiarity phenomenology. However, under certain conditions, gist-based reconstruc-tion supports illusory vivid phenomenology that emulates verbatim-based recollective phenomenology, a phenomenon known as *phantom recollection*. According to FTT's ac-count of semantic false recognition (Brainerd et al., 2001; Reyna, 1998; Reyna & Brainerd, 1995), when many items that share meaning are studied, like those in the DRM procedure, the gist memories of those meanings are very strong. Strongly cueing this meaning at study causes gist memories to take on phenomenological vividness and inten-sity when they are retrieved at test (Reyna & Lloyd, 1997). The administration of distractors that are excellent retrieval cues for those gist memories has been shown to produce illusions of item-specific recollection (i.e., ersatz verbatim traces are formed at retrieval; Brainerd et al., 2001; Lampinen, Meier, Arnal, & Leding, 2005). That is, critical distrac-tors provide the physical stimuli for redintegration to coalesce around. Phantom recollection is well established in recognition experiments that satisfy these conditions (e.g., Roediger & McDermott, 1995; Singer & Remillard, 2008; Stahl & Klauer, 2009), but little is known about it in false recall.

It is quite conceivable that phantom recollection could be negligible in recall experiments that satisfy the same con-ditions, owing to a key difference between recognition and recall: In recognition, false memory items are physically presented in memory tests, which may be necessary for triggering phantom recollection. There are a number of other reasons to expect that the phenomenological experi-ence of false recall may be less vivid than that of false recognition. First, according to dual-recall theory (Barnhardt et al., 2006; Brainerd et al., 2009; Payne et al., 1996), early output is dominated by direct access to verbatim traces and recollective phenomenology, but later output is dominated by reconstruction from gist and nonrecollective phenome-nology. Intrusions of semantic associates, in both standard and DRM recall, occur chiefly at the end of output sequen-ces (e.g., Brainerd, Reyna, Harnishfeger, & Howe, 1993; Payne et al., 1996; Roediger & McDermott, 1995), suggest-ing that they are reconstructions not accompanied by recol-lective phenomenology. Second, introspective reports of recall phenomenology have indicated that intrusions of se-mantic associates may not induce phantom recollection (Brainerd et al., 2001). Third, a process model of phantom recollection, proposed by Reyna and Titcomb (1997), assumes that when recognition probes are presented that

are excellent retrieval cues for strong gist memories, such as DRM semantic associates, fragmented surface information from disintegrated verbatim traces redintegrates, producing illusory verbatim traces when distractors are physically presented (i.e., recognition tests). In recall, however, there are no physical stimuli for redintegration to coalesce around, which may lead to very low levels of phantom recollection. However, according to FTT (Brainerd et al., 2003), because of the reconstructive nature of the critical distractor, that distractor becomes available to conscious awareness, which provides the "stimuli" needed for the formation of simulated verbatim traces.

In the present study, we used a common methodology for studying false recall and phantom recollection in recognition, the DRM paradigm. When nonpresented semantic associates appear as distractors on a recognition test, the false-alarm rate approaches (Payne et al., 1996; Roediger & McDermott, 1995), and sometimes exceeds (Brainerd et al., 2001), the hit rate for targets, and participants report high levels of recollective experience for false alarms (e.g., Gallo, McDermott, Percer, & Roediger, 2001; Payne et al., 1996; Roediger & McDermott, 1995). On free recall tests, these semantic associates intrude in a substantial proportion of protocols (Payne et al., 1996; Roediger & McDermott, 1995), and the intrusion rate sometimes approaches the true-recall rate (McDermott, 1996).

DRM lists vary in their ability to generate false recall and/or recognition, and according to FTT, this occurs because the strengths of gist memories vary. The theory assumes that lists that produce very strong gist memories can cause semantic content to take on phenomenological vividness and intensity when such memories are retrieved. Thus, according to FTT, phantom recollection may occur for recall without the presence of physical stimuli, given the right conditions. These illusory vivid experiences should be more common with lists that produce higher levels of false recall than with lists that produce lower levels. In contrast, owing to verbatim–gist dissociation, manipulating the strength of DRM lists is not expected to affect the direct access and true recollective phenomenology for presented items.

Measuring dual-retrieval processes in recall

How are dual-retrieval processes measured? Researchers have traditionally relied on item recognition, whereby participants make old/new decisions about studied items and are then asked to make metacognitive judgments, which are believed to indicate the form of retrieval responsible for recall. The most common measurement techniques are remember/know judgments (Tulving, 1985); inclusion/exclusion judgments, from which the recollective and nonrecollective parameters

of the process dissociation model are estimated (Jacoby, 1991); and confidence judgments, from which recollective and nonrecollective components of the receiver operating characteristic (ROC) are derived (Yonelinas, 1994). However, as has been argued for some time (see Brainerd & Reyna, 2010; Dunn, 2008), subjective measures of phenomenological experience are beset with many interpretive problems. For instance, such measures are based on the assumption that participants understand how to introspect on the phenomenological qualities of their memories, as well as how to perform those introspections reliably. Challenges to metacognitive judgments simply do not arise with the recall approach to measuring dual processes, because only recall data themselves are used to effect the measurements. Specifically, well-developed mathematical models, in the form of two-stage Markov chains, exist for separating and quantifying the contributions of recollective and nonrecollective retrieval to such data (for reviews, see Brainerd, Howe, & Desrochers, 1982; Greeno, 1974). Those models provide purer measures of recollective and nonrecollective retrieval (see Brainerd & Reyna, 2010, for details). Also, such models deliver good fits to conventional recall data (for a review, see Brainerd et al., 2009), without the necessity of enriching the data with metacognitive judgments. We now turn to an overview of the dual-retrieval conception of recall and to two dual-recall models: repeated recall and conjoint recall.

According to dual-retrieval theory (see Brainerd et al., 2009), the probability of successful recall before the first study cycle is basically zero because, on a recall test, participants must generate items on their own, rather than simply recognizing them. Hence, the items in a recall experiment always begin in a no-recall state U, in which the correct response probability is zero. Learning to recall consists of the following discrete changes over trials: Items could pass into a partial-recall state P from U, where the average probability of successful recall is $0 < p < 1$; or they could pass into a perfect-recall state L from U, where the probability of successful recall is 1; or they could pass from U to P and then from P to L on a later trial.

One type of learning (entry into state P) produces non-recollective retrieval, and another (entry into state L) produces recollective retrieval. According to the theory, participants store two types of episodic traces—traces of individual list items (verbatim memories) and traces of some of their features, typically semantic features (gist memories). Recollective retrieval involves direct access to verbatim traces of prior presentations of study-list items. That is, it retrieves an item's trace without first activating, searching through, and comparing the traces of other items (see Kintsch, 1970, and Clark & Gronlund, 1996, for similar conceptualizations), and therefore is the faster of the two operations. These verbatim traces reinstate items' surface forms, which allow rememberers to read items out from consciousness. Direct access therefore

supports the errorless performance of state *L*. Traditional conceptualizations of recollection have at their core this kind of vivid, realistic phenomenology, and thus, direct access is the recollective form of recall.

