Psychological Review
2007, Vol. 114, No. 4, 954–993

0033-295X/07/$12.00 DOI: 10.1037/0033-295X.114.4.954

The fSAM Model of False Recall

Daniel R. Kimball and Troy A. Smith
University of Texas at Arlington

Michael J. Kahana
University of Pennsylvania

The authors report a new theory of false memory building upon existing associative memory models and implemented in fSAM, the first fully specified quantitative model of false recall. Participants frequently intrude unstudied critical words while recalling lists comprising their strongest semantic associates but infrequently produce other extralist and prior-list intrusions. The authors developed the theory by simulating recall of such lists, using factorial combinations of semantic mechanisms operating at encoding, retrieval, or both stages. During encoding, unstudied words' associations to list context were strengthened in proportion to their strength of semantic association either to each studied word or to all co-rehearsed words. During retrieval, words received preference in proportion to their strength of semantic association to the most recently recalled single word or multiple words. The authors simulated all intrusion types and veridical recall for lists varying in semantic association strength among studied and critical words from the same and different lists. Multiplicative semantic encoding and retrieval mechanisms performed well in combination. Using such combined mechanisms, the authors also simulated several core findings from the Deese–Roediger–McDermott paradigm literature, including developmental patterns, specific list effects, association strength effects, and true–false correlations. These results challenge existing false-memory theories.

Keywords: associative memory, false memory, memory model, semantic memory, search of associative memory (SAM)

One of the areas of most interest in memory research recently has been false memory—mistaken memory for an event that never occurred or that occurred but not as remembered. One paradigm that has reliably produced a robust false-memory effect in the laboratory is the Deese–Roediger–McDermott (DRM) paradigm (Deese, 1959b; Roediger & McDermott, 1995; see also Read, 1996). In this paradigm, participants study a list of words that are all strong semantic associates of an unpresented critical word. For example, participants may study the words *mad, hate, fear, ire, rage,* and so on, all of which are semantically related to the critical word *anger.* Participants later falsely recall and falsely recognize the critical word at rates that often approximate the rate of veridical recall. This effect is robust, having been replicated in many subsequent studies (for reviews, see Brainerd & Reyna, 2005; Gallo, 2006; Roediger, Balota, & Watson, 2001).

However, although intrusion of the critical word during recall occurs frequently, intrusions of other words—including words appearing in previously studied lists (prior-list intrusions) and words not appearing in any previously studied lists (extralist intrusions)—occur relatively rarely considering the much larger set of words from which such intrusions might be drawn (see, e.g., Kimball & Bjork, 2002). The combination of the high rate of critical word intrusions and the low rates of other intrusions provides an important constraint on models of false recall. A key goal of our modeling effort is to provide a global account of all such intrusions, not just intrusions of the critical word.

Existing Theories of False Memory

A number of theories incorporating associative memory processes have been applied to explain false recall in the DRM paradigm (for reviews, see Brainerd & Reyna, 2005; Gallo, 2006; Roediger, McDermott, & Robinson, 1998; Roediger, Watson, McDermott, & Gallo, 2001). For example, the implicit associative response hypothesis (Underwood, 1965) assumes that studying a word brings to mind a strongly associated word, such as the critical word. Similarly, spreading activation theories (Anderson & Bower, 1973; Collins & Loftus, 1975) assume that accessing words during study causes their memory representations to be activated and this activation to spread to other associated words across connections in a semantic network, with the critical word being repeatedly, and therefore strongly, activated during study of a DRM list. The source monitoring framework (Johnson, Hashtroudi, & Lindsay, 1993) similarly assumes that a word such as the critical word is internally generated during study of strongly associated words and that there then is source confusion at test in determining whether the reason for the word's salience in memory is due to its external presentation or internal generation at study (Mather, Henkel, & Johnson, 1997). Roediger, Balota, and Watson (2001) offered a combination of the activation and monitoring theories as the best explanation for false memory, and the

Daniel R. Kimball and Troy A. Smith, Department of Psychology, University of Texas at Arlington; Michael J. Kahana, Department of Psychology, University of Pennsylvania.

We acknowledge support from National Institutes of Health Research Grants MH55687 and MH079357 and Training Grant NS07292. Thanks are due to Mark Steyvers for providing us with access to the word association space database. The code for the fSAM model is available for download from the fSAM Web site (http://fsam.fusion.uta.edu).

Correspondence concerning this article should be addressed to Daniel R. Kimball, Department of Psychology, Box 19528, University of Texas at Arlington, Arlington, TX 76019. E-mail: dkimball@uta.edu

activation–monitoring theory is now one of the leading accounts of false memory in the DRM effect.

The other leading account for this phenomenon is fuzzy trace theory (Brainerd & Reyna, 1998, 2005; Reyna & Brainerd, 1995). This theory assumes that a verbatim trace and a gist trace are formed at encoding. The verbatim trace consists of surface details of studied words, including such aspects as phonology, orthography, and contextual information. The gist trace results from the retrieval and processing during study of semantic information, including semantic associations to unstudied words. Accessing the gist trace at retrieval promotes intrusion of the critical word, whereas accessing the verbatim trace promotes retrieval of the studied words and rejection of intrusions (the latter being termed *recollection rejection*). The gist trace is assumed to be more robust and stable than the verbatim trace. Fuzzy trace theory forms the basis of the only other published quantitative model of false recall of which we are aware (Brainerd, Payne, Wright, & Reyna, 2003; see also Brainerd & Reyna, 2005). However, the quantitative model covers only the theory's decision processes at test and not its encoding processes, including those involved in the creation of gist and verbatim traces.[1]

Both leading accounts of false memory in the DRM paradigm—the activation–monitoring theory and fuzzy trace theory—posit processes operating at both encoding and retrieval. There is evidence showing that false memory in this paradigm arises as a result of processes operating at both encoding and retrieval (for reviews, see Brainerd & Reyna, 2005; Gallo, 2006; Roediger et al., 1998; Roediger, Watson, et al., 2001). An example of evidence implicating encoding processes is the finding by Smith, Gerkens, Pierce, and Choi (2002) that the critical word showed evidence of indirect priming on a poststudy stem completion task, under conditions intended to minimize deliberative, conceptually driven retrieval of list words (see also McDermott, 1997; McKone & Murphy, 2000). On the other hand, a role for retrieval processes in the occurrence of false memories is implicated by such evidence as the finding that providing items from a studied DRM list as part-list cues during a recall test reduced false recall of the critical word (Bäuml & Kuhbandner, 2003; Kimball & Bjork, 2002, Experiment 2; Reysen & Nairne, 2002).

Overview of the fSAM Model

In this article, we present a new theory of false memory and report tests of its capacity to simulate patterns of veridical and false recall in several studies drawn from the DRM literature and a new experiment. The theory is implemented in a quantitative model, fSAM, which we developed within the framework of the search of associative memory (SAM) model of episodic recall (Gillund & Shiffrin, 1984; Kahana, 1996; Raaijmakers & Shiffrin, 1981; Shiffrin & Raaijmakers, 1992; Sirotin, Kimball, & Kahana, 2005). SAM is an associative model of memory positing that, during study, list items become episodically associated with each other and with the study context in proportion to the amount of time the items spend in a limited-capacity rehearsal buffer. SAM further assumes that retrieval from long-term memory (LTM) is cue dependent, with the list context and previously recalled items serving as retrieval cues for other items and the probability of retrieving an item being determined by strength-dependent competition among all items associated to a given set of cues. SAM has been applied to a broad range of free-recall phenomena, including the effects of presentation rate and list length (Raaijmakers & Shiffrin, 1980), part-set cuing (Raaijmakers & Shiffrin, 1981), word frequency (Gillund & Shiffrin, 1984), interference and forgetting (Mensink & Raaijmakers, 1988), list strength (Shiffrin, Ratcliff, & Clark, 1990), generation (Clark, 1995), and temporal contiguity (Kahana, 1996).

In addition to such episodically formed interitem and contextual associations, fSAM explicitly represents preexperimental semantic associations between pairs of words in a large-scale lexicon that includes both studied and unstudied words (see Sirotin et al., 2005). In the simulations described in this article, these pairwise association strengths were derived in two different ways. For the first three simulations, we used values derived from word association space (WAS; Steyvers, Shiffrin, & Nelson, 2005), which are in turn based on word association norms (Nelson, McEvoy, & Schreiber, 2004). For the last two simulations, we found it expedient to generate abstract association strength values so that we could illustrate more clearly the theoretical operation of the model. In the General Discussion, we discuss issues regarding WAS and other semantic association metrics (e.g., latent semantic analysis, Landauer & Dumais, 1997; Wordnet, Miller, 1996).

We developed several versions of fSAM that differ in the ways that these semantic associations exert an influence on episodic recall. One critical way that the model versions differ is in whether they incorporate a semantic mechanism at encoding, at retrieval, or at both stages. In addition, the particular semantic mechanism used at encoding can be one of three different versions, and there are also three different versions of the semantic retrieval mechanism. In the different versions of the semantic encoding mechanism, each word in the lexicon becomes associated to the list context in proportion to the word's strength of semantic association either to the most recently presented word alone or to all of the studied words jointly occupying the rehearsal buffer at a given time; if the latter, the association strengths combine either additively or multiplicatively. In a similar way, at test, the probability of retrieving a word is in part a function of its strength of semantic association either to the last recalled word alone or to all of the most recently recalled words jointly; if the latter, the strengths combine either additively or multiplicatively. These mechanisms share certain features with spreading activation (Collins & Loftus, 1975; Quillian, 1968) and compound cuing (Dosher & Rosedale, 1989; Ratcliff & McKoon, 1988), which we address in the General Discussion. By factorially combining the encoding and retrieval mechanisms, we generated 16 model versions comprising a 4 (semantic encoding mechanism: none, single-item, additive, multiplicative) × 4 (semantic retrieval mechanism: none, single-item, additive, multiplicative) design. We compared the performance of these 16 model versions in Simulation 1, and in Simulations 2–5, we used the best performing model version from Simulation

[1] Arndt and Hirshman (1998) have used the MINERVA2 model to simulate false recognition in the DRM paradigm. The search of associative memory model (SAM), on which our fSAM model is based, has been used to model recognition processes (Gillund & Shiffrin, 1984), and we discuss the possibility of using it to simulate false recognition in the General Discussion. Our current modeling focuses on simulating false recall using SAM's recall processes.

1—the version that combines the multiplicative encoding and retrieval mechanisms.

Overview of Simulated Experiments

To provide a strong test of fSAM's capabilities, we sought a strong set of constraints. We report the results of five simulations that cover several DRM experiments. In Simulation 1, we sought to fit the patterns of veridical recall, critical word intrusions, extralist intrusions, and prior-list intrusions in conditions drawn from Kimball and Bjork (2002, Experiment 2) and a new experiment designed especially for purposes of testing the model. The simulated conditions in this initial fit included conditions using typical DRM lists, tested immediately and at a delay, as well as a condition in which the words from particular DRM lists were distributed across studied lists and a condition in which studied words were not systematically related. We used a single parameter set for all conditions to test whether the intricate pattern of veridical and false recall across the conditions could be captured by the model's mechanisms, rather than by changes in parameter values.

Simulation 2 addressed the developmental pattern of veridical and false recall in children. Younger children produce low rates of both veridical recall and semantically induced intrusions, such as critical word intrusions, both of which subsequently increase with development; by contrast, prior-list intrusions decrease with development (Brainerd, Reyna, & Forrest, 2002; Dewhurst & Robinson, 2004).

In Simulation 3, we fit the pattern of veridical-recall and critical word intrusion rates across specific DRM lists. Despite substantial disagreement in the literature as to these rates for particular lists and sets of lists, we sought to simulate the patterns reported in two studies, Stadler, Roediger, and McDermott (1999) and Gallo and Roediger (2002). We tested our model's capacity to simulate the mean rates of veridical recall and critical word intrusions, collapsing across lists, as well as the pattern of individual list means and the correlation across lists between veridical recall and critical word intrusions in these studies.

In Simulations 4 and 5, we used abstract association strengths to simulate the effects on veridical and false recall of differences in backward association strength (the propensity of studied words to elicit the critical word in a free-association task), forward association strength (the propensity of the critical word to elicit studied words), connectivity (the propensity of one studied word to elicit another studied word), and number of critical word associates appearing in the studied list. Except for backward association strength, reports of the effects of these variables have been mixed and, in some cases, confounded with other variables, as we describe more fully in the preamble for Simulation 4. Our purpose in these simulations was to generate theoretical predictions for these effects unconfounded by other factors, to understand better how the model operates.