Despite being fast and highly accurate, direct access rapidly degrades, because verbatim traces are sensitive to proactive and concurrent interference (Barnhardt et al., 2006; Brainerd & Reyna, 1993; Payne et al., 1996). However, if sufficient learning takes place, verbatim traces can be stable enough to survive well past initial storage. The nonrecollective operation *reconstruction* decays much less rapidly because it regenerates items from stable episodic traces of partial identifying information, particularly traces of semantic information (e.g., "soft drink" and "cola" for *Coke*). This nonrecollective form of retrieval consists of a reconstruction operation along with a slave familiarity judgment operation. Reconstructive retrieval is a delimited search process (see Crowder, 1976) that uses episodic traces of some of the targets' features (e.g., "farm animal") to constrain the generation of candidate sets to ones that are restricted enough to be rapidly searched (e.g., *cow*, *sheep*, *horse*, or *pig*). Hence, reconstruction uses the second type of episodic trace to *regenerate* list items. Such traces are not item specific (e.g., "farm animal" is not a specific animal, "European country" is not a specific nation, and "vehicle" is not a specific form of transportation), which means that a process is needed to get from such traces to items that could potentially be output. Brainerd et al. (2009) proposed that this can be accomplished by constructing sets of candidate items that are small enough to be searched within the time limitations of a recall test. To decide which members of a set to output, another operation is required and this is where familiarity comes in. Constructed items generate familiarity signals, and those that generate strong familiarity signals have a greater likelihood of being output as compared to those that do not. Because familiarity judgment may fail to output constructions that are actually list items and may output constructions that are not list items (intrusions), the nonrecollective form of retrieval supports the imperfect recall of state *P*. By fitting a three-state absorbing Markov chain to the error–success data of recall experiments, we can measure direct access, reconstruction, and familiarity judgment.

In terms of repeated recall, Brainerd et al. (2009) showed that the outcome space of experiments in which participants participate in only three test trials per list is adequate to define a two-stage Markov model that contains two direct-access parameters, two reconstruction parameters, and two familiarity judgment parameters. This model is described in the Appendix (Eqs. A1–A11), and the retrieval processes that it measures are summarized in Table 1. From the table, it can be seen that a set of three parameters (D_1, R_1, and J_1) measure direct access (recollection), reconstruction, and familiarity

Table 1 Retrieval processes measured by the Markov chain with repeated-recall data

Process/Parameter	Definition
Recollective Retrieval (True Recollection)	
D_1	The probability that a verbatim trace of an item's presentation on a study list can be directly accessed on Trial 1
D_2	For any item whose verbatim trace could not be accessed on Trial 1, the probability that such a trace can be directly accessed on Trial 2 or 3
Nonrecollective Retrieval	
Reconstruction:	
R_1	For any item whose verbatim trace cannot be accessed on Trial 1, the probability that it can be reconstructed on Trial 1
R_2	For any item for whose verbatim trace cannot be accessed on Trial 1 and that could not be reconstructed on Trial 1, the probability that it can be reconstructed on Trial 2 or 3
Familiarity judgment:	
J_1	For any item that is reconstructed on Trial 1, the probability that the reconstruction is judged to be familiar enough to output
J_2	For any item that is reconstructed on Trial 2 or 3, the probability that the reconstruction is judged to be familiar enough to output
Recollective Retrieval (Phantom Recollection)	
D_{CN1}	The probability that a critical nonpresented (CN) item can be directly accessed on Trial 1
D_{CN2}	For any critical nonpresented item that could not be accessed on Trial 1, the probability that it can be accessed on Trial 2 or 3
Nonrecollective Retrieval	
Reconstruction:	
R_{CN1}	For any critical nonpresented item that cannot be accessed on Trial 1, the probability that it can be reconstructed on Trial 1
R_{CN2}	For any critical nonpresented item that cannot be accessed on Trial 1 and that could not be reconstructed on Trial 1, the probability that it can be reconstructed on Trial 2 or 3
Familiarity judgment:	
J_{CN1}	For any critical nonpresented item that is reconstructed on Trial 1, the probability that the reconstruction is judged to be familiar enough to output
J_{CN2}	For any critical nonpresented item that is reconstructed on Trial 2 or 3, the probability that the reconstruction is judged to be familiar enough to output

judgment on Trial 1, and a second set (D_2, R_2, and J_2) measures these processes on Trials 2 and 3. For example, D_1 is the proportion of items in state U that participants learned to directly access on Trial 1, while D_2 is the proportion of items

in state U or state P that participants learned to directly access on Trials 2 and 3. A parallel set of parameters define the recall of nonpresented critical items, with a set of three parameters (D_{CN1}, R_{CN1}, and J_{CN1}) measuring phantom recollection, reconstruction, and familiarity judgment on Trial 1, and a second set (D_{CN2}, R_{CN2}, and J_{CN2}) measuring these processes on Trials 2 and 3. Note that verbatim traces are not available for critical nonpresented items, implying that phantom recollection must be due to constructive processing. Thus, for targets, the D parameters are measures of verbatim retrieval (with perhaps some slight contribution from reconstructive retrieval), whereas the R parameters measure reconstructive retrieval. For intrusions, however, the D_{CN} and R_{CN} parameters (the reconstruction parameter for intrusion data) are both measures of reconstructive retrieval, the difference between them being whether intrusions are subjected to a further metacognitive check. Intrusions that are accompanied by phantom recollection of their "presentation" (with probability D_{CN}) are simply read out of consciousness, but intrusions that are not accompanied by this illusory phenomenology (with probability R_{CN}) must pass the metacognitive check.

Rather than recall items across three trials, conjoint-recall experiments involve a manipulation whereby participants recall items under one of three types of instructions: verbatim (V), meaning (M), and verbatim + meaning (VM). In the V condition, the emphasis is on recalling only studied items (i.e., directly accessing verbatim traces), whereas M instructions place a premium on reconstructing meaning-preserving items for which verbatim traces are not available. VM instructions allow recall to be based on either direct access of verbatim traces or reconstruction from gist. Brainerd et al. (2009) showed that the outcome space of experiments in which participants participate in one of these three recall conditions can be represented by a two-stage Markov model that contains a direct-access parameter D, a reconstruction parameter R, and a familiarity judgment parameter J. Again, a parallel set of parameters define the recall of nonpresented critical items, with three parameters (D_{CN}, R_{CN}, and J_{CN}) that measure phantom recollection, reconstruction, and familiarity judgment. This model is described in the Appendix (Eqs. A16–A21), and the retrieval processes that it measures are summarized in Table 2.

As can be seen in the Appendix, these models obtain estimates of all of the parameters and conduct fit tests by simultaneously solving Eqs. A2–A9 for the repeated-recall data and Eqs. A16–A21 for the conjoint-recall data. There is empirical support for parameter validity, in that manipulations that embody the process definition of direct access selectively affect parameters that measure the difficulty of transitioning to L from U or P, and manipulations that embody the process definition of reconstruction selectively affect parameters that measure the difficulty of transitioning to P from U (see Brainerd et al., 2009, for a review).

Table 2 Retrieval processes measured by the Markov chain with conjoint-recall data

Process/Parameter	Definition
Recollective Retrieval	
D	The probability that a verbatim trace of an item's presentation on a study list can be directly accessed (explicitly recollected)
Nonrecollective Retrieval	
Reconstruction:	
R	For any item for whose verbatim trace cannot be accessed, the probability that it can be reconstructed
Familiarity judgment:	
J	For any item that is reconstructed, the probability that the reconstruction is judged to be familiar enough to output

In the present research, we used dual-recall theory to investigate whether phantom recollection occurs at high levels with DRM false recall, as it does with false recognition, and whether true recollection and phantom recollection can be dissociated by experimental manipulations, as they have been with recognition. We conducted two experiments. The first implemented the dual-recall model in a repeated recall design, and the second implemented a recall variant of the conjoint recognition paradigm (Brainerd, Reyna, & Mojardin, 1999). DRM list strength was expected, on theoretical grounds, to both elevate phantom recollection and dissociate it from true recollection. We also examined how reconstruction and familiarity judgment were influenced by this manipulation. Although we expected to find convergent results between the first and second experiments with respect to the pattern of parameter effects for the list strength manipulation, the parameter values were not expected to be the same, given the differences in methodology (e.g., repeated vs. single-trial recall).

Experiment 1

Using a repeated-recall design in which the strength of DRM lists was manipulated, we estimated the parameters D, R, and J to measure true recollection, reconstruction, and familiarity judgment for correct recall, and D_{CN}, R_{CN}, and J_{CN} to measure phantom recollection, reconstruction, and familiarity judgment for intrusions of semantic associates. DRM lists of higher strength were expected to increase phantom recollection (the D_{CN} parameters) more so than true recollection (the D parameters). Given that the DRM intrusion rates have been found to drift upward across a series of tests (Payne et al., 1996)—which may be due to a change in how compelling the

accompanying phenomenology is—we also examined whether there was a shift toward or away from phantom recollection over a series of separate recall tests.

Method

Participants A group of 38 undergraduates ($M = 21.29$, $SD = 4.80$) participated in Experiment 1.

Materials Twelve 10-word DRM lists (Roediger, Watson, McDermott, & Gallo, 2001) were used to elicit, on the basis of norms for the 15 associates, low ($X = .24$), intermediate ($X = .36$), or high ($X = .62$) levels of false recall. The four high lists contained the first 10 associates for the critical items *sleep*, *smell*, *doctor*, and *window*; the four intermediate lists contained the first 10 associates for the critical items *pen*, *spider*, *foot*, and *car*; and the four low lists contained the first 10 associates for the critical items *army*, *lion*, *man*, and *thief*.

Design and method The study featured a within-subjects design, with participants receiving three levels of list strength (high vs. intermediate vs. low) and three recall tests. Participants were told that they would be listening to several short lists of words and would be given three memory tests after a number of lists were presented, and that this procedure would be repeated twice. Each participant studied a total of twelve 10-word DRM lists, presented in counterbalanced blocks of four lists, with each block of lists eliciting (on the basis of norms) either high, intermediate, or low false recall. Within lists, words were presented in the "forward" order (strongest to weakest associates of the critical distractor) specified in Roediger and McDermott (1995). A 15-s interval separated consecutive lists, and the items on each list were presented at a 2.5-s rate. After each block of four lists was presented, participants solved arithmetic problems for 1 min, completed a written recall test for 3 min, and finally inserted the recall form into the envelope provided. This recall procedure was immediately repeated two more times for the remaining two blocks of four lists.

Results

The proportions of presented items and of critical nonpresented items that were recalled across the three test trials are reported in Table 3. Although the recall of presented items did not increase across trials, recall of critical nonpresented items did, but this varied depending on the strength of the list. A 3 (strength: high, intermediate, or low) × 3 (test: Trial 1, 2, or 3) repeated measures analysis of variance (ANOVA) for presented items revealed no significant effects. However, a similar ANOVA with critical nonpresented items revealed significant effects of strength, $F(2, 74) = 11.45$, $p < .001$, $\eta^2 =$

Table 3 Mean recall proportions (*M*) and standard deviations (*SD*) for presented and critical nonpresented words for Experiment 1

Item and Illusion Strength	Test 1		Test 2		Test 3		Means	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Presented:								
High	.45	.17	.45	.16	.46	.16	.45	.16
Intermediate	.44	.15	.44	.16	.45	.17	.44	.16
Low	.45	.14	.45	.13	.45	.13	.45	.13
Means	.45	.15	.45	.15	.45	.15	.45	.15
Critical Nonpresented:								
High	.33	.25	.39	.26	.40	.26	.37	.26
Intermediate	.24	.25	.26	.28	.29	.30	.26	.28
Low	.11	.16	.16	.20	.18	.22	.15	.19
Means	.23	.22	.27	.25	.29	.26	.26	.24

.24, and trial, $F(2, 74) = 7.59$, $p < .001$, $\eta^2 = .17$. The reporting of nonpresented critical items increased with list strength, with a significant difference between the high-strength lists and both the intermediate-, $t(37) = 2.65$, $p < .01$, and low-, $t(37) = 4.57$, $p < .001$, strength lists. The reporting of nonpresented critical items also increased from the first to the third trial, $t(37) = 3.50$, $p < .001$.

The recall model described above was used to measure true recollection, reconstruction, and familiarity judgment for correct recall, as well as phantom recollection, reconstruction, and familiarity judgment for semantic intrusions. We estimated the parameters of the model separately for the three list-strength conditions and evaluated goodness of fit by implementing the model in a likelihood function and then maximizing the function for the sample data using numerical methods (for the statistical details, see Riefer & Batchelder, 1988). (Refer to Table 4 for the estimates of the parameters for each condition, as well as for their standard deviations.) The goodness-of-fit-statistic evaluates the evidence for the null hupothesis that the data were generated by the dual-retrieval model against that for the alternative hypothesis that the data were generated by some other process that was not isomorphic to the model. The total value of the fit statistics across the six conditions did not exceed the critical value to reject the null hypothesis (3.84).

Inspection of the parameter values suggests that the list-strength manipulation had effects on the direct-access parameters for false but not for true memory. The estimates of D_1 for the presented items did not vary as a function of list strength, whereas the estimates of D_{CN1} for the critical nonpresented items increased with list strength. Direct access for presented items was greater than phantom recollection, regardless of the illusion strength of the lists.[1] As for the rates of true and

---

[1] All parameter value differences were significant beyond the .05 level, by the appropriate likelihood ratio test.

Table 4 Parameter estimates (P) and standard deviations (SD) for Experiment 1

| Item/Statistic | Illusion Strength | | | | | |
| | High | | Intermediate | | Low | |
	P	SD	P	SD	P	SD
Presented:						
D_1	.41	.01	.41	.02	.41	.02
D_2	.04	.01	.03	.00	.03	.01
R_1	.16	.02	.13	.02	.16	.02
R_2	.00	.02	.00	.00	.00	.02
J_1	.71	.06	.96	.02	.83	.05
J_2	.38	.05	.50	.06	.49	.06
Critical Nonpresented:						
D_{CN1}	.24	.05	.17	.03	.09	.04
D_{CN2}	.01	.05	.04	.02	.00	.04
R_{CN1}	.38	.09	.27	.09	.08	.06
R_{CN2}	.00	.08	.00	.00	.05	.05
J_{CN1}	.42	.12	.34	.15	.67	.56
J_{CN2}	.61	.06	.27	.14	.60	.29

phantom recollection across trials, D_2 and D_{CN2} measure additional increases in true and phantom recollection that take place on Trial 2. Using the expression $D_1 + (1 − D_1)D_2$ to calculate the degree of true recollection on the second trial, and using the corresponding expression for phantom recollection, only small increases were found in both true (from an average of .41 to .43) and phantom (from an average of .17 to .18) recollection, indicating little additional direct access of items beyond the first trial.

With respect to the reconstuction and familiarity judgment parameters, list strength did not affect reconstructive processing for the presented items, as reflected in the similar R_1 values for the different list strengths. However, reconstructive processing was affected by list strength for the critical nonpresented items. The R_{CN1} values were larger, the higher the list strength, with the values for the intermediate- and high-strength lists exceeding the reconstruction estimates for presented items. In the rates of reconstruction for the presented and the critical nonpresented items across trials, as we saw with direct access, there was no reliable increase in reconstruction from the first to the second trial (from an average of .15 to .15 and from .24 to .25 for the presented and critical nonpresented items, respectively). The judgment parameter was consistently higher for the presented (J_1) than for the critical nonpresented (J_{CN1}) items, regardless of list strength, and it increased from Trials 1 to 2.

With models like the dual-recall model, between-condition differences in parameter values are evaluated for

statistical significance with a three-step procedure (see, e.g., Brainerd et al., 2009). The first step is analogous to the omnibus F test in ANOVA: For all of the conditions of an experiment, an experiment-wise likelihood ratio test, which is a G^2 statistic that is asymptotically distributed as χ^2, is computed to evaluate the null hypothesis that parameter values do not vary between conditions. This null hypothesis was rejected for the critical nonpresented item conditions in the present experiment, $G^2(10) = 16.96$. Condition-wise tests were not necessary for the presented items, as parameter values did not vary. The second step is analogous to paired comparisons in ANOVA. For each pair of conditions, a likelihood ratio test is computed to evaluate the null hypothesis that parameter values do not differ between that specific pair of conditions. This null hypothesis was rejected for the critical nonpresented items in the high- versus the intermediate-list-strength conditions, $G^2(8) = 41.52$, in the intermediate- versus the low-list-strength conditions, $G^2(8) = 59.73$, and in the high- versus the low-list-strength conditions, $G^2(8) = 56.00$. No condition-wise tests were performed for the presented items, as there were no consistent differences in the parameters of interest. This null hypothesis was also rejected for the presented versus the critical nonpresented items in the high-list-strength condition, $G^2(8) = 22.18$, the intermediate-list-strength condition, $G^2(8) = 25.82$, and the low-list-strength condition, $G^2(8) = 42.59$.