We next describe in more detail the basic SAM model, the new mechanisms incorporated in the fSAM model, and the simulations. In the General Discussion, we examine the ramifications of our new mechanisms for theories of memory, including spreading activation theory, the source monitoring framework, fuzzy trace theory, compound cue theory, and SAM.

The SAM Model

In this section, we describe the simplified simulation recall model first reported by Raaijmakers and Shiffrin (1980, 1981), which forms the foundation of SAM, and we also describe a number of subsequent modifications to that model. To the extent applicable, we distinguish these instantiations from the general SAM theory, which contemplates certain features that were not implemented in the original simplified model.

Memory Stores

The SAM model assumes the existence of two memory stores: short-term memory (STM) and LTM. Within STM, rehearsal processes are idealized in the form of a limited-capacity buffer in which studied words become associated through a rehearsal process, as described below. LTM contains values for the strengths of two types of associations: the associations formed at study between each list word and the list context, and the pairwise episodic associations formed among list words during study.

In the basic SAM model, the strengths of item-to-context and interitem associations formed during study are stored in an episodic matrix (e.g., Gillund & Shiffrin, 1984; Mensink & Raaijmakers, 1988; Raaijmakers & Shiffrin, 1981; Shiffrin & Raaijmakers, 1992). List context is conceptualized as the temporal and situational setting for a particular list. For the sake of simplicity, the basic SAM model assumes that all associations in LTM are episodically created in the course of rehearsal during study, so the strengths in the episodic matrix are set to zero prior to study (although these associative strengths are later reset to a residual value for pairs of words that are not rehearsed together during study). $S(i, \mathit{context})$ denotes the strength of association between word i and the list context, and $S_e(i, j)$ denotes the strength of association that is generated episodically between words i and j.

Storage Process

During study of a list, SAM assumes that, as each list item is presented, it enters the STM buffer and is rehearsed along with other items occupying the buffer at any given time, thereby increasing the strengths of the items' episodic associations in LTM. In particular, rehearsal increases the strength of association between each item in the buffer and the list context; for each unit of time, the strengths of the associations between the context and all items then occupying the buffer are incremented in LTM by an equal proportion of the value of parameter a.

Rehearsal also increases the strength of the association in LTM between any two items that simultaneously occupy the buffer; for each unit of time, the interitem episodic associative strengths for the pairs of items then occupying the buffer are incremented by an equal proportion of the value of parameter b. Kahana (1996) substituted two parameters in lieu of b, one being used to increment interitem strengths in the forward direction—that is, from earlier presented items to later presented items (b_1)—and the other being used to increment strengths in the backward direction, from later presented items to earlier presented items (b_2). This enabled Kahana to simulate the bias in output order that favors sequential output during recall of items studied close to each other in time, particularly recall transitions from earlier to later studied items.

Our model incorporates this feature, although only b_1 is a free parameter in our model; b_2 is fixed at $.5*b_1$, as in Sirotin et al. (2005).

SAM also represents the association of an item to itself—that is, autoassociation—and includes parameters that increment an item's autoassociative strength when the item occupies the buffer in STM during study (parameter c) and when it is output during recall (parameter g). In our simulations, all autoassociative strengths and parameters c and g are set to zero, as they were in Sirotin et al. (2005).

The amount of time that each item spends in the STM buffer during study is determined by the presentation rate, the size of the buffer (the maximum number of items that can simultaneously occupy the buffer), and the rule for displacement of items from the buffer. In Raaijmakers and Shiffrin (1981) and other early implementations, the size of the STM rehearsal buffer, r, was set at a single fixed value for all subjects, with $r = 4$ typically providing the best fit to free-recall data. Kahana (1996) found it useful to allow the size of the buffer to vary for each subject, with r being randomly selected from a distribution having mean μ_{r} and standard deviation σ_{r}. Our model uses such a distribution of buffer sizes, with $\mu_{r} = 4$ for our simulations of adult recall and lower values for children, as described in connection with Simulation 2.

Once the buffer is full, each new item displaces one of the items then occupying the buffer. The general SAM theory is silent concerning the particular rule governing displacement. The simulation model of Raaijmakers and Shiffrin (1981) assumes that each item in STM has an equal probability of being displaced by the new item. Kahana (1996) found that an alternative displacement rule proposed by Phillips, Shiffrin, and Atkinson (1967) provided a better fit to data on free recall. The Phillips et al. rule assumes a bias in favor of displacing items that have been in the buffer longer than others. Under this rule, the probability that the ith buffer item is to be displaced is given by

$$P(i\ displaced) = \frac{q(1-q)^{i-1}}{1-(1-q)^{r}}, \tag{1}$$

where q is a fixed parameter of the model that determines the degree of bias favoring displacement of older items. Later presented items occupy higher ordinal positions in the buffer than earlier-presented items, thus insuring a bias under the displacement rule favoring displacement of earlier presented items. When a new item displaces an old item, each old item that remains in the buffer and occupies a higher ordinal position than the displaced item shifts to occupy the next lower ordinal position, and the new item enters the buffer in the highest ordinal position. Our model incorporates the Phillips et al. displacement rule, as implemented by Sirotin et al. (2005).

For a pair of list items that are never rehearsed together in the STM buffer during study, the basic SAM model assigns the pair a residual interitem strength of episodic association in LTM equal to the value of parameter d. Our model incorporates the general concept of residual episodic strength but implements the concept by initializing the episodic associative strengths in the episodic matrix to random values drawn from a normal distribution with mean μ and standard deviation σ, which are fixed parameters of the model, instead of initializing the strengths to zero (see Sirotin et al., 2005). The same distribution is used to initialize the contextual association strengths contained in the context vector.

Retrieval Process

SAM posits a two-stage retrieval process for immediate free recall, the first stage involving the output of items in the STM buffer at the beginning of recall and the second stage involving retrieval of items from LTM. According to SAM, items in the STM buffer are always available for recall, and the items in the buffer at the end of list presentation are always output first during immediate free recall. This assumption allows SAM to account for the pronounced recency effect observed in classic immediate recall experiments in which participants are presented with a large number of lists (e.g., Murdock, 1962), and tend to develop a last-in-first-out strategy on their own after a few study–test trials (see also Murdock & Okada, 1970, who discarded the first several trials). A pronounced recency effect is also observed when participants are explicitly instructed to begin recall with the last few items (e.g., Roediger & McDermott, 1995). However, SAM is likely to overestimate the recency effect in immediate free-recall data to the extent that human participants do not begin recall output with recency items, as occurred for a substantial proportion of the trials that we simulated from Kimball and Bjork (2002, Experiment 2). We address this point in our discussion of that experiment in Simulation 1 and in the General Discussion.

In delayed free recall, STM is assumed to have been emptied during the retention interval, and recall therefore begins with retrieval from LTM. The simulation model of Raaijmakers and Shiffrin (1981) assumes that, following the end of the study period, the buffer empties at the same rate as items were displaced during study. This method of gradually emptying the contents of STM enables SAM models to account for data showing that delaying free recall eliminates positive recency but that such a delay does not produce negative recency (Postman & Phillips, 1965).

Search of LTM

Retrieval of items from LTM results from strength-dependent competition among all items associated to a given set of retrieval cues. Each cycle of the search process includes at least two phases: First, an item is *sampled,* and then it may or may not be *recovered,* that is, identified as a particular word.

SAM begins the search of LTM using context as the sole retrieval cue. The probability of sampling an item i when using context alone as a retrieval cue is

$$P_s(i|context) = \frac{S(i,\ context)}{\sum_{k}^{k\in N} S(k,\ context)}, \tag{2}$$

where N is the set of items stored in LTM. This equation ensures that items with greater strengths of association to the list context are more likely to be sampled. Once sampled, the probability that item i is recovered is

$$P_r(i|context) = 1 - e^{-S(i,context)}. \tag{3}$$

Thus, recovery also depends on the strength of association between the item and the list context.

Once an item is recalled, it is then used in combination with context to cue recall of another item. In the basic SAM model, the probability of sampling item *i,* given that both context and the just-recalled item *j* serve as retrieval cues, is

$$P_s(i|context, j) = \frac{S(i, context)S_e(i, j)}{\sum_{k}^{k \in N} S(k, context)S_e(k, j)}, \quad (4)$$

and the probability of recovering the item is

$$P_r(i|context, j) = 1 - e^{-[S(i, context) + S_e(i, j)]} \quad (5)$$

However, regardless of an item's strength of association either to the context, to other retrieved items, or to both, the item cannot be recalled if the same retrieval cues failed to recover the item previously or if the item has previously been recalled. Instead, such retrieval attempts are counted as retrieval failures for purposes of the stopping rules discussed later (Raaijmakers & Shiffrin, 1981).

Increment in Strengths Following Retrieval

When retrieval cues are successful in recovering an item, the strengths of the item's associations to the retrieval cues are incremented in LTM. The strength of association between the recovered item and the list context is incremented by the value of parameter *e,* and the episodic strength of association between the recovered item and any other item then being used as a retrieval cue is incremented by the value of parameter *f.* Thus, different parameters are used at test than at encoding to increment associative strengths in LTM. As with parameter *b,* Kahana (1996) bifurcated parameter *f* into f_1 for forward associations (i.e., from earlier to later studied items) and f_2 for backward associations, and our model adopts this bifurcation.

Stopping Rules

There are two rules determining when a subject stops searching, one governing when a subject abandons search with a particular set of retrieval cues and a second governing when the subject abandons search altogether. When there have been L_{max} consecutive failures at recovery using a particular item together with context as retrieval cues, SAM assumes that the subject reverts to using the context alone as a retrieval cue. Search stops altogether when K_{max} recovery failures have accumulated over all sets of retrieval cues (Raaijmakers & Shiffrin, 1981).

Contextual Drift

Sirotin et al. (2005) used a contextual drift mechanism in which the item-to-context strengths decay probabilistically after a subject finishes recalling each list according to an exponential decay rule:

$$S(i, context)_l = \rho S(i, context)_{l-1} + \varepsilon, \quad (6)$$

where *l* is a counter that represents the index of the current list; $S(i, context)_{l-1}$ and $S(i, context)_l$ are the levels of item-to-context strength for lists $l - 1$ and *l,* respectively; ρ represents the proportion of the item-to-context strength that is conserved between lists; and ε is a noise term taken from a normal distribution with mean μ and standard deviation σ (cf. Mensink & Raaijmakers, 1989). Our model adopts this contextual drift mechanism.

The fSAM Model

In addition to the features of the SAM framework discussed above, the fSAM model includes explicit representation of preexperimental semantic associations among words, which are used in the new semantic mechanisms that operate at encoding and retrieval.

Semantic Matrix

Preexperimental semantic association strengths are stored in a separate semantic matrix (see Sirotin et al., 2005). The semantic matrix incorporates a large lexicon of words, including those presented on different lists during the experiment and those not presented during the experiment at all. The strength of the semantic association between words *i* and *j* is denoted as $S_s(i, j)$ and remains fixed during the course of the experiment, reflecting a simplifying assumption that semantic associations are not significantly affected by episodic experience on the scale of a single experiment. In our simulations, we used two different methods of specifying semantic association strengths, either basing the values on behavioral word association norms or selecting the values from abstract distributions having certain properties. We discuss these methods in more detail in the simulations section.

Semantic Encoding Mechanisms

The new semantic encoding mechanisms provide for the incrementing during study of the item-to-context association strength of each word in the lexicon in proportion to either (a) the word's strength of semantic association to the most recently presented word (single-item encoding mechanism), (b) the sum of the strengths of the word's semantic associations to all the items being rehearsed in STM at a given time (additive encoding mechanism), or (c) the product of those strengths (multiplicative encoding mechanism). We refer to the additive and multiplicative encoding mechanisms collectively as multiple-item encoding mechanisms.

The incremented contextual association strength is determined by the following equations for these three versions of the semantic encoding mechanism:

(a) Single-item semantic encoding:

$$S(i, context)_t = S(i, context)_{t-1} + a_s S_s(i, j), \quad (7)$$

(b) Additive semantic encoding:

$$S(i, context)_t = S(i, context)_{t-1} + a_s \sum_{j}^{j \in M} S_s(i, j), \quad (8)$$

(c) Multiplicative semantic encoding:

$$S(i, context)_t = S(i, context)_{t-1} + a_s \prod_{j}^{j \in M} S_s(i, j). \quad (9)$$

In the above equations, i is an index for all items in the lexicon, M is the set of items in STM at a given time, j is an index for the items in M at a given time, $S(i, context)$ is the strength of the association between item i and context, $S_s(i, j)$ is the strength of the semantic association between items i and j, t is an index for time increments, and new parameter a_s serves to scale the semantically related increment in contextual association strength.

The incrementing of contextual association strengths reflects an assumption that unpresented words such as the critical word become associated to the study episode in a global way and not to individual words encountered in the study episode. The latter approach would be most straightforwardly implemented in our framework by incrementing the strengths of interitem episodic associations to the critical word, as might occur if the critical word were consciously intruded into study-phase rehearsal. There is some evidence that such conscious intrusion does take place during encoding (Goodwin, Meissner, & Ericsson, 2001). However, our model makes the simplifying assumption that any effect of encoding on false recall arises due to automatic, unconscious processing.

We considered other alternatives by which associative strengths might reflect semantic encoding of unpresented words. One possibility was creation of a new hybrid semantic–episodic association strength. We rejected this option because it would have involved an additional matrix of association strengths, as well as an additional retrieval weight parameter, and we thought this lacked parsimony. We also chose not to increment semantic association strengths because we made the simplifying assumption that such associations are not significantly affected by experience on the scale of a single experiment. Thus, we assume that any changes in association strengths involving unpresented words are transiently attributable to the particular set of words being studied in the context of a given episode. Although we chose to increment contextual association strength in proportion to semantic association strength for the foregoing reasons, we do not rule out alternative formulations. Such formulations may prove useful or necessary to simulate other effects reported in the literature.

Semantic Retrieval Mechanisms

For the semantic retrieval mechanisms, we modified the sampling and recovery rules in SAM to reflect the use of preexisting semantic associations when cuing with context plus recovered items. The probability of sampling and recovering item i when using context alone is the same as in SAM and is calculated using the item's contextual association strength.

The new semantic retrieval mechanisms operate after a word has been recovered using context-only cuing. In SAM, cuing after recovery of a word involves the product of two types of associative strength: a retrieval candidate's strength of association to list context and its episodic strength of association to the just-recovered word. Our new semantic retrieval mechanisms multiply that product further by either (a) the strength of the candidate's semantic association to the most recently recalled word (single-cue retrieval mechanism), (b) the sum of the strengths of the candidate's semantic associations to the recently recalled words then occupying the STM buffer (additive retrieval mechanism), or (c) the product of those strengths (multiplicative retrieval mechanism). We refer to the additive and multiplicative retrieval mechanisms collectively as multiple-item retrieval mechanisms.

In model versions that combine semantic encoding and semantic retrieval mechanisms, the contextual association strengths available at retrieval reflect the effects of semantic encoding at study. Thus, during either context-only or context-plus-items cuing at retrieval in those models, unstudied items (e.g., the critical word) will be more or less retrievable as a function of the degree to which they were encoded semantically at study.

The W_s parameter is used to weight semantic retrieval cues relative to the weights of contextual and episodic retrieval cues (W_c and W_e, respectively). The single-cue retrieval mechanism uses the same sampling and recovery rules as were used by Sirotin et al. (2005). The probability of sampling item i following the recovery of item j is

$$P_s(i|context, j) = \frac{S(i, context)^{W_c} S_e(i, j)^{W_e} S_s(i, j)^{W_s}}{\sum_{k}^{k \in N} S(k, context)^{W_c} S_e(k, j)^{W_e} S_s(k, j)^{W_s}}, \tag{10}$$

and the probability of recovering a sampled word is

$$P_r(i|context, j) = 1 - e^{-[W_c S(i, context) + W_e S_e(i, j) + W_s S_s(i, j)]}, \tag{11}$$

where $S(i, context)$ represents the strength of the association of item i to context; $S_e(i, j)$ represents the strength of the episodic association between the most recently recalled item (item j) and item i; $S_s(i, j)$ represents the strength of the semantic association between the most recently recalled item and item i; W_c, W_e, and W_s are parameters for weighting of item-to-context associations, interitem episodic associations, and interitem semantic associations, respectively; and N is the set of all items in LTM.

For the multiple-item retrieval mechanisms, the most recently retrieved items (including any intrusions) are assumed to be stored in STM and are used as multiple-item retrieval cues, with the maximum number of such cues being equal to r, the size of the STM buffer. After r words have been recalled from LTM, the earliest recalled item then in the buffer is displaced by each successive recalled word. This first-in-first-out displacement rule is different from that operating at encoding, reflecting an absence of rehearsal processes.

The first step in the operation of the additive retrieval mechanism involves calculating, for each word in the lexicon, the sums of the semantic and episodic association strengths between that word and all of the recently retrieved items then occupying the STM buffer. These sums are then used to calculate the sampling and recovery probabilities when cuing LTM with item and context information. Thus, the probability of sampling item i following the recovery of items j_1, j_2, j_3, and so on becomes

$$P_s(i|context, j \in M)$$

$$= \frac{S(i, context)^{W_c}\left(\sum_{j}^{j \in M} S_e(i, j)\right)^{W_e}\left(\sum_{j}^{j \in M} S_s(i, j)\right)^{W_s}}{\sum_{k}^{k \in N}\left[S(k, context)^{W_c}\left(\sum_{j}^{j \in M} S_e(k, j)\right)^{W_e}\left(\sum_{j}^{j \in M} S_s(k, j)\right)^{W_s}\right]}, \tag{12}$$

and the probability of recovering a sampled word becomes

$$P_r(i|context, j \in M) = 1 - e^{-[W_c S(i,\, context) + W_e(\Sigma_j^{j \in M} S_e(i,\, j)) + W_s(\Sigma_j^{j \in M} S_s(i,\, j))]}, \tag{13}$$

where M is the set of recently recovered items (j_1, j_2, j_3, etc.) then occupying STM. The multiplicative retrieval rule is similar, except that it involves calculation of products of episodic and semantic association strengths rather than sums, so that the probability of sampling item i becomes

$$P_s(i|context,j \in M) = \frac{S(i,\, context)^{W_c}\left(\prod_j^{j \in M} S_e(i,\, j)\right)^{W_e}\left(\prod_j^{j \in M} S_s(i,\, j)\right)^{W_s}}{\sum_k^{k \in N}\left[S(k,\, context)^{W_c}\left(\prod_j^{j \in M} S_e(k,\, j)\right)^{W_e}\left(\prod_j^{j \in M} S_s(k,\, j)\right)^{W_s}\right]}, \tag{14}$$

and the probability of recovering a sampled word becomes

$$P_r(i|context, j \in M) = 1 - e^{-[W_c S(i,\, context) + W_e(\Pi_j^{j \in M} S_e(i,\, j)) + W_s(\Pi_j^{j \in M} S_s(i,\, j))]}. \tag{15}$$

It is evident from these equations that the semantic retrieval cuing mechanisms do not operate in isolation, but rather, they work in combination with contextual and episodic cuing. Words will have a higher probability of retrieval if they have been studied—giving them stronger associations to context and stronger episodic associations to other studied words—and if they are semantically associated to studied words. Each type of association acts to modulate the influence of the other types of associations. Thus, there are trade-offs in the extent to which the different types of associations play a role in cuing retrieval of a particular word. For example, a word that has not been studied and that consequently has lower strengths of association to the study context or to studied words episodically may nevertheless be retrieved if its semantic associations to recently recalled words are strong enough (e.g., the critical word). On the other hand, a studied word that has strong episodic associations to other items and to the study context may not be successfully retrieved because its semantic associations to recently recalled words are comparatively weak.

Expectations for Simulations of Critical Word Intrusions and Other Intrusions

A key difference among the three versions of each mechanism is in the degree to which they selectively target unstudied words—such as the critical word—that are strongly related to a high proportion of the words that either are rehearsed together or have been output during recall. Of course, the versions that use only a single pairwise semantic association, to either the last-studied or last-recalled word, accord no such selective targeting of words that have strong associations to multiple words. By contrast, both the additive and the multiplicative versions of each mechanism selectively target such multiply associated words. All other things being equal, the multiplicative mechanisms can be expected to target such words more selectively than the additive mechanisms because multiplying the semantic strengths creates a larger proportional difference between multiply associated words and other words than does adding the semantic strengths. Accordingly, we expected that the simulation of critical word intrusions would be easier for the mechanisms using multiple semantic associations than for those using single semantic associations and easier for the multiplicative mechanisms than for the additive mechanisms.

By contrast, we surmised that words produced as extralist and prior-list intrusions, to the extent such intrusions are semantically induced, would tend to be strongly related to one or two studied words but generally not to more than that. Accordingly, we expected that models incorporating the single-association mechanisms would favor production of these noncritical intrusions. Therefore, all other things being equal, we predicted that the mechanisms would trade off in their ability to produce critical word intrusions versus other intrusions. Given the high rate of critical word intrusions and the low rates of other intrusions in DRM lists, we expected that the multiple-item mechanisms would work best in simulating recall of those lists, with an advantage for the multiplicative over the additive mechanisms.

Model Parameters

Table 1 lists the 10 free parameters used in the fSAM model and provides a brief explanation of their function. Eight of the parameters listed in Table 1 were inherited from previous instantiations

Table 1
Free Parameters and Their Functions

Process	Parameter	Description
Encoding	a	Increment in item-context strength with rehearsal during study
	a_s	Scaling factor for incrementing item-context strength as a function of semantic associations to studied items
	b_1	Increment in forward interitem episodic strength with rehearsal during study
Retrieval	W_c	Retrieval weight for item-context strength
	W_e	Retrieval weight for interitem episodic strength
	W_s	Retrieval weight for interitem semantic strength
	K_{max}	Maximum number of cumulative recovery failures during recall
Output encoding	e	Increment in item-context strength after recovery during recall
	f_1	Increment in forward interitem episodic strength after recovery during recall
Forgetting	ρ	Proportion of item-context strength that is preserved between lists

of SAM. Only two new parameters are added, one for each new mechanism: the a_s parameter for semantic encoding and the W_s parameter for semantic retrieval (see Sirotin et al., 2005). Thus, we succeeded in keeping the number of free parameters within the range used in previous SAM-type models (cf. Gillund & Shiffrin, 1981, 1984; Gronlund & Shiffrin, 1986; Mensink & Raaijmakers, 1988; Raaijmakers, 2003; Raaijmakers & Shiffrin, 1981; Sirotin et al., 2005). The number of parameters per simulated process was also quite small: three for encoding (a, a_s, b_1), four for retrieval (W_c, W_e, W_s, K_{max}), two for output encoding (e, f_1), and one for forgetting (ρ).

Simulation 1: Fit of Kimball and Bjork (2002, Experiment 2) and the New Experiment

In this simulation, we used a single parameter set to fit four experimental conditions simultaneously—the free-recall condition from Kimball and Bjork (2002, Experiment 2) and all three conditions in the new experiment. For convenience, we refer to Kimball and Bjork and the new experiment's standard condition collectively as the DRM conditions and to the other two conditions in the new experiment (mixed and control) collectively as the non-DRM conditions.

Kimball and Bjork (2002, Experiment 2)

Kimball and Bjork (2002, Experiment 2) presented 12 DRM lists of 15 words each by audiotape at a rate of 1.5 s per word. The 12 lists were based on the following 12 critical words, presented in the following randomly determined order to all participants: *smell, anger, sleep, sweet, cold, slow, smoke, rough, needle, soft, chair,* and *window*. Within each list, the words were presented in the same order as in Roediger and McDermott (1995), roughly in descending order of strength of association to the critical word. Each list was followed by an immediate free-recall test, for which participants were instructed to write as many words within 90 s as they could be confident they had heard on that list. Eight of the lists were tested with part-list cues present, and the remaining four lists were tested with no such cues present. Our current simulations focus on the latter, uncued lists.

Intrusions of the critical word occurred at an average rate of $M = .54$, commensurate with other findings reported in the literature. Also consistent with other reports in the literature, the critical word tended to be output late in recall, with an average output percentile of $M = 66$. In contrast to the high rate of critical word intrusions, the number of extralist intrusions and prior-list intrusions per list averaged only $M = .32$ and $M = .01$, respectively, despite the much larger set of words from which they might be drawn, compared with the single critical word. The mean proportion of studied words that were recalled was $M = .53$. For reasons described in the *Simulation Method* section, we fit veridical recall for the first 11 serial positions only ($M = .51$), but we also report the behavioral and simulated means for the last four serial positions.