The third step, which has no analogue in ANOVA, determines which parameters were affected by these manipulations. For each pair of conditions, a likelihood ratio test is computed for each parameter in order to evaluate the null hypothesis that the parameter has the same value in that pair of conditions. This final series of tests produced the following results of interest: D_1 and R_1 were larger with the high- than with the intermediate-strength lists, with the intermediate- than with the low-strength lists, and with the high- than with the low-strength lists; $D_1 > D_2$ for the presented items for all three lists strengths, and for the critical nonpresented items for the high- and intermediate-strength lists; $R_1 > R_2$ for the presented items for all three lists strengths and for the critical nonpresented items for the high-strength lists; $D_1 > R_1$ for the presented items for all three lists strengths; $J_1 < J_2$ for the presented items for all three lists strengths; $D_1 > D_{CN1}$ for all three list strengths; and $R < R_{CN}$ for all three list strengths.

Discussion

Given that the model provided acceptable fits to the data, we were able to answer the main questions of interest. The first question concerned the extent to which false recall arises from phantom recollection. Although the probability of recollection was much higher for presented items than for nonpresented items, phantom recollection did occur with recall, especially

with lists believed to elicit high levels of illusory vivid phenomenology. The second aim was to determine whether true recollection and phantom recollection can be dissociated with manipulations that involve the strength of DRM lists and repeated testing. Here, as expected, phantom recollection reliably increased from the low- to the high-strength lists. In contrast, list strength did not affect true recollection. Reconstruction rates were much higher for critical nonpresented items than for presented items with the high and intermediate lists, but list strength did not influence reconstruction for presented items. Thus, phantom recollection and the true recollection that accompanies presented items appear to involve different retrieval processes.

The third and final aim of this experiment was to determine whether recall shifts from verbatim-based to gist-based recollection across trials. Consistent with Brainerd et al. (2003), false recall increased over the three test trials; however, unlike Brainerd et al. (2003), who found a slight increase across trials for true recall, we found no change. However, this was also the case for critical nonpresented items. Thus, both true recollection and phantom recollection were unaffected by repeated recall. Repeated recall also had little effect on the rates of reconstruction for both presented and critical nonpresented items. The increase in false recall across trials was due to the greater likelihood that reconstructed critical nonpresented items would be output on subsequent trials. Using the expression $J_1 + (1 - J_1)J_2$ to measure the probability that reconstructed presented and critical nonpresented items would be judged familiar enough to be output, we found a smaller increase from Trials 1 to 2 for presented items (from .71 to .82, .96 to .98, and .83 to .91 for the high-, intermediate-, and low-strength lists, respectively) than for critical nonpresented items (from .42 to .77, .34 to .52, and .67 to .87 for the high-, intermediate-, and low-strength lists, respectively). Thus, although recall did not shift toward or away from phantom recollection over trials, reconstruction-based intrusions were more likely to be output on subsequent trials.

Experiment 2

The estimates of phantom recollection in Experiment 1 demonstrated that in the DRM paradigm, false recall generates illusory vivid phenomenology. The two main objectives of this study were to determine whether we would obtain convergent results with respect to the existence of phantom recollection in recall and whether list strength would dissociate true- from phantom-recollection parameters using a different recall methodology: conjoint recall.

The conjoint-recall methodology is expected to produce higher levels of phantom recollection than were produced in the first experiment, because repeated recall should strengthen verbatim memories, increasing direct access to studied items and the rejection of false items, as compared to those verbatim memories formed with the conjoint-recall methodology. The third purpose of the study was to secure further data on the relationship between the reliance on verbatim and gist traces and recollective phenomenology. Recall that according to FTT, relying on strong gist traces can produce high levels of illusory vivid phenomenology, and because the conjoint-recall instructions manipulate reliance on gist traces, phantom recollection should vary in predictable ways across the three recall conditions. In the V condition, the emphasis is on recalling only studied items (i.e., directly accessing verbatim traces), which means that fewer critical nonpresented items should be recalled, as compared to in the M and VM conditions, where both recollective and nonrecollective forms of recall are acceptable. However, this may not be the case with lists of higher strength, where reliance on nonrecollective recall in the V condition may be greater. Given the emphasis on verbatim recall, critical nonpresented items that are recalled should be accompanied by the realistic conscious experience of their prior "presentation." M instructions, however, place a premium on reconstructing meaning-preserving items for which verbatim traces are not available, which should increase the number of critical nonpresented items recalled, as compared to the V condition. Because recall will be based on reconstruction from the start, we would expect critical nonpresented items to be reported earlier in the M condition than in the V condition. Finally, VM instructions allow recall to be based on either direct access to verbatim traces or reconstruction from gist, so critical nonpresented items can be accompanied by either recollective or nonrecollective phenomenology. The average output position of critical nonpresented items should be intermediate between the V and M conditions.

Method

Participants A group of 120 undergraduates ($M = 20.34$, $SD = 5.52$) participated in Experiment 2.

Materials, design, and method This was a 3 (list strength: high, intermediate, or low) × 3 (instruction: V, M, or VM) mixed-model design, with list strength manipulated within subjects and instruction manipulated between subjects. The materials and methodology were similar to those of Experiment 1, with the exception of the instructions that participants received for recall and the fact that items were

recalled only once. In the V condition ($N = 40$), participants were asked to report only words that were actually presented. In the M condition ($N = 39$), they were asked to report only nonpresented words whose meanings were similar to the presented words. Finally, participants in the VM condition ($N = 41$) were asked to report presented words as well as nonpresented similar words. Following presentation of each of the three blocks of lists, participants solved mathematics problems for 1 min before commencing written recall.

Results

The proportions of presented and critical nonpresented items recalled are reported by condition in Table 5. Overall, with the V instructions, more presented items were recalled with lists of low strength, and more critical nonpresented items were recalled with lists of high strength. Critical nonpresented items were output significantly later in the V condition than in either the VM or the M condition, and they were recalled earlier with lists of high strength in the V and VM conditions.

A 3 (strength: high, intermediate, or low) × 3 (instruction: V, M, or VM) ANOVA of the total recall data for presented items revealed effects of strength, $F(2, 234) = 8.36$, $p < .001$, $\eta^2 = .07$, and instruction, $F(2, 117) = 170.48$, $p < .001$, $\eta^2 = .75$, as well as a Strength × Instruction interaction, $F(4, 234) = 3.50$, $p < .01$, $\eta^2 = .06$. Post-hoc analyses of the strength main effect revealed that slightly more presented items were recalled with the low- than with the intermediate-

Table 5 Mean recall proportions (M) and standard deviations (SD) for presented and critical nonpresented words for Experiment 2

Item and Illusion Strength	Instructional Condition							
	V		M		VM		Means	
	M	SD	M	SD	M	SD	M	SD
Presented:								
High	.38	.14	.08	.10	.42	.12	.29	.12
Intermediate	.37	.14	.05	.05	.37	.13	.26	.11
Low	.44	.12	.06	.07	.43	.15	.31	.11
Means	.40	.13	.06	.07	.41	.13	.29	.11
Critical Nonpresented:								
High	.42	.31	.26	.26	.68	.25	.45	.27
Intermediate	.31	.29	.31	.30	.59	.27	.40	.29
Low	.14	.19	.26	.26	.41	.25	.27	.23
Means	.29	.26	.28	.27	.56	.26	.38	.26

strength lists, $t(119) = 4.46$, $p < .001$. As for the instruction main effect, more items were recalled in the V and VM conditions than in the M condition, $p < .05$. The Strength × Instruction interaction indicated that recall varied as a function of strength under V instructions, but not under M or VM instructions. In the V condition, more presented items were recalled in the low-strength condition than in the intermediate-, $t(39) = 4.54$, $p < .001$, or the high-, $t(39) = 3.31$, $p < .005$, strength condition.

A similar ANOVA was conducted with the total recall data for critical nonpresented items. There were effects of strength, $F(2, 234) = 19.02$, $p < .001$, $\eta^2 = .14$, and instruction, $F(2, 117) = 30.51$, $p < .001$, $\eta^2 = .34$, as well as a Strength × Instruction interaction, $F(4, 234) = 4.45$, $p < .005$, $\eta^2 = .07$. Fewer critical nonpresented items were recalled with the low lists than with the intermediate, $t(119) = 4.48$, $p < .001$, or the high, $t(119) = 5.46$, $p < .001$, lists. Recall was greater in the VM condition than in the M or the V condition, $p < .05$. The Strength × Instruction interaction indicated that recall differed as a function of strength in the V and VM conditions, but not in the M condition. In the V condition, fewer critical nonpresented items were recalled in the low-strength than in the intermediate-, $t(39) = 3.66$, $p < .001$, or the high-, $t(39) = 5.54$, $p < .001$, strength condition. Similarly, in the VM condition, fewer critical nonpresented items were recalled in the low-strength condition than in the intermediate-, $t(40) = 3.62$, $p < .001$, or the high-, $t(40) = 4.88$, $p < .001$, strength condition.

In terms of the positions at which critical nonpresented items were recalled, a 3 (strength: high, intermediate, or low) × 3 (instruction: V, M, or VM) ANOVA was carried out on the proportional output position data (output position relative to the positions of all other items). Both the strength, $F(2, 234) = 31.63$, $p < .001$, $\eta^2 = .21$, and instruction, $F(2, 117) = 199.75$, $p < .001$, $\eta^2 = .77$, main effects were significant, as was the interaction, $F(4, 234) = 22.00$, $p < .001$, $\eta^2 = .27$. The average output position for the intermediate lists ($M = .41$, $SD = .09$) was significantly earlier than for either the high- ($M = .44$, $SD = .11$), $t(119) = 3.33$, $p < .001$, or the low- ($M = .48$, $SD = .14$), $t(119) = 6.17$, $p < .001$, strength lists, and the output position was significantly earlier for the high lists than for the low lists, $t(119) = 3.98$, $p < .001$. Items were output significantly later in the V condition ($M = .56$, $SD = .08$) than in either the VM ($M = .41$, $SD = .08$) or the M ($M = .37$, $SD = .06$) condition, and significantly later in the VM than in the M condition (all $ps < .05$).

The interaction revealed that with the lists of low strength, items were recalled later in the V condition ($M = .65$, $SD = .05$) than in either the VM ($M = .44$, $SD = .08$) or the M ($M = .35$, $SD = .05$) condition, and later in the VM condition than in the M condition. For the intermediate lists,

items were recalled later in the V condition ($M = .46$, $SD = .09$) than in either the VM ($M = .39$, $SD = .09$) or the M ($M = .36$, $SD = .05$) condition, but the VM and M conditions did not differ. This latter pattern was also observed for the high lists, in which critical nonpresented recall was output significantly later in the V condition ($M = .55$, $SD = .10$) than in either the VM ($M = .39$, $SD = .08$) or the M ($M = .38$, $SD = .07$) condition (all $ps < .05$). The interaction also revealed that with the V instructions, the average output position for the intermediate lists ($M = .46$, $SD = .09$) was significantly earlier than those for both the high ($M = .55$, $SD = .10$), $t(39) = 4.43$, $p < .001$, and the low ($M = .65$, $SD = .05$) lists, $t(39) = 11.63$, $p < .001$, and the output position was significantly earlier for the high lists than for the low lists, $t(39) = 6.26$, $p < .001$. With the M instructions, the mean output position for critical nonpresented items did not vary significantly across the high ($M = .38$, $SD = .07$), intermediate ($M = .36$, $SD = .05$), or low ($M = .35$, $SD = .05$) lists (all $ts < 2.16$, all $ps > .04$). With the VM condition, the output position was significantly earlier for the high lists ($M = .39$, $SD = .08$) than for the low lists ($M = .44$, $SD = .08$), $t(40) = 3.07$, $p < .005$.

Turning to the rates of true and phantom recollection, we used Eqs. A16–A21 to measure the probabilities that presented and critical nonpresented items were directly accessed (D and D_{CN} for true and phantom recollection, respectively), the probabilities that presented and critical nonpresented items were reconstructed (R and R_{CN}, respectively), and the probabilities that the reconstructions of presented and critical nonpresented items were familiar enough to output the reconstructed item (J and J_{CN}, respectively). Because we had the same number of parameters as there were degrees of freedom (i.e., three), there were no degrees of freedom remaining for model fitting. However, Brainerd et al. (2003) showed that the conjoint-recall model

Table 6 Parameter estimates for Experiment 2

Item/Statistic	Illusion Strength		
	High	Intermediate	Low
Presented:			
D	.34	.32	.37
R	.12	.07	.10
J	1.00	1.00	1.00
Critical Nonpresented:			
D_{CN}	.42	.28	.15
R_{CN}	.45	.43	.31
J_{CN}	.00	.10	.00

fits data produced in experiments with similar conjoint-recall methodology.

Inspection of the parameter values in Table 6 suggests that the list-strength manipulation had clear effects on phenomenology, replicating Experiment 1. The estimates of D for the presented items indicated that true recollection was not affected by list strength. However, the estimates of D_{CN} for the critical nonpresented items increased monotonically with list strength. As in Experiment 1, illusory vivid phenomenology increased as list strength increased.

Turning to reconstuction and metacogntive judgment, list strength did not affect either process for presented items. However, the values of the reconstruction parameter, R_{CN}, were larger with the high and intermediate lists than with the low lists, with the values for all list strengths greatly exceeding the reconstruction estimates for presented items. The metacognitive judgment parameter for intrusions was not affected by list strength.

The proportions of true and false recall that provoked recollective experience could be computed from the ratios $D/[D + (1 - D)RJ]$ and $D_{CN}/[D_{CN} + (1 - D_{CN})R_{CN}J_{CN}]$, respectively. Both proportions were quite high and varied little across lists of different strengths (for the low, intermediate, and high lists, respectively, the values were .85, .87, and .81 for true recollection, and 1.00, .90, and 1.00 for phantom recollection).

Discussion

The purpose of this second experiment was to use conjoint-recall instructional conditions to study phantom recollection in DRM recall. As expected, this manipulation influenced the frequency of false recall and its output position. We found that recall of critical nonpresented items was more common under VM instructions than under V instructions with low- and intermediate-strength lists, and was more common under V than under M instructions with low-strength lists. The VM instructions produced the greatest recall of critical nonpresented items, regardless of list strength. Because VM instructions allowed recall to be based on either direct access to verbatim traces or reconstruction from gist, critical nonpresented items could be accompanied by either recollective or nonrecollective phenomenology, which increased the likelihood that they would be reported. We expected recall of critical nonpresented items to be greater in the M than in the V condition, because these instructions emphasized reconstructing meaning-preserving items for which verbatim traces were unavailable; however, this was only the case with low-strength lists. It appears, then, that as list strength increases and gist memory strengthens, recall becomes dominated by reconstruction from

meaning, under all three types of instructions. Greater reconstruction from meaning in the intermediate- and high-strength lists may make the critical nonpresented items more similar to one another in the V and M conditions than they are with low-strength lists.