The New Experiment

In this section, we summarize the method and results for the new experiment; they are described in detail in Appendix A. The three between-subjects conditions—standard, mixed, and control—differed in the way the lists were constructed. For the standard condition, we used the same 12 DRM lists as in Kimball and Bjork (2002, Experiment 2) plus three additional lists based on the following critical words: *doctor, mountain,* and *trash*. The lists were presented auditorily by computer, and the presentation orders of lists and of words within lists were randomized anew for each participant. Each list was followed by a filler task (math problems) for 30 s and then a delayed free-recall test. The materials and procedure were otherwise similar to those for the uncued lists in Kimball and Bjork.

The mixed condition used the same 225 studied words as in the standard condition, but exactly one word from each of the 15 DRM lists was included in each of the 15 studied lists. Subject to that constraint, the assignment of words to lists was randomized anew for each participant, as was the presentation order of words within lists. The control condition used 225 words that bore no systematic semantic relationship to each other and were matched with the 225 studied words in the other two conditions on several dimensions—word frequency, number of letters, number of syllables, and normed judgments of concreteness, imageability, and familiarity. Words in the control condition were randomly assigned anew to lists and serial positions for each participant. The materials and procedure were otherwise similar to those for the standard condition.

Results showed that veridical recall in the standard condition ($M = .47$) was comparable to that in Kimball and Bjork (2002) and significantly higher than in the mixed condition ($M = .31$) and the control condition ($M = .32$), which did not differ reliably. The critical word intrusion rate in the standard condition ($M = .49$) was also comparable to that in Kimball and Bjork and significantly higher than in the mixed condition ($M = .07$), which was in turn slightly but significantly higher than in the control condition ($M = .01$). Extralist intrusions were higher than in Kimball and Bjork and did not differ reliably across the three conditions (Ms = .55, .60, and .55, in the standard, mixed, and control conditions, respectively). Prior-list intrusions in the standard condition ($M = .08$) were slightly higher than in Kimball and Bjork, slightly but significantly higher still in the control condition ($M = .13$), and significantly higher still in the mixed condition ($M = .33$).

Simulation Method

Measures of Goodness of Fit

Our basic goodness-of-fit measure was the root-mean-square deviation (RMSD). RMSD has the advantage of being measured in the same units as the dependent measures that are being fit and is interpretable as a global measure of the difference between the model's predictions and the observed data averaged across all the dependent measures. RMSD values are weighted averages of the differences between the means of each dependent variable for human versus simulated subjects.

In addition, because we were comparing the fits of models that have different numbers of parameters, we calculated Schwarz's (1978) Bayesian information criterion (BIC) for each model's fit:

$$\mathrm{BIC} = k \ln n + n \ln(\mathrm{RMSD})^2, \; n > 1, \qquad (16)$$

where k is the number of parameters and n is the number of dependent variables being fit. Lower BIC values indicate better fits. Using BIC adjusts for the improvement in fit that typically results from adding parameters, although it does not directly adjust for differences in the number or complexity of mechanisms. The number of parameters varied from 8 for the model version without any semantic mechanisms to 9 for those versions incorporating a semantic encoding or retrieval mechanism, but not both, to 10 for those versions incorporating both semantic encoding and retrieval mechanisms. In the tables reporting RMSD values, we also report the BIC rank order for all models, although it is only for models with different numbers of parameters that these rank orders provide information not already reflected in the RMSD values. The BIC values are reported in full in Appendix B.

Fitting Procedure

We used a genetic algorithm (Mitchell, 1996) to generate parameter sets. For the first generation of parameter sets, the value of each parameter in a set was randomly selected from a predetermined range of values. Each parameter set was then used with the model to simulate the behavioral results. The mean values for the dependent variables were calculated across the simulated subjects and compared with the behavioral means to calculate RMSD.

The first-generation parameter sets that yielded the lowest RMSD values were then used to create the next generation of parameter sets through the processes of mutation and recombination. Mutation creates a particular second-generation parameter set by randomly selecting one of the best fitting first-generation parameter sets and then randomly copying or varying the value of each parameter within a specified range. During mutation, parameters were allowed to mutate by up to 0.5% of the parameter's range, in either direction, subject to the parameter's upper and lower range limits. Recombination creates a second-generation parameter set by randomly selecting two of the best fitting first-generation parameter sets as "parents" and, for each parameter, randomly selecting one parent's values for that parameter as the "child's" value.

These processes iterated through 10 generations or more as needed until the minimum RMSD value per generation reached asymptote. The best fitting parameter sets generated by the genetic algorithm were then each used to run the model again, using samples of 200 simulated subjects to generate statistically stable predictions and allow for regression to the mean. All values we report are from the large-sample run that produced the best quantitative fit to the data.

Dependent Variables

We simultaneously fit the recall patterns for four types of dependent variables: veridical recall, critical word intrusions, extralist intrusions, and prior-list intrusions.

Veridical recall. We fit both the overall level of veridical recall and the serial position curve, weighting the two equally. We eventually decided not to include in RMSD calculations the differences in means for the recency portion of the serial position curve for immediate recall in Kimball and Bjork (2002). Early attempts to simulate recall of the recency items failed because, contrary to the assumption of SAM, not all participants on all trials in Kimball and Bjork started their recall with the recency items. Participants started recall with an item from one of the last three serial positions on only 37% of the lists and started recall with an item from one of the first three serial positions on 36% of the lists. As a consequence, there was a much less pronounced recency effect than SAM predicts. Because fitting the recency portion of the curve was not of central importance for our present purposes, we excluded recall of the last four serial positions in our fits of Kimball and Bjork to prevent the fitting algorithm from laboring in vain to fit the low levels of recency item recall in the Kimball and Bjork behavioral data. Nevertheless, we report the behavioral and simulated means for the complete serial position curve, including those four serial positions.

Critical word intrusions. We included two measures related to critical word intrusions. For all conditions, we fit the proportion of critical words intruded. In the DRM conditions, there were a sufficient number of such intrusions to permit stable estimates of output percentiles, so we also fit the output percentile of the critical word when it was intruded in those conditions. Not only must a model fit the high level of critical word intrusions for DRM lists but it must also produce the critical word relatively late in the recall output sequence, as was observed in the behavioral data. Only those lists for which a subject intruded the critical word were included in calculating the subject's mean output percentile, and only those subjects who intruded at least one critical word were included in calculating the overall mean across subjects. In the fit, we accorded the percentile measure only one fifth of the weight accorded to the critical word intrusion rate. We found this reduced weighting of the output percentile to be necessary to avoid fits that generated extremely low numbers of critical word intrusions that nevertheless occurred relatively late in the output sequence. We considered it more important theoretically for the model to generate appropriate levels of critical word intrusions than to generate them at the appropriate point in the output sequence. As it turned out, the model was able to fit the output percentiles fairly well even with this reduced weighting.

Extralist and prior-list intrusions. Extralist and prior-list intrusions are an often overlooked but extremely important element of false recall. Sirotin et al. (2005) demonstrated that a SAM-type model could simulate appropriate levels of extralist intrusions and prior-list intrusions using single-item semantic cuing at recall and a mechanism permitting contextual drift across lists. However, the ability of a SAM-type model to account for the complete pattern of false recall in the DRM paradigm—critical words along with extralist intrusions and prior-list intrusions—had not been tested until now. Simultaneously simulating the relatively low levels of extralist intrusions and prior-list intrusions along with the relatively high level of critical word intrusions for DRM lists is an important test for any computational model of false recall. Of course, the model should also be able to fit the pattern of extralist intrusions and prior-list intrusions in lists of words that are not systematically related, as well as in sets of lists with strong semantic relations among items from different lists, as in the control and mixed conditions of the new experiment, respectively.

Metric for Semantic Association Strength

Although our theory is neutral as to the measure of semantic association strength to be used, we used WAS (Steyvers et al., 2005) in Simulations 1–3. The WAS metric uses a mathematical technique called singular value decomposition to transform the free-association word norms collected by Nelson et al. (2004) into a multidimensional semantic space. The asymmetric associative strengths given by the norms are made symmetric by summing the forward and backward associative strengths (e.g., the tendency of *dog* to evoke *cat* and the tendency of *cat* to evoke *dog*). Each word is then represented as a vector of the strengths of its associations to other words, and singular value decomposition is applied to reduce the dimensionality of the resulting matrix. The relatedness of two words can be calculated as the cosine of the angle between their vectors in semantic space. Using this method, words that are directly associated or that share associates have large cosine values. Words that do not directly share associations may still have reasonably high relatedness values because of their indirect associations, although the strengths of indirect associations are lower than those of direct associations, all other things being equal. The method used to construct WAS is discussed in greater detail in Steyvers et al. (2005).

We do not rule out the use of other metrics, such as latent semantic analysis (Landauer & Dumais, 1997) or Wordnet (Miller, 1996). However, WAS has three main advantages for our purposes. First, WAS is based on normed probabilities of producing associates of words, which seems quite similar to producing words in free recall (see Steyvers et al., 2005). Second, WAS provides a value for each pair of words in the word association norm database, allowing us to avoid problems that would arise from using zero values in our formulas. Third, as discussed by Sirotin et al. (2005), WAS discriminates semantically related words from unrelated words better than at least one other candidate metric, latent semantic analysis.

Notwithstanding these advantages, WAS was not ideal for our purposes. For one thing, not all DRM list words are included in the WAS database. Also, the incorporation of indirect, mediated associations into WAS values is suboptimal for simulation of free recall. Steyvers et al. (2005) showed that, unlike for recognition and cued recall, direct associations are better predictors of semantic intrusions for free recall than is a combination of direct and indirect associations. Another issue involves the summing of forward and backward association strengths, which could have added noise to our simulations given that backward associations may be more important than forward associations in generating false recall (cf. Brainerd & Wright, 2005; Roediger, Watson, et al., 2001).

Nevertheless, the advantages of WAS seemed to us to outweigh the disadvantages, and we therefore used it as our semantic association metric in this simulation and the next two. We avoided these issues with WAS in our final two simulations by generating an abstract semantic matrix comprising association strength values drawn from abstract distributions rather than being based on real words. Use of the abstract matrix allowed direct manipulation in those simulations of mean association strengths among studied words and between studied words and the critical word, as well as distribution of such strengths across studied words. Of course, there is a cost in ecological validity for using the abstract association values rather than behaviorally determined values, but doing so allowed us to examine the model's operations more clearly.

Lexicon

We used a large-scale lexicon that included many words not presented to participants in the experiment (see Sirotin et al., 2005). Such a lexicon enables fSAM to simulate a variety of semantic effects in recall, including semantically induced intrusions. There were 750 words in the lexicon, consisting of the 225 studied words and 15 critical words from the new experiment's 15 DRM lists (which also included the 12 lists from Kimball & Bjork, 2002), the new experiment's 225 control words, and 285 additional words randomly chosen from the WAS database. Because different list types were presented to different groups of subjects, the control words served as additional unstudied words for Kimball and Bjork (2002), the standard condition, and the mixed condition; the DRM list words served as additional unstudied words for the control condition.

Of the 49 DRM studied words and 38 control words that were not in the WAS database, we used substitutes that varied in tense, number, or part of speech for 27 and 6 words, respectively. For the other 54 experimental words that were not in the WAS database, the strengths of the semantic associations to other lexicon words were set equal to the average WAS value for all experimental and filler words that were included in the WAS database (.02). The average strengths of association among studied words on a particular DRM list (connectivity) and between those words and the applicable critical word were .28 and .49, respectively, for the 12 lists in Kimball and Bjork (2002); these values were .28 and .50, respectively, for the new experiment's 15 lists. The average strengths of association between words on different DRM lists, between DRM list words and critical words for different lists, among control words, and between control words and critical words all ranged between .020 and .026.

Factorial Combination of Semantic Encoding and Retrieval Mechanisms

To address the possibility that our semantic encoding and retrieval mechanisms may interact in their capacity to simulate the data, we tested 16 different versions of fSAM in Simulation 1, comprising the cells in a 4 (semantic encoding mechanism: none, single-item, additive, multiplicative) × 4 (semantic retrieval mechanism: none, single-item, additive, multiplicative) factorial design.

Simulation 1 Results

Goodness-of-Fit Statistics

Set forth in Table 2 are the RMSD values and BIC rank orders for the overall fits for all 16 model versions (Subtable A) and the fits for the four individual conditions that are constituent elements of the overall fit for each model (Kimball & Bjork, 2002, and the standard, mixed, and control conditions from the new experiment in Subtables B, C, D, and E, respectively).

Examining first the overall goodness-of-fit statistics reported in Subtable A of Table 2, the model version that yielded the worst fit overall, perhaps unsurprisingly, was the version that lacked any semantic encoding or retrieval mechanism at all (upper left corner; RMSD = .26). Of the six single-mechanism model versions (the last three cells in the first column and the last three cells in the first row), the best fit was obtained with the multiplicative retrieval mechanism alone (upper right corner; RMSD = .08), followed by the multiplicative encoding mechanism alone (lower left corner;

Table 2
Goodness-of-Fit Statistics Overall and by Experimental Condition in Simulation 1

	Semantic retrieval			
Semantic encoding	None	Single item	Additive rule	Multiplicative rule
A. All conditions combined				
None	0.26_{16}	0.18_{15}	0.13_{10}	0.08_{6}
Single item	0.17_{13}	0.14_{11}	0.08_{7}	0.07_{2}
Additive rule	0.17_{14}	0.15_{12}	0.08_{4}	0.08_{5}
Multiplicative rule	0.11_{9}	0.09_{8}	0.07_{3}	0.06_{1}
B. Kimball and Bjork (2002, Experiment 2)				
None	0.31_{16}	0.22_{11}	0.08_{3}	0.05_{1}
Single item	0.27_{15}	0.23_{13}	0.10_{6}	0.09_{5}
Additive rule	0.27_{14}	0.23_{12}	0.10_{7}	0.12_{9}
Multiplicative rule	0.14_{10}	0.12_{8}	0.08_{4}	0.07_{2}
C. Standard condition				
None	0.30_{16}	0.23_{15}	0.17_{11}	0.10_{8}
Single item	0.19_{13}	0.15_{10}	0.07_{3}	0.06_{1}
Additive rule	0.20_{14}	0.18_{12}	0.07_{4}	0.07_{2}
Multiplicative rule	0.16_{9}	0.09_{7}	0.09_{6}	0.08_{5}
D. Mixed condition				
None	0.21_{16}	0.16_{15}	0.12_{14}	0.08_{13}
Single item	0.07_{10}	0.07_{11}	0.04_{5}	0.06_{9}
Additive rule	0.06_{8}	0.04_{1}	0.05_{7}	0.04_{3}
Multiplicative rule	0.05_{4}	0.07_{12}	0.04_{6}	0.04_{2}
E. Control condition				
None	0.19_{16}	0.08_{13}	0.12_{15}	0.08_{11}
Single item	0.03_{1}	0.05_{4}	0.09_{14}	0.06_{7}
Additive rule	0.06_{5}	0.06_{9}	0.06_{10}	0.05_{6}
Multiplicative rule	0.04_{2}	0.07_{12}	0.06_{8}	0.04_{3}

Note. Each version of the fSAM model was fit to veridical- and false-recall data drawn from Kimball and Bjork (2002, Experiment 2) and our new experiment, using a single parameter set. Goodness of fit is reported as the root-mean-square deviation between behavioral and simulated means. Subscripts indicate the rank ordering of the Bayesian information criterion (BIC) for models within each subtable; see Appendix B for BIC values.

RMSD = .11). Poorer fits were obtained for the versions incorporating only additive retrieval (RMSD = .13), only additive encoding (RMSD = .17), only single-item retrieval (RMSD = .18), or only single-item encoding (RMSD = .17).

Relative to the fits for the single-mechanism model versions, equal or better fits were obtained for the six model versions that combined either of the multiple-item retrieval mechanisms with one of the encoding mechanisms (last three rows in each of the last two columns in Subtable A of Table 2; RMSDs = .06–.08). Despite the addition of a parameter, these six model versions obtained six of the top seven rankings for the BIC, which corrects for differences in number of parameters (see subscripts in Subtable A of Table 2).

Examining Subtables B through E of Table 2, it is clear that the goodness of the overall fits for the six top-fitting model versions was attributable largely to their superior ability to fit the two DRM conditions, Kimball and Bjork (2002) and the standard condition in the new experiment. These model versions obtained six of the top nine fits (RMSDs = .07–.12) for Kimball and Bjork and all six top fits (RMSDs = .06–.09) for the standard condition. The fits of these six model versions for the mixed and control conditions in the new experiment were also quite good (RMSDs = .04–.09).

Another way to examine the results of the fit is by word type, collapsing across the conditions, as set forth in Table 3. All model versions incorporating a semantic mechanism performed quite well in fitting veridical recall (RMSDs = .04–.10). However, for critical word intrusions, there was a fairly sharp division between the good fits for model versions that included one of the multiple-item retrieval mechanisms (last two columns in Subtable B; RMSDs = .03–.07) and the worse fits for those versions that did not (first two columns in Subtable B; RMSDs = .09–.34). Extralist intrusions and prior-list intrusions were also generally better fit by the model versions with one of the multiple-item retrieval mechanisms, although the differences in fits were not as dramatic or as consistent as for critical word intrusions. The most notable exception was for the model version incorporating only additive retrieval, which fit these noncritical intrusions quite poorly.

Two of the dual-mechanism model versions are of particular interest: the one that incorporates both of the additive mechanisms and the one that incorporates both of the multiplicative mechanisms. These model versions are of interest because they each posit similar mechanisms at encoding and retrieval and thus are more parsimonious and plausible than other model versions that posit different mechanisms at encoding and retrieval. The overall fits for these two model versions were quite good, but the fit was better for the combined multiplicative mechanisms (RMSD = .06; bottom right corner of Subtable A of Table 2) than for the combined additive mechanisms (RMSD = .08; third column, third row of Subtable A of Table 2). Another reason to prefer the combined

Table 3
Goodness-of-Fit Statistics by Word Type in Simulation 1

	Semantic retrieval			
Semantic encoding	None	Single item	Additive rule	Multiplicative rule
A. Studied words				
None	0.12_{16}	0.07_{11}	0.05_{2}	0.10_{15}
Single item	0.04_{1}	0.06_{7}	0.06_{8}	0.07_{13}
Additive rule	0.06_{3}	0.06_{6}	0.06_{4}	0.06_{9}
Multiplicative rule	0.06_{5}	0.08_{14}	0.07_{10}	0.07_{12}
B. Critical word				
None	0.34_{16}	0.25_{14}	0.07_{8}	0.05_{4}
Single item	0.24_{12}	0.20_{11}	0.03_{1}	0.04_{3}
Additive rule	0.25_{15}	0.22_{13}	0.04_{2}	0.06_{6}
Multiplicative rule	0.17_{10}	0.09_{9}	0.06_{7}	0.05_{5}
C. Extralist intrusions				
None	0.36_{16}	0.20_{13}	0.26_{15}	0.12_{3}
Single item	0.21_{14}	0.17_{12}	0.16_{10}	0.11_{4}
Additive rule	0.19_{11}	0.16_{9}	0.14_{7}	0.14_{8}
Multiplicative rule	0.11_{2}	0.13_{6}	0.12_{5}	0.08_{1}
D. Prior-list intrusions				
None	0.12_{13}	0.15_{16}	0.12_{15}	0.04_{4}
Single item	0.08_{9}	0.10_{14}	0.04_{6}	0.03_{2}
Additive rule	0.08_{10}	0.07_{11}	0.08_{12}	0.05_{7}
Multiplicative rule	0.04_{3}	0.06_{8}	0.04_{5}	0.02_{1}

Note. Each version of the fSAM model was fit to veridical- and false-recall data drawn from Kimball and Bjork (2002, Experiment 2) and our new experiment, using a single parameter set. Goodness of fit is reported as the root-mean-square deviation between behavioral and simulated means. Subscripts indicate the rank ordering of the Bayesian information criterion (BIC) for models within each subtable; see Appendix B for BIC values.

multiplicative mechanisms to the combined additive mechanisms is that the model versions incorporating the individual multiplicative mechanisms were the best fitting single-mechanism model versions. Note that the combined multiplicative mechanisms fit each word type and condition much the same as, or better than, either of those mechanisms alone.

For the foregoing reasons, in Simulations 2–5, we focused on the model version that combines the multiplicative encoding and retrieval mechanisms, and we refer to this model as the fSAM multiplicative model for the sake of convenience. However, in Appendix C, we report means for veridical and false recall in all four conditions for each model version in Simulation 1.

Fit of Means for Dependent Variables

An examination of the fits for each word type in each condition reveals in more detail the specific elements of the data pattern that the fSAM multiplicative model handled with relative ease or difficulty.

Veridical recall. Figure 1 shows the serial position curves for each of the four conditions in the behavioral data and in the fit of the fSAM multiplicative model. The model captured the general shapes of the curves, as well as almost all of the more local patterns. The exceptions were the overestimation of the primacy effect for Serial Position 1 in the new experiment's conditions and the previously noted overestimation of the recency effect across the last few serial positions in Kimball and Bjork (2002, Experiment 2).

Panel A of Figure 2 shows the mean veridical recall for the four conditions in the behavioral data and in the fit of the fSAM multiplicative model. The model qualitatively captured the ordinal ranking of means in the behavioral data—Kimball and Bjork (2002) > standard > mixed = control—and the absolute deviation from each mean is quite small.

Critical word intrusions. Panel B of Figure 2 shows the mean rate of critical word intrusions in the four conditions for the behavioral data and for the fit of the fSAM multiplicative model. The model fit the intrusion rates well quantitatively and qualitatively. The model also captured the relatively late output percentile for the critical word, as well as the ordinal ranking of percentiles for Kimball and Bjork (2002; behavioral $M = 66$; simulated $M = 78$) and the standard condition (behavioral $M = 63$; simulated $M = 59$).

We explored whether the later simulated output percentile for Kimball and Bjork (2002) was attributable to the assumption by SAM that subjects always initiated immediate free recall with the recency items still in STM at the end of study, an assumption that was inconsistent with the data in Kimball and Bjork for reasons previously mentioned. We adjusted simulated output percentiles for Kimball and Bjork to reflect the initiation of recall with nonrecency items in 63% of the lists, consistent with the behavioral data. The output percentile for the critical word following this adjustment was 65, a much better fit to the data.

Extralist intrusions. Panel C of Figure 2 shows the mean number of extralist intrusions in the four conditions for the behavioral data and for the fit of the fSAM multiplicative model. The qualitative and quantitative fits were excellent for the mixed and control conditions. The fits of extralist intrusions for the DRM conditions were fairly good quantitatively but failed to capture the

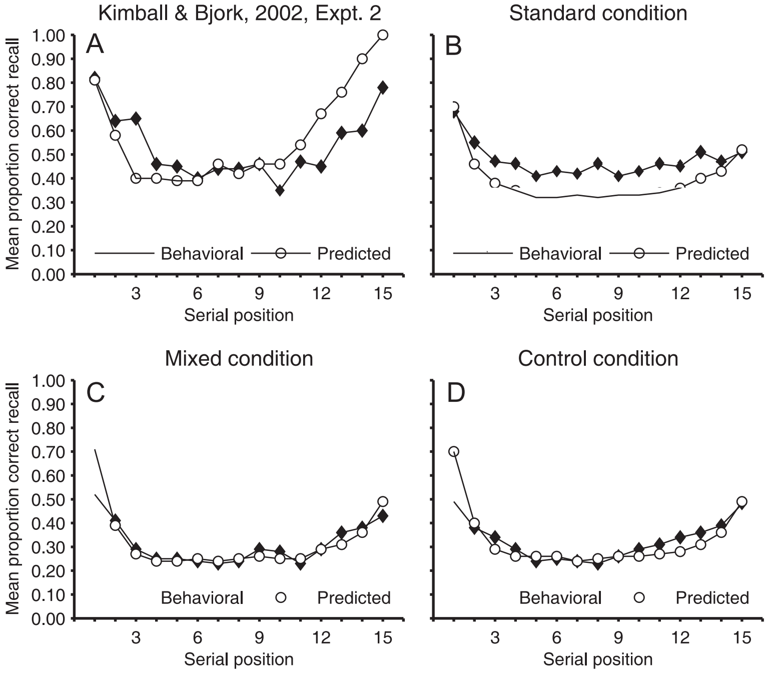

Figure 1. Serial position curves for the fSAM multiplicative model in Simulation 1. Behavioral data are from Kimball and Bjork (2002, Experiment 2) and the three conditions in our new experiment (standard, mixed, and control).