Concerning output position, it was expected that critical nonpresented items would be recalled later in the V condition than in either the M or the VM condition. The data fell out as expected. According to dual-recall theory (Brainerd et al., 2002), early output of standard lists is dominated by direct access to verbatim traces and recollective phenomenology, while later output is dominated by reconstruction from gist and nonrecollective phenomenology. DRM recall is different, according to the results of this experiment. Intrusions of critical nonpresented items in the V condition typically occurred midway through the output protocol, not near the end, and were accompanied by recollective phenomenology (phantom recollection).

True recall and false recall were dissociated by the list-strength and instruction manipulations. The list-strength manipulation drove true and false recall in opposite directions in the V and VM conditions. Similarly, list strength influenced phantom recollection, but not true recollection or reconstruction for nonpresented items, and had no effect on presented items. Finally, as predicted, false recall was accompanied by high rates of phantom recollection, consistent with the results of Experiment 1, with recollective phenomenology dominating both true and false recall.

General discussion

The primary purpose of this study was to use dual-retrieval theory, and its associated models, to investigate the degree to which phantom recollection occurs in recall, as well as whether phantom recollection and true recollective phenomenology can be dissociated by a gist manipulation, DRM list strength. Both experiments showed that phantom recollection accompanies false recall, especially with lists that encourage reliance on gist memories (i.e., high-strength lists), and that true recollection and phantom recollection are dissociated processes. Given the differences in methodology across the two experiments (e.g., single study–test vs. repeated recall), the levels of false recall, and consequently phantom recollection, were expected to be higher with the conjoint methodology. Indeed, whereas the level of phantom recollection was nearly half the true recollection level with the repeated recall methodology, it exceeded it with the conjoint-recall methodology. False memory items, therefore, do not need to be physically presented on memory

tests for phantom recollection to be triggered. Under certain conditions (e.g., when target experiences repeatedly cue the same meaning, as they do in DRM lists), gist-based reconstruction can support illusory vivid phenomenology that emulates verbatim-based recollective phenomenology. This has been shown to be the case with recognition (e.g., Roediger & McDermott, 1995; Stahl & Klauer, 2009; Singer & Remillar, 2008), and from the results of the present study, it is also the case with false recall. Although many nonpresented critical items were accompanied by vivid illusory experience, reconstruction-based false recall was more common than phantom recollection. Unlike in recognition, then (Brainerd et al., 2001), phantom recollection was not the main contributor to DRM false recall.

Our findings favor a gist-processing basis for phantom recollection. First, although phantom recollection did not increase over recall tests, repeated recall increased the likelihood that reconstructed critical nonpresented items would be output on subsequent trials. Second, there was evidence of higher false recall in the VM and M conditions than in the V condition with the low-strength lists, and of higher false recall in the VM than in the V condition with intermediate-strength lists, consistent with the idea that the task would provide more opportunity to report such items when recall could be based on either recollective or nonrecollective phenomenology and when there was no need for a metacognitive check to decide whether reconstructed items should be reported. Third, if constructive processing underlies the recall of critical nonpresented words, their mean output positions should be later than those for targets, which are assumed to be directly accessed. This is, in fact, what has been found (e.g., Brainerd et al., 1993; Payne et al., 1996; Roediger & McDermott, 1995). However, when recall of semantic associates is accompanied by recollective phenomenology, we would expect them to be treated as if they were targets and, consequently, to be output earlier. In this experiment, the mean output positions of critical nonpresented words were altered by instructions that recall should be based only on either constructive processing (M instructions) or on both constructive processing and verbatim retrieval (VM instructions). The fact that critical nonpresented items were output significantly earlier with the high- than with the low-strength lists in both the V and VM conditions, but did not vary across list strengths in the M condition, is consistent with the claim that phantom recollection is more likely to occur with strong DRM lists, and as such will result in earlier output.

True and false recall, as well as true and phantom recollection, were dissociated by the strength of DRM lists. Because high-strength DRM lists are believed to induce stronger gist memories, the presentation of high lists should

increase phantom recollection without affecting true recollection; this was the finding in both studies. As described earlier, according to FTT, whereas true recollection results primarily from direct access to verbatim traces, phantom recollection is due to constructive processing of strong gist traces. Thus, phantom recollection and the true recollection that accompanies presented items appear to involve different retrieval processes. These results are consistent with the claim that both false recall and false recognition involve dual processes (e.g., Brained & Reyna, 1988; McDermott & Watson, 2001; Roediger & McDermott, 2000; Seamon, Luo, Shulman, Toner, & Caglar, 2002).

If it is in fact the case that true and false phenomenology are byproducts of different retrieval processes, it may be possible to use remembering phenomenology to discriminate true from false reports (Brainerd & Reyna, 2005). We know that there are subtle phenomenological differences between true and false memory and that phenomenology may be a useful means of differentiating experienced from nonexperienced events (e.g., Marche, Brainerd, & Reyna, 2010). Dual-retrieval models of recall and recognition are needed to secure quantitative evidence on the phenomenologies that accompany the reporting of false memories, and they provide an attractive technology for measuring the different phenomenological bases of true and false memory.

Appendix

Repeated recall

Consider an experiment in which there is one study trial for each target list, with three recall tests following each study cycle—that is, an experiment of the form $S_1T_1T_2T_3$. Because there are three recall tests, each item on the list can have one of eight different patterns of successes and errors over these tests: $C_1C_2C_3$, $C_1C_2E_3$, $C_1E_2C_3$, ..., $E_1E_2E_3$, where C_i designates that the item was recalled on the ith test, and E_i designates that it was not recalled. Similarly, each critical nonpresented (CN) item from the list can have one of eight different patterns of successes and errors over these tests: $C_{CN1}C_{CN2}C_{CN3}$, $C_{CN1}C_{CN2}E_{CN3}$, $C_{CN1}E_{CN2}C_{CN3}$, ..., $E_{CN1}E_{CN2}E_{CN3}$, where C_{CNi} designates that the critical nonpresented item was recalled on the ith test and E_{CNi} designates that it was not recalled. The model's parameters in Table 1 can be estimated for any set of such data by expressing the observed probability of each of the eight error–success patterns in a two-stage Markov process that contains the parameters. The states of the process are U (an initial no-recall state), P (an intermediate partial-recall state, with a correct recall substate P_C and an incorrect recall substate P_E, in which the probability of successful recall has some average value $0 < p < 1$), and L, a terminal criterion-recall state. The two-stage Markov process for these states consists of a starting vector W_1 and a transition matrix M (a similar vector and matrix can be developed for the critical nonpresented items by substituting the parameters in the following equations with their corresponding CN parameters):

$$W_1 = [L(2), P_C(2), P_E(2), U(2), L(1), P_C(1), P_E(1), U(1)]$$
$$= [0, 0, 0, 0, D_1, (1-D_1)R_1J_1, (1-D_1)R_1(1-J_1),$$
$$(1-D_1)(1-R_1)];$$

		$L(2)$	$P_E(2)$	$P_C(2)$	$U(2)$	$L(3)$	$P_E(3)$	$P_C(3)$	$U(3)$	
	$L(2)$	0	0	0	0	1	0	0	0	
	$P_E(2)$	0	0	0	0	D_2	$(1-D_2)(1-J_2)$	$(1-D_2)J_2$	0	
	$P_C(2)$	0	0	0	0	0	$1-J_2$	J_2	0	
$M =$	$U(2)$	0	0	0	0	D_2	$(1-D_2)(1-J_2)$	$(1-D_2)J_2$	$1-D_2$	(A1)
	$L(1)$	1	0	0	0	0	0	0	0	
	$P_E(1)$	D_2	$(1-D_2)(1-J_2)$	$(1-D_2)J_2$	0	0	0	0	0	
	$P_C(1)$	0	$1-J_2$	J_2	0	0	0	0	0	
	$U(1)$	D_2	$(1-D_2)(1-J_2)$	$(1-D_2)J_2$	$1-D_2$	0	0	0	0	