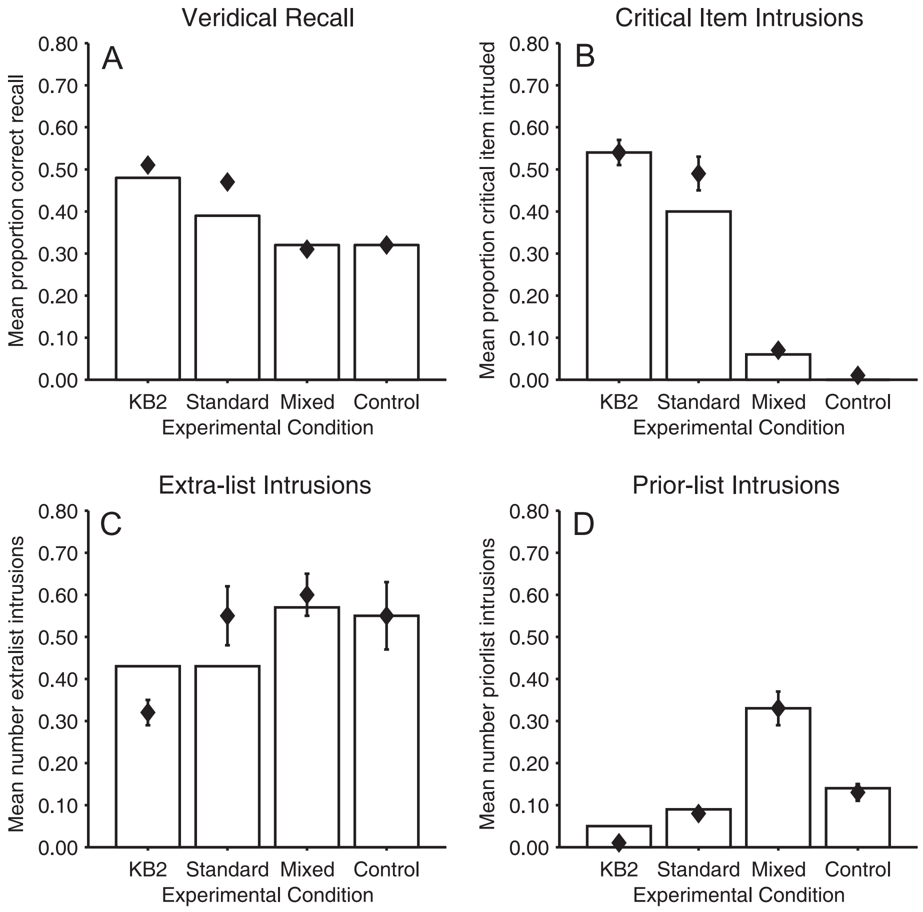

Figure 2. Mean veridical and false recall for the fSAM multiplicative model in Simulation 1. Predicted means are represented by the bars. Behavioral data from Kimball and Bjork (2002, Experiment 2 [KB2]) and the three conditions in our new experiment (standard, mixed, and control) are represented by diamonds with error bars.

qualitative difference in the behavioral rates for the two conditions, instead generating means for both that were midway between the behavioral rates. This was likely attributable to the lack of any mechanism that would allow a difference in extralist intrusions for immediate versus delayed free recall. The only mechanism in SAM-type models that treats immediate and delayed free recall differently is the initiation of recall with the emptying of the STM buffer in immediate recall. We address this point in the General Discussion.

Prior-list intrusions. Panel D of Figure 2 shows the mean number of prior-list intrusions in the four conditions for the behavioral data and for the fit of the fSAM multiplicative model. The fits are excellent both qualitatively and quantitatively.

Summary of the Simulation 1 Results

The multiplicative encoding and retrieval mechanisms were the two best single semantic mechanisms, and the combination of those two mechanisms performed the best overall. The fSAM multiplicative model proved quite capable of fitting the intricate pattern of means in Kimball and Bjork (2002) and the new experiment, fitting all of the dependent variables well across the four conditions, using a single parameter set.

Parameter Values

The parameter values for the fits for all model versions appear in Appendix D. We note that the values for parameters inherited from SAM remained within the ranges for such parameters that have been reported historically for SAM-type models (cf. Gillund & Shiffrin, 1981, 1984; Gronlund & Shiffrin, 1986; Mensink & Raaijmakers, 1988; Raaijmakers, 2003; Raaijmakers & Shiffrin, 1981; Sirotin et al., 2005). Using historical parameter value ranges is not strictly necessary, but it bears mentioning because it means that our model versions did not indirectly change the underlying operation of the basic SAM mechanisms by way of setting parameter values outside of historical bounds. Such an indirect change can thus be ruled out as an alternative explanation for the success of our model.

Simulation 1A: Increased Weighting of Critical Word Intrusions

A risk for theories of false memory in the DRM paradigm is to focus exclusively on predicting critical word intrusions. As we have seen, requiring prediction of other intrusions and veridical recall in addition to critical word intrusions imposes constraints on the generality of any such theory. To demonstrate further the importance of these additional constraints, we conducted a set of fits of the initial data set in which we weighted the fit of critical word intrusions 100 times as much as the fit of veridical recall or other intrusions. We conducted simulations using the nonsemantic model version and the six single-mechanism model versions to examine the ability of each mechanism to fit the data on its own under conditions highly favorable to simulating the critical word intrusion rates. As reported previously, many of these model versions had provided poor overall fits when the fits of critical word intrusions were weighted equally with the fits of veridical recall and noncritical intrusions. The present simulations sought to examine whether the fits of the critical word intrusions would improve for these single-mechanism model versions and, if so, at what cost in terms of worsened fits for the other dependent variables. Evidence of such a trade-off between fitting the critical word intrusions and the other dependent variables as a result of preferentially weighting the former would underscore the importance of modeling veridical recall and noncritical intrusions, not just critical word intrusions. The results of these simulations are set forth in Table 4.

Four aspects of the pattern in Table 4 are noteworthy. First, unsurprisingly, the nonsemantic model version proved utterly incapable of generating critical word intrusions, even when such intrusions were heavily weighted in the fit. Second, somewhat more surprisingly, the single-item retrieval model version also fit the critical word intrusions poorly, despite the heavy weighting of that variable. Clearly, the single-item retrieval mechanism—the mechanism used in the eSAM model (Sirotin et al., 2005)—is not well suited to simulating false recall in the DRM paradigm.

Third, versions incorporating semantic encoding were all able to fit the critical word intrusions well. If all we were concerned about were simulating critical word intrusions, any of these semantic encoding model versions would suffice. However, the good fits of the critical word intrusions for these model versions were achieved at the expense of quite poor fits of extralist intrusions, prior-list intrusions, and even veridical recall. The results for these versions most clearly illustrate the need to test the capacity of theories of false memory to predict not just critical word intrusions but also veridical recall, extralist intrusions, and prior-list intrusions.

Finally, the versions incorporating a multiple-item retrieval mechanism—especially the multiplicative retrieval mechanism—not only fit the critical word intrusions exceedingly well but were able to do so while also providing reasonably good fits for most of the other dependent variables. That the fits of these other dependent variables were so good notwithstanding the extremely low weights accorded to them in the fitting procedure suggests that the multiple-item retrieval mechanisms are not subject to the same trade-offs as are the other mechanisms and thus are more general and more powerful as explanations of human memory processes.

We have thus far demonstrated the ability of the fSAM multiplicative model to fit the data from Kimball and Bjork (2002) and the new experiment. The remaining four simulations tested the generality of the model by assessing its capacity to simulate other false-recall effects reported in the literature, including, in particular, developmental data (Simulation 2), specific list effects (Simulation 3), and the effects of backward association strength, forward association strength, connectivity, and number of studied associates of the critical word (Simulations 4 and 5).

Simulation 2: Developmental Data

In this simulation, we sought to test the capacity of fSAM to simulate veridical and false recall in children. We briefly review the major findings in the literature regarding the patterns of veridical and false recall as children age. We then describe how the fSAM multiplicative model can account for these developmental changes and apply the model to data from Brainerd et al. (2002), a key study of false recall by children in the DRM paradigm.

Free Recall and the DRM False-Memory Effect in Children

The DRM paradigm has been used extensively to test recognition memory in both children and adults (for reviews, see Brainerd & Reyna, 2005; Gallo, 2006). The DRM paradigm has been used less frequently to examine free recall in children, and the number of empirical studies that have reported data in sufficient detail to provide a good set of constraints for computational models of free recall is quite small. Nevertheless, these studies show important changes in the pattern of veridical and false recall as children develop.

Table 4
Goodness-of-Fit Statistics for Simulation 1A, Using Increased Weighting of Critical Word Intrusions

			Word type			
Semantic encoding	Semantic retrieval	Overall	Studied words	Critical word	Extralist intrusions	Prior-list intrusions
None	None	0.26	0.13	0.33	0.38	0.12
None	Single-item	0.24	0.14	0.23	0.28	0.33
None	Additive	0.17	0.20	0.02	0.31	0.06
None	Multiplicative	0.13	0.18	0.02	0.17	0.12
Single-item	None	0.63	0.41	0.08	1.39	0.32
Additive	None	0.58	0.16	0.10	1.37	0.32
Multiplicative	None	0.25	0.19	0.05	0.51	0.20

Note. Each model version was fit to veridical- and false-recall data drawn from Kimball and Bjork (2002, Experiment 2) and our new experiment, using a single parameter set. Goodness of fit is reported as the root-mean-square deviation between the mean values observed in the behavioral data for each dependent variable and the mean values for those variables calculated across sets of simulated subjects.

Developmental Changes in Veridical Recall

In general, children show reduced levels of veridical recall compared with adults, with the level of recall increasing almost linearly with age until it levels off during early adolescence (Moely, 1977; Schneider, 2002; Schneider & Pressley, 1997). This general pattern has been found with a wide variety of stimuli, including names of pictured objects (Cole, Frankel, & Sharp, 1971), lists of unrelated words (Hall & Tinzmann, 1985), categorized lists (Howe, 2006), and DRM lists (Brainerd et al., 2002; Dewhurst & Robinson, 2004; Howe, 2006). A number of theoretical explanations have been offered for this pattern (for reviews, see Schneider, 2002; Schneider & Pressley, 1997), but the explanations all agree on several basic points. First, there is an overall increase in processing efficiency and effective memory span between the ages of 5 and 14 years old that allows older children and adults to take advantage of elaborative and cumulative rehearsal strategies (Schneider & Pressley, 1997). Second, children begin to use these rehearsal strategies spontaneously between ages 8 and 11 (Schneider, 2002). Third, because adults and older children have a larger knowledge base, they are able to make more effective use of preexisting semantic associations during encoding and retrieval (Schneider & Pressley, 1997).

Developmental Changes in Critical Word Intrusion Rates

Like veridical recall, false recall of critical words for DRM lists typically increases with age. Brainerd et al. (2002) found that 5- and 7-year-olds generally failed to show the DRM false-memory effect for free recall, with critical word intrusion rates near the floor (*M*s = .05–.11 in their three experiments), whereas 11-year-olds intruded the critical word considerably more often ($M = .23$) than did their younger counterparts. Brainerd et al. reported that Price, Metzger, Williams, Phelps, and Phelps (2001) found similar patterns in an unpublished study. Howe (2006) also found a similar pattern of false recall for 5-, 7-, and 11-year-olds, although the levels of critical word intrusions were somewhat higher than in Brainerd et al.

Developmental Changes in the Rates of Other Intrusions

Brainerd et al. (2002) reported that, unlike veridical recall and critical word intrusions, intrusions of other nonstudied but semantically related items (extralist intrusions) and intrusions of items from previously studied lists (prior-list intrusions) did not increase with age. The extralist intrusion rate, measured as the mean proportion of lists for which there was at least one such intrusion, did not reliably differ between 5-, 7-, and 11-year-olds in any of their experiments. The mean proportion of lists with at least one prior-list intrusion actually decreased with age, with 5- and 7-year-olds intruding words from previous lists more often than they intruded either the critical word or semantically related extralist intrusions and 11-year-olds recalling almost no words from previous lists.

Theoretical Accounts of Developmental Changes

The most complete theoretical account of these patterns to date is based on fuzzy trace theory (for reviews, see Brainerd & Reyna, 2004, 2005). To explain the developmental pattern for veridical recall, the theory assumes that verbatim storage and retrieval mechanisms develop relatively rapidly between early childhood and adolescence. On the other hand, to explain the developmental pattern for critical word intrusions, fuzzy trace theory assumes that gist storage and retrieval mechanisms have not yet fully developed among young children, impairing the ability to extract meaning from events (Brainerd et al., 2002). The theory is thus able to account for the finding by Dewhurst and Robinson (2004) that the nature of intrusions becomes increasingly semantic and decreasingly phonological as children mature from age 5 to age 11. To explain the higher rate of prior-list intrusions by younger children, fuzzy trace theory assumes that the weaker gist traces provide less competition for erroneous verbatim traces from previous episodes.