The probabilities of the eight individual error–success patterns for the presented items are obtained by simply multiplying the vector and the matrix together. Those expressions are

$$P(C_1C_2C_3) = D_1 + (1-D_1)R_1J_1(J_2)^2; \qquad (A2)$$

$$P(C_1C_2E_3) = (1-D_1)R_1J_1J_2(1-J_2); \qquad (A3)$$

$$P(C_1E_2C_3) = (1-D_1)R_1J_1(1-J_2)[D_2 + (1-D_1)J_2]; \qquad (A4)$$

$$P(C_1E_2E_3) = (1-D_1)R_1J_1(1-J_2)^2; \qquad (A5)$$

$$P(E_1C_2C_3) = (1 - D_1)(1 - R_1)D_2 + (1 - D_1)R_1(1 - J_1)$$
$$\times \left[D_2 + (1 - D_2)(J_2)^2 \right]; \qquad (A6)$$

$$P(E_1C_2E_3) = (1 - D_1)(1 - R_1)(1 - D_2)R_2J_1(1 - J_2)$$
$$+ (1 - D_1)R_1(1 - J_1)(1 - D_2)J_2(1 - J_2); \qquad (A7)$$

$$P(E_1E_2C_3) = (1 - D_1)(1 - R_1) \qquad (A8)$$
$$\times \{ D_2 + (1 - D_2)R_2(1 - J_2)[D_2 + (1 - D_2)J_2] \}$$
$$+ (1 - D_1)R_1(1 - J_1)(1 - D_2)$$
$$\times (1 - J_2)[D_2 + (1 - D_2)J_2];$$

$$P(E_1E_2E_3) = (1 - D_1)(1 - R_1)\{ [(1 - D_2)(1 - R_2)]^2$$
$$+ (1 - D_2)(1 - R_2)(1 - D_2)R_2(1 - J_2)$$
$$+ (1 - D_2)R_2(1 - J_2)^2 \} \qquad (A9)$$
$$+ (1 - D_1)R_1(1 - J_1)(1 - J_2)^2$$

A corresponding set of probabilities of the eight error–success patterns for the critical nonpresented items can easily be written by substituting the equations above with those corresponding to recall of critical nonpresented items. The likelihood of any sample of data (and estimates of the parameters in Table 1, for the presented and critical nonpresented

items separately) is then obtained by maximizing the following likelihood function:

$$L_{11} = \text{Pi}(p_i)^{N(i)}. \qquad (A10)$$

The p_i are the eight expressions on the right sides of Eqs. A2–A9, and because the terms in Eq. A10 are multiplied, this expression can only be maximized if all of the $p_i > 0$, which means that all of the exponents must be >0. The exponent of each p_i is an empirical data count that corresponds to one of the eight error–success sequences and is the number of times that error–success sequence was observed in sample data. Because six memory parameters are estimated, the likelihood value in Eq. A10 is computed with 1 degree of freedom. A goodness-of-fit test is then carried out by computing a likelihood ratio statistic that compares the likelihood in Eq. A10 with the likelihood of the same data when all seven observable probabilities are free to vary. This test evaluates the null hypothesis that no more than two retrieval processes are required to account for the data. That test statistic, which is asymptotically distributed as $\chi^2(1)$, is

$$G^2 = -2\ln[L_6/L_7], \qquad (A11)$$

where L_7 is the likelihood of the data when all sev00en observable probabilities are free to vary.

Another goodness-of-fit test can be calculated that assesses the null hypothesis that learning to recall involves only a single nonrecollective process. This one-process model is obtained from the two-process model by simply removing one of the Markov states:

$$W_1' = [L(1), P_E(1), P_C(1), L(2), P_E(2), P_C(2),] = [D'_1,$$
$$(1 - D'_1)(1 - J'_1), (1 - D'_1)J'_1, 0, 0, 0];$$

$$M' = \begin{array}{c} \\ L(2) \\ P_E(2) \\ P_c(2) \\ L(1) \\ P_E(1) \\ P_c(1) \end{array} \begin{array}{c} L(2) \\ 0 \\ 0 \\ 0 \\ 1 \\ D'_2 \\ D'_2 \end{array} \begin{array}{c} P_E(2) \\ 0 \\ 0 \\ 0 \\ 0 \\ (1 - D'_2)(1 - J'_2) \\ (1 - D'_2)(1 - J'_2) \end{array} \begin{array}{c} P_c(2) \\ 0 \\ 0 \\ 0 \\ 0 \\ (1 - D'_2)J'_2 \\ (1 - D'_2)J'_2 \end{array} \begin{array}{c} L(3) \\ 1 \\ D'_2 \\ D'_2 \\ 0 \\ 0 \\ 0 \end{array} \begin{array}{c} P_E(3) \\ 0 \\ (1 - D'_2)(1 - J'_2) \\ (1 - D'_2)(1 - J'_2) \\ 0 \\ 0 \\ 0 \end{array} \begin{array}{c} P_c(3) \\ 0 \\ (1 - D'_2)J'_2 \\ (1 - D'_2)J'_2 \\ 0 \\ 0 \\ 0 \end{array} \qquad (A12)$$

The likelihood of any set of data over which Eq. A11 can be defined can also be estimated for the one-process model in A12 by maximizing a simplified version of Eq. A11 that contains only the parameters in Eq. A12. The revised fit statistic is then

$$G^2 = -2\ln[L_4/L_7], \qquad (A13)$$

which is asymptotically distributed as $\chi^2(3)$ because the one-process model contains only four memory parameters.

A third goodness-of-fit test of a one-process model can also be computed, which assumes that the process is *recollective* rather than nonrecollective retrieval. The recollective one-process model is also obtained from the two-process model by simply removing one of the Markov states, as follows:

$$W_1'' = [L(1), U(1), L(2), U(2)] = \left[D'_1, \left(1 - -D'_1 \right), 0, 0 \right];$$

$$M'' = \begin{array}{c} \\ L(2) \\ U(2) \\ L(1) \\ U(1) \end{array} \begin{array}{cccc} L(2) & U(2) & L(3) & U(3) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & D_2' & 1 - D_2' \\ 1 & 0 & 0 & 0 \\ D_2' & 1 - D_2' & 0 & 0 \end{array} \qquad (A14)$$

The likelihood of any set of data over which Eq. A11 can be defined can also be estimated for this one-process model by maximizing a simplified version of Eq. A11 that contains only the parameters in Eq. A12. The revised fit statistic is then

$$G^2 = -2\ln[L_2/L_7], \qquad (A15)$$

which is asymptotically distributed as $\chi^2(5)$ because this one-process model contains only two memory parameters.

Note that if the fit statistic in Eq. A13 produces a null hypothesis rejection (showing that recall involves more than a single nonrecollective process), it is unnecessary to then compute the fit test in Eq. A15 (to determine whether the recall involves a single recollective process) because the recollective one-process model in Eq. A14 is a submodel of the nonrecollective one-process model in Eq. A12.

Conjoint recall

Now consider an experiment that involves an instructional manipulation whereby participants recalled the DRM lists under one of three recall conditions: verbatim (V), meaning (M), and verbatim + meaning (VM). In the V condition, the emphasis is on recalling only studied items (i.e., directly accessing verbatim traces). In the M condition, participants are instructed to recall only items that are related to those presented (i.e., reconstructing gist traces). VM instructions allow recall to be based on either direct access to verbatim traces or reconstruction from gist.