A second explanation is based on the activation–monitoring account. In this account, the DRM effect is driven primarily by associative relations rather than by thematic relations (Howe, 2006).[2] This theory attributes the pattern of changes in veridical-recall and critical word intrusion rates to developmental increases in the number and strength of associative relations in LTM (Bjorklund, 1987) and to corresponding increases in the automaticity with which these associations are activated (Bjorklund & Jacobs, 1985). These developmental changes are theorized to lead to increased activation of the studied items, promoting an increase in veridical recall, which in turn leads to increased activation—and therefore increased intrusions—of critical words (Howe, 2006). To date, no activation-based explanation has been offered for the developmental pattern of extralist and prior-list intrusion rates.

A third possibility is that the developmental changes in veridical and false recall are both driven by changes in rehearsal style. During encoding, young children tend to use a single-item rehearsal strategy (Ornstein, Naus, & Liberty, 1975), thus reducing the amount of relational processing during encoding and, consequently, the rates of both veridical recall and critical word intrusions. Children typically begin to spontaneously use cumulative rehearsal strategies between the ages of 8 and 11 (Dempster, 1981), the same age range in which veridical-recall and critical word intrusion rates dramatically increase.

The developmental changes in the pattern of veridical and false recall might also be driven by changes in the efficiency of retrieval and corresponding changes in retrieval strategies. Very young children often fail to make use of retrieval cues even when they are provided by the experimenter (Ritter, Kaprove, & Fitch, 1973). Around age 5 or 6, most children spontaneously begin using simple retrieval cues such as external objects or the last retrieved item; by age 10 or 11, most children have begun using more sophisticated retrieval strategies such as using category labels and multiple previously retrieved items as retrieval cues (Kobasigawa, 1977). There is also a developmental change in how exhaustively children search memory during a retrieval attempt, with children often giving up much earlier than adults (Kobasigawa, 1977).

In this simulation, we simulated the data for 5-, 7-, and 11-year-olds from Brainerd et al. (2002), which included means for all four word types of interest: veridical recall, critical word intrusions,

[2] Howe (2006) used the term *semantic* somewhat differently than we do. Howe defined semantic relations as “thematic or categorical relations between concepts” (Howe, 2006, p. 1112). Our semantic matrix is a reflection of generalized prior experiences and maps directly to his definition of interitem associative relations.

extralist intrusions, and prior-list intrusions. Because there is ample evidence of many changes in processing with development, we allowed the values of multiple parameters to vary across age groups, as described below. In doing so, we constrained the parameters to vary only in theoretically plausible ways in light of the above-described developmental changes in processing.

Simulating Developmental Changes With fSAM

Changes in Processing Efficiency

In the fSAM model, increases in general processing efficiency during encoding can be implemented by increasing the values of the contextual and episodic incrementing parameters during encoding (a and b_1). Given that such processing efficiency increases with age, it seems plausible to set these parameters to relatively low values for young children and increase them monotonically with development. Similarly, the general increase in retrieval efficiency with age can be simulated by monotonically increasing the retrieval weight parameters (W_c and W_e) and the output encoding parameters (e and f_1) with development. Also, to reflect the developmental increase in the exhaustiveness of memory search, the maximum number of retrieval failures prior to recall termination, K_{max}, can increase with age.

Changes in Degree of Semantic Processing

The increase in semantic processing that is known to occur during childhood can be implemented by increases in the value of the semantic retrieval weight parameter, W_s, and increases in the value of the semantic encoding scaling parameter, a_s.

Changes in Rehearsal and Retrieval Strategies

Changes in memory span and the set size for cumulative rehearsal can be implemented by varying parameter r, the STM buffer size. For example, setting $r = 1$ in the model implements single-item rehearsal and limits the set size for the retrieval cue to one item. Letting r increase with age simulates the increased use of cumulative rehearsal strategies and the use of multiple items to cue retrieval.

Changes in the Conservation of Contextual Strength

The literature on source monitoring in children shows that younger children do not distinguish information from different episodes as well as adults and older children do (Schneider & Pressley, 1997). We implemented this effect by allowing the amount of contextual association strength that is preserved from one list to another—that is, the value of parameter ρ—to decrease with age.

Simulation Method

Experimental Procedure Simulated

We simulated the data for the three age groups from Experiment 2 (5- and 7-year-olds) and Experiment 3 (5- and 11-year-olds) in Brainerd et al. (2002). In these experiments, a total of 16 DRM lists taken from Stadler et al. (1999) were auditorily presented to each participant at a rate of 2 s per word. The order of the lists was randomized for each participant, but the words in each list were presented in the same highest-to-lowest associate order that was used in Stadler et al. One min was allowed for immediate free recall following presentation of each list.

We simulated the combined recall data from the two experiments for the 16 lists, averaging the means for the 5-year-olds. We did not simulate the results for a group of adults in Experiment 3 because the reported rate of extralist intrusions (M = .03 lists with at least one extralist intrusion) was anomalously low compared with other results reported in the literature, including Kimball and Bjork (2002; M = .32 extralist intrusions per list), McDermott (1996; M = .22 and .32 extralist intrusions per list for immediate and delayed recall, respectively), Roediger and McDermott (1995; M = .14 lists with at least one extralist intrusion), and our new experiment (M = .55 extralist intrusions per list).

Fitting Procedure

We used the best fitting parameter set for the fSAM multiplicative model from Simulation 1 as a base parameter set and then used a genetic algorithm to find values for the parameters for each age group. The parameter values for each age group were determined by multiplying the parameters from the next highest age group by an algorithmically determined scaling factor. So, the parameter values for the 11-year-old group were determined by applying a set of scaling factors to the adult parameters from Simulation 1 and so on. The exception was that we fixed the value of r at 1 for the 5-year-old group to reflect the fact that individuals in this age group almost exclusively use a single-item rehearsal strategy. The scaling factors ensured monotonicity in the change of parameters across age groups, consistent with the theoretical constraints mentioned above. As in Simulation 1, we ran the best fitting parameter sets from the genetic fit with 200 simulated subjects each. The values we report are from the large-sample run that provided the best quantitative fit to the data.

Dependent Variables

For each age group, we fit veridical recall and the mean numbers of lists for which the critical word was intruded, for which there was at least one extralist intrusion, and for which there was at least one prior-list intrusion. To determine goodness of fit, we calculated RMSD, weighting each of the dependent variables for all three groups equally.

Lexicon

We used a lexicon of 779 words based on the 55 DRM lists used by Roediger, Watson, et al. (2001) that included the 16 lists used by Brainerd et al. (2002). The lexicon consisted of the 15 studied words and the critical word from each of the 55 lists, net of 101 words that were duplicated across lists. Of the remaining 779 words, 103 were not in the WAS database. We used substitutes that varied in tense, number, or part of speech for 32 of those 103 words. For the other 71 words that were not in the WAS database, the strengths of the semantic associations to other lexicon words were set equal to the average WAS value for all words that were included in the WAS database (.02). The average strengths of association among words on a particular DRM list and between those words and the applicable critical word were .21 and .41, respectively, across all 55 lists; these values were .26 and .48, respectively, for the 16 lists used by Brainerd et al. and as studied lists in our simulation. The average strengths of association between words on different DRM lists, between DRM list words and critical words for different lists, and among all words in the lexicon were all equal to .02.

Simulation Results

The results of our fits to the developmental data from Brainerd et al. (2002) are shown in Table 5. The fit is excellent, both quantitatively (RMSD = .04) and qualitatively. The fSAM model fit the developmental pattern for each of the dependent variables. Both veridical-recall and critical word intrusion rates increased with age. Despite these increases, the model generated extralist intrusions at the same rate for all three age groups. Finally, the model generated high levels of prior-list intrusions for the 5-year-olds, with a decrease in the rate as age increased. The specific parameters for each age group are set forth in Appendix D, reflecting the constraints we mentioned previously.

Discussion

The simulation results show that fSAM can account for the pattern of developmental changes in veridical and false recall using parameter sets that reflect the changes in cognitive processes that are known to occur as children develop. Our goal for the present article is to demonstrate that fSAM can do so when all developmental changes hypothesized to impact model operations are implemented simultaneously. Further work could elucidate which of these changes, alone or in combination, are necessary or sufficient to simulate the developmental pattern, or particular aspects of it.

Simulation 3: List Effects in Roediger, Watson, McDermott, and Gallo (2001)

Another set of results from the DRM literature pertains to the relative rates of veridical recall and critical word intrusions for specific DRM lists based on different critical words, as well as the correlation between those rates—that is, the true–false correlation. In Simulation 3, we simulated the patterns of these effects for the 55 DRM lists incorporated in the multiple regression analysis reported by Roediger, Watson, et al. (2001).

Table 5
Mean Veridical and False Recall for the Fit to Brainerd, Reyna, and Forrest (2002) in Simulation 2

Word type	Age (years)	fSAM	Behavioral
Studied words	5	0.23	0.25
	7	0.31	0.38
	11	0.41	0.41
Critical word intrusions	5	0.09	0.07
	7	0.14	0.09
	11	0.23	0.23
Extralist intrusions	5	0.09	0.13
	7	0.09	0.12
	11	0.09	0.13
Prior-list intrusions	5	0.21	0.21
	7	0.12	0.18
	11	0.04	0.02

The behavioral mean rates of veridical recall and critical word intrusions for the 55 DRM lists used in Roediger, Watson, et al. (2001) came from two other studies. Stadler et al. (1999) reported veridical-recall and critical word intrusion rates for 36 lists, including the 24 lists originally used by Roediger and McDermott (1995) and 12 new lists. Gallo and Roediger (2002) reported rates for 28 lists, including nine that had been reported by Stadler et al. and 19 new lists. Our goal in this simulation was to simulate the list effects reported in these studies. However, we discovered that simulating these list effects was less straightforward than it might appear at first blush because of disagreement in the literature as to the veridical-recall and critical word intrusion rates for particular lists and as to the correlation between those two rates across sets of lists. We next describe such disagreement in more detail.

Differences in Critical Word Intrusion Rates

One example of the lack of agreement in the literature as to critical word intrusion rates involves the 10 lists that were used in each of four studies: Stadler et al. (1999), Deese (1959b), Kimball and Bjork (2002, Experiment 2), and the new experiment. The critical word intrusion rates reported in Stadler et al. correlated poorly with the rates reported in the other three studies (*r*s = .18, .13, and .08, respectively). There was greater, but not outstanding, agreement among critical word intrusion rates in the latter three studies, ranging from .60 to .71.

There were also differences in the mean levels of critical word intrusions for the same lists across different studies. For the 17 lists shared by Deese (1959b) and Stadler et al. (1999), the critical word intrusion rate was lower for Deese than for Stadler et al. (*M* difference = .11). The same was true for the 10 lists shared by Deese and Kimball and Bjork (2002; *M* difference = .19) and for the 11 lists shared by Deese and the new experiment (*M* difference = .13). One possibility is that these differences were due in part to Deese's use of lists with only 12 associates rather than the 15 associates used in the other studies (see Robinson & Roediger, 1997). However, Gallo and Roediger (2002) used 15 associates in their lists, and the critical word intrusion rates in that study differed very little from the rates reported by Deese for the subset of 10 lists that were used in both studies (*M*s = .19 and .17, respectively). There was also a substantial difference in mean critical word intrusion rates for the subset of nine lists used in both Stadler et al. and Gallo and Roediger (*M*s = .41 and .27, respectively).

Differences in Veridical-Recall Rates

Veridical-recall rates for particular lists in the new experiment correlated poorly with those for sets of lists that experiment shared with Stadler et al. (1999), Gallo and Roediger (2002), and Kimball and Bjork (2002), *r*s = .41, .53, and .42, respectively. However, the veridical-recall rates among the latter three studies correlated well, ranging from .90 to .97. Deese (1959b) did not report veridical-recall rates.

Differences in True–False Correlations

The true–false correlation—relating mean veridical-recall and critical word intrusion rates across lists—has also varied across studies. For example, true–false correlations were −.54, −.24, −.43, .08, and .28 for Gallo and Roediger (2002); Stadler et al. (1999), Roediger, Watson, et al. (2001); Kimball and Bjork (2002); and the new experiment, respectively. All of the true–false correlations reported in Brainerd et al. (2003) were also positive, ranging from .02 to .36. Some of this variation is surely attributable to the use of different lists across the studies. However, there is not always agreement across studies even for the same lists. For example, there is a substantial difference in the true–false correlation obtained by Kimball and Bjork across the 12 lists they used ($r = .08$) and that obtained by Stadler et al. across the same 12 lists ($r = .46$).