Because V is a standard recall condition, the dual-retrieval model's expression for correct recall is

$$PV(C) = D + (1 - D)RJ, \qquad (A16)$$

where PV(C) is the probability of correct recall under V instructions. Under M instructions, a target whose presentation is explicitly recollected should not be recalled. However, if it is reconstructed without being recollected, this is used as evidence that the item was not presented and the item will then be recalled, and no metacognitive check is needed to pass the item on for output. Thus, the dual-retrieval model's expression for the probability that a target is (incorrectly) recalled is

$$PM(E) = (1 - D)R. \qquad (A17)$$

Finally, meaning-relatedness, not presentation, is the criterion used for recall under the VM instructions, which means that participants will output targets when they are either recollected or reconstructed. Because prior presentation is not a recall criterion, there is no need for a metacognitive check under VM instructions. Hence, the dual-retrieval model's expression for the probability that a target is (correctly) recalled is

$$PVM(C) = D + (1 - D)R. \qquad (A18)$$

A parallel series of equations can be written that expresses the probability of incorrect recall of critical distractors (under V instructions) and the probability of correct recall of critical distractors (under M and VM instructions). Those expressions are

$$PV_{CN}(E) = D_{CN} + (1 - D_{CN})R_{CN}J_{CN}, \qquad (A19)$$

$$PM_{CN}(C) = (1 - D_{CN})R_{CN}, \qquad (A20)$$

and

$$PVM_{CN}(C) = D_{CN} + (1 - D_{CN})R_{CN}, \qquad (A21)$$

from which estimates of phantom recollection in free recall can be obtained.

Both the repeated-recall and conjoint-recall models are identifiable (i.e., all parameters can be measured) and are process-specific measures. The identifiability proof for the model in Experiment 1 is provided in Brainerd et al. (2009), and the identifiability proof for the model in Experiment 2 is provided in Brainerd et al. (2003). Evidence for process specificity is also provided in Brainerd et al. (2009), and the issue is considered at length in Brainerd and Reyna (2010).

References

Atkinson, R. C., & Juola, J. F. (1973). Factors influencing speed and accuracy in word recognition. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 583–612). New York, NY: Academic Press.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Learning, memory, and thinking, Vol. 1, pp. 243–293). San Francisco, CA: Freeman.

Atkinson, R. C., & Westcourt, K. T. (1975). Some remarks on a theory of memory. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance V* (pp. 485–498). New York, NY: Academic Press.

Barnhardt, T. M., Choi, H., Gerkens, D. R., & Smith, S. M. (2006). Output position and word relatedness effects in a DRM paradigm: Support for a dual-retrieval process theory of free recall and false memories. *Journal of Memory and Language, 55,* 213–231. doi:10.1016/j.jml.2006.04.003

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

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. (1988). When things that never happened are easier to remember than things that did. *Psychological Science, 9,* 484–489.

Brainerd, C. J., & Reyna, V. F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review, 100,* 42–67. doi:10.1037/0033-295X.100.1.42

Brainerd, C. J., & Reyna, V. F. (1996). Mere memory testing creates false memories in children. *Developmental Psychology, 32,* 467–478. doi:10.1037/0012-1649.32.3.467

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual-processes in memory, reasoning, and cognitive neuroscience. *Advances in Child Development and Behavior, 28,* 41–100. doi:10.1016/S0065-2407(02)80062-3

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory.* New York, NY: Oxford University Press.

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., Harnishfeger, K. K., & Howe, M. L. (1993). Is retrievability grouping good for recall? *Journal of Experimental Psychology. General, 122,* 249–268. doi:10.1037/0096-3445.122.2.249

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. doi:10.1037/a0016963

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106,* 160–179. doi:10.1037/0033-295X.106.1.160

Brainerd, C. J., Wright, R., Reyna, V. F., & Mojardin, A. H. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 307–327. doi:10.1037/0278-7393.27.2.307

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.

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

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale, NJ.: Erlbaum.

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

Dunn, J. C. (2008). The dimensionality of the remember–know task: A state-trace analysis. *Psychological Review, 115,* 426–446. doi:10.1037/0033-295X.115.2.426

Gallo, D. A., McDermott, K. B., Percer, J. M., & Roediger, H. L., III. (2001). Modality effects in false recall and false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 339–353. doi:10.1037/0278-7393.27.2.339

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

Greeno, J. G. (1974). Representation of learning as discrete transition in a finite state space. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Learning, memory, and thinking, Vol. 1, pp. 1–43). San Francisco, CA: Freeman.

Horton, D. L., Pavlick, T. J., & Moulin-Julian, M. W. (1993). Retrieval-based and familiarity-based recognition and the quality of information in episodic memory. *Journal of Memory and Language, 32,* 39–55.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541. doi:10.1016/0749-596X(91)90025-F

Kintsch, W. (1970). *Learning, memory, and conceptual processes.* New York, NY: Wiley.

Lampinen, J. M., Meier, C. R., Arnal, J. D., & Leding, J. K. (2005). Compelling untruths: Content borrowing and vivid false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 954–963.

Lampinen, J. M., Neuschatz, J. S., & Payne, D. G. (1998). Memory illusions and consciousness: Examining the phenomenology of true and false memories. *Current Psychology, 16,* 181–224.

Marche, T. A., Brainerd, C. J., & Reyna, V. F. (2010). Distinguishing true from false memories in forensic contexts: Can phenomenology tell us what is real? *Applied Cognitive Psychology, 24,* 1168–1182.

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. doi:10.1080/09658210902939355

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language, 35,* 212–230.

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: The impact of presentation duration. *Journal of Memory and Language, 45,* 160–176. doi:10.1006/jmla.2000.2771

Murphy, G. L., & Shapiro, A. M. (1994). Forgetting of verbatim information in discourse. *Memory & Cognition, 22,* 85–94.

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language, 35,* 261–285.

Reyna, V. F. (1998). Fuzzy-trace theory and false memory. In M. Intons-Peterson & D. Best (Eds.), *Challenges and controversies: Memory distortions and their prevention* (pp. 15–27). Mahwah, NJ: Erlbaum.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7,* 1–75. doi:10.1016/1041-6080(95)90031-4

Reyna, V. F., & Kiernan, B. (1994). The development of gist versus verbatim memory in sentence recognition: Effects of lexical familiarity, semantic content, encoding instructions, and retention interval. *Developmental Psychology, 30,* 178–191.

Reyna, V. F., & Kiernan, B. (1995). Children's memory and interpretation of psychological metaphors. *Metaphor and Symbolic Activity, 10,* 309–331.

Reyna, V. F., & Lloyd, F. (1997). Theories of false memory in children and adults. *Learning and Individual Differences, 9,* 95–123. doi:10.1016/S1041-6080(97)90002-9

Reyna, V. F., & Mills, B. A. (2007). Interference processes in fuzzy-trace theory: Aging, Alzheimer's disease, and development. In C. MacLeod & D. Gorfein (Eds.), *Inhibition in cognition* (pp. 185–210). Washington, DC: American Psychological Association.

Reyna, V. F., & Titcomb, A. L. (1997). Constraints on the suggestibility of eyewitness testimony: A fuzzy-trace theory analysis. In D. G. Payne & F. G. Conrad (Eds.), *A synthesis of basic and applied approaches to human memory* (pp. 157–174). Hillsdale, NJ: Erlbaum.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95,* 318–339. doi:10.1037/0033-295X.95.3.318

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. doi:10.1037/0278-7393.21.4.803

Roediger, H. L., & McDermott, K. B. (2000). Tricks of memory. *Current Directions in Psychological Science, 9,* 123–127.

Roediger, H. L., III, Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8,* 385–407. doi:10.3758/BF03196177

Seamon, J. G., Luo, C. R., Shulman, S., Toner, S. K., & Caglar, S. (2002). False memories are hard to inhibit: Differential effects of directed forgetting on accurate and false recall in the DRM procedure. *Memory, 10,* 225–237.

Singer, M., & Remillard, G. (2008). Veridical and false memory for text: A multiprocess analysis. *Journal of Memory and Language, 59,* 18–35. doi:10.1016/j.jml.2008.01.005

Stahl, C., & Klauer, K. C. (2009). Measuring phantom recollection in the simplified conjoint recognition paradigm. *Journal of Memory and Language, 60,* 180–193. doi:10.1016/j.jml.2008.08.001

Strong, E. K. (1913). The effect of time-interval upon recognition memory. *Psychological Review, 20,* 339–371.

Tulving, E. (1985). How many memory systems are there? *American Psychologist, 40,* 385–398. doi:10.1037/0003-066X.40.4.385

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341–1354. doi:10.1037/0278-7393.20.6.1341