Moreover, true–false correlations can be substantially affected by the range of critical word intrusion rates represented among the lists used in a particular study. It matters a good deal whether the study includes lists with high, low, or both high and low critical word intrusion rates. For a number of studies, the true–false correlation was likely affected by the inclusion of the same two lists with the lowest critical word intrusion rates—those with the critical words *fruit* ($M = .20$) and *king* ($M = .10$). This can be seen rather dramatically for the 24 lists used in Roediger and McDermott (1995). The true–false correlation for those lists ($r = -.23$) actually becomes slightly positive ($r = .05$) with the exclusion of the *fruit* and *king* lists, based on the means in Stadler et al. (1999). Similarly, the overall true–false correlation was −.24 for the entire set of 36 lists included in Stadler et al. and −.28 for the 16 lists used in Brainerd, Forrest, Karibian, and Reyna (2006), but these correlations were only −.09 and .03, respectively, with the exclusion of the *fruit* and *king* lists, based on the means from Stadler et al. The fragility of the negative true–false correlation in these last two studies is particularly important because they have been offered as evidence of a globally negative true–false correlation when pooling across lists and participants, respectively (see Brainerd et al., 2006).

A similar pattern can be observed for the set of all 55 lists in Roediger, Watson, et al. (2001): The overall true–false correlation, −.43, becomes only −.07 with the exclusion of the 19 lists with critical word intrusion rates at or below $M = .20$. The true–false correlation actually becomes positive ($r = .19$) for the 18 lists reported in Roediger, Watson, et al. (taken from Stadler et al., 1999) with critical word intrusion rates above $M = .40$. This last result is consistent with the positive true–false correlations found in Kimball and Bjork (2002) and in the new experiment given that those experiments used lists selected from among those 18 high-intrusion lists. In Simulation 5, we discuss model predictions that offer an explanation for this shift in true–false correlations from negative to positive as critical word intrusion rates increase.

Methodological Differences

Differences in method could account for some of these differences in true- and false-recall rates and correlations. Examples of methodological differences that might be expected to matter in this connection include the following: differences in the number of lists presented to each subject; differences in the average backward association strength across the entire set of lists presented to a subject; differences in the order of lists, including constant versus random orders across subjects, and for constant list orders, differences in the list sequence position for particular lists; differences in the order of words within lists, including constant versus random orders across subjects; differences in presentation rate; and differences in retention interval.

Although the effects of such methodological differences may merit exploration in controlled studies and although some have already been explored, their effects on critical word intrusions and veridical recall for particular DRM lists have not been examined as yet. In addition, although some of the differences could be incorporated into our simulations, some of the differences were not sufficiently well specified for that purpose. For example, Stadler et al. (1999) failed to disclose the list sequence position of particular lists, and Deese (1959b) failed to report the presentation rate.

Simulation Method

Accordingly, rather than attempting to incorporate a priori all the potential methodological differences into our simulation of list effects, we opted to simulate data for the 55 lists reported in Roediger, Watson, et al. (2001) by separately fitting the data contributed to that study by Stadler et al. (1999) and Gallo and Roediger (2002). We assumed that the methodological differences across these studies would influence parameter values obtained in the fits. We simulated this set of results because they were obtained in other labs, thus allowing us to test further the generalizability of our model's predictions, and because the authors reported mean critical word intrusion rates and veridical-recall rates for all lists. (Deese, 1959b, did not report veridical-recall rates; Deese, 1959a, did not report critical word intrusions and noncritical extralist intrusions separately). As in the fit of the developmental data in Simulation 2, we used the fSAM multiplicative model in this fit.

Fitting Procedure

We used the same fitting procedure as in Simulation 1 and separately fit the sets of 36 lists drawn from Stadler et al. (1999) and 19 unique lists drawn from Gallo and Roediger (2002). In the simulation, words were presented in the same order within each list as in those two studies. The simulation implemented the different presentation rates in those two studies—2 s per word for Stadler et al. and 1.5 s per word for Gallo and Roediger—by multiplying the encoding incrementing parameters a and b_1 by factors of 2 and 1.5, respectively. The studies did not report list presentation orders, so we randomized those orders.

Dependent Variables

We equally weighted the following dependent variables in our fits, with the values for each variable being calculated separately across the two sets of lists: (a) the mean veridical-recall rate in the behavioral data, collapsing across lists; (b) the correlation between veridical-recall rates in the behavioral and simulated data across lists; (c) the mean critical word intrusion rate in the behavioral data, collapsing across lists; (d) the correlation between critical word intrusion rates in the behavioral and simulated data across lists; and (e) the correlation in the behavioral data between the critical word intrusion rate and the veridical-recall rate.

Lexicon

This simulation used the same lexicon as did Simulation 2, comprising the words used in all 55 DRM lists used in Roediger, Watson, et al. (2001). The average WAS strengths of association among words on a particular DRM list and between those words and the applicable critical word were .25 and .46, respectively, for the 36 lists taken from Stadler et al. (1999); these values were .12 and .31, respectively, for the 19 lists taken from Gallo and Roediger (2002).

Simulation Results

Veridical Recall

As shown in Table 6, mean veridical-recall rates, collapsing across lists, were fit well both quantitatively and qualitatively. The correlations between behavioral and simulated veridical-recall means, calculated across individual lists, were not as high as we had hoped, but they were within the range of correlations for mean veridical recall among different behavioral studies.

Critical Word Intrusions

Mean critical word intrusion rates, collapsing across lists, were fit well qualitatively in that the rate was somewhat higher for the Stadler et al. (1999) lists than for the Gallo and Roediger (2002) lists, as was true in the behavioral data. Quantitatively, the fit for the Gallo and Roediger mean rate of critical word intrusions was fairly good, but the mean simulated critical word intrusion rate for the Stadler et al. lists was quite a bit lower than in the behavioral data. The correlation between behavioral and simulated critical word intrusion means, calculated across individual lists, was somewhat low. However, as with veridical recall, this correlation was comparable to correlations for mean critical word intrusion rates among different behavioral studies.

True–False Correlations

The true–false correlations were also fit well qualitatively in that the simulated correlations were negative for both the Gallo and Roediger (2002) and Stadler et al. (1999) list sets, with the correlation for the former being substantially more negative than that for the latter, as in the behavioral data. Quantitatively, the absolute values of the two correlations were not as great as in the behavioral data, but they were well within the range of true–false correlation values reported in the literature.

Parameter Values

The parameter values for the fits are set forth in Appendix D. Most of the variation in parameter values across experiments might reflect differences in methodology between the experiments. However, we note that the retrieval weight parameter, W_s, is more than an order of magnitude lower for the fit of Gallo and Roediger (2002) than for the fit of Stadler et al. (1999). This parameter difference has the effect of increasing competition for the critical word in the Gallo and Roediger fit and thus reducing critical word intrusions. Even with that parameter difference, the critical word intrusion rate was higher in the simulation than in the behavioral data. To the extent that this parameter adjustment was needed to keep the critical word intrusion rate low, it is telling that the model was unable to accomplish that goal solely on the basis of differences in association strengths for the sets of lists used in the two experiments, rather than by adjusting parameter values. We address possible explanations for this inability in the *Discussion* section.

Another set of parameter differences between the two fits involves the semantic encoding parameter, a_s; the contextual decay parameter, ρ; and the contextual retrieval weight parameter, W_c. These parameters involve incrementing, preserving, and cuing with contextual association strengths, respectively. Compared with the fit of Stadler et al. (1999), in the fit of Gallo and Roediger (2002), semantically related contextual strength increments were higher, as was the proportion of contextual association strength preserved across lists, but contextual association strengths received less weight at retrieval. Thus, these parameter differences tended to cancel each other out, such that the net effect of these differences seems likely to be minimal.

Discussion

The fSAM multiplicative model provided fairly good qualitative fits and, in some cases, good quantitative fits to the patterns of veridical-recall and critical word intrusion rates across the individual lists reported in Gallo and Roediger (2002) and Stadler et al. (1999), as well as true–false correlations calculated across the lists in those studies. That the quantitative fits were less than outstanding in many cases was likely due in large part to two causes, which we have discussed previously. The first of these causes is the intrinsic variability suggested by disagreement among the reported

Table 6
Mean Veridical Recall, Mean False Recall, and Correlations for Fits of List Effects in Simulation 3

		Veridical recall		Critical word intrusions		
Study/model simulation	*N*	*M*	*r*	*M*	*r*	True–false *r*
Gallo & Roediger (2002)	19	0.63		0.10		−.61
fSAM	19	0.61	.59	0.18	.24	−.33
Stadler, Roediger, & McDermott (1999)	36	0.60		0.40		−.24
fSAM	36	0.63	.48	0.23	.22	−.13

Note. The fSAM multiplicative model was fit separately to the data for the 19 lists from Gallo & Roediger (2002) and the 36 lists from Stadler, Roediger, & McDermott (1999). r = Pearson correlation coefficient.

statistics in the literature for individual lists and sets of lists, perhaps due in substantial part to methodological differences.

The second likely cause relates to the use of WAS values as our metric of semantic association strengths and, in particular, its incorporation of indirect associations rather than direct associations alone. As noted previously, use of indirect associations dilutes the influence of direct associations, which Steyvers et al. (2005) determined were most important in simulating free recall. Presumably in part because of this dilution, WAS values were only moderately correlated with three key measures: (a) the levels of backward association strength (i.e., the propensity of a word to elicit the critical word in a word association task) that Roediger, Watson, et al. (2001) reported for all studied words in the 55 lists they used ($r = .58$); (b) the mean backward association strength per list for those 55 lists ($r = .73$); and (c) the mean critical word intrusion rate per list for those 55 lists ($r = .44$). These last two correlations are particularly telling in that Roediger, Watson, et al. reported a correlation of .73 between the mean critical word intrusion rates and the mean backward association strengths for their 55 lists. Mean WAS strength per list is thus not as strongly predictive of critical word intrusion rates as is mean backward association strength, presumably in large part because the correlation between those two mean strength measures is somewhat less than perfect.

The dilution of the effect of direct association strengths in WAS values can also help explain the fitting algorithm's selection of such a low value for the semantic retrieval weight parameter, W_s, for the 19 lists taken from Gallo and Roediger (2002). The dilution is particularly apparent for those lists in that the mean WAS value for the studied words on those lists ($M = .31$) was approximately 15 times as large as the mean backward association strength ($M = .02$). By comparison, for the 36 lists taken from Stadler et al. (1999), the mean WAS value for the studied words ($M = .46$) was only approximately 2.5 times as large as the mean backward association strength ($M = .18$). Using WAS as the semantic association metric thus made it relatively more difficult to produce the low levels of critical word intrusions for the Gallo and Roediger lists because the WAS values more substantially overestimated the mean backward association strength for those lists. The fitting algorithm may have adopted the much lower value for W_s to compensate for this greater difficulty in producing lower levels of critical word intrusions for those lists.

Nevertheless, WAS appears to be the best available metric of semantic association strength for our purposes. None of the other metrics available as yet provides a measure that is based on word production, provides values for all word pairs in a large lexicon, and discriminates as well between semantically related and unrelated words. Accordingly, for us to be able to simulate the effects of direct association strengths at this time unconfounded by the influences of indirect association strengths, we needed values for the semantic matrix other than those provided by one of the available metrics. Ideally, we wanted association strength values that we could control and vary. To that end, we decided to conduct simulations using association strength values drawn from abstract distributions of values, rather than basing them on one of the available metrics for real words. We describe these simulations next.

Simulation 4: Effects of Mean Forward, Backward, and Intralist Association Strength

In this simulation, we used the fSAM multiplicative model to generate predictions regarding the effects on veridical and false recall of variations in the mean levels per list of three types of association strength: forward association strength (the propensity of the critical word to elicit one of the studied words in a word association task), backward association strength (the propensity of a studied word to elicit the critical word), and intralist association strength, or connectivity (the propensity of a studied word to elicit another of the studied words). To do so, we used abstract semantic matrices comprising hypothetical pairwise association strengths, which allowed us to construct lists that varied along these dimensions factorially. Using such abstract semantic matrices allowed us to use direct pairwise association strengths, rather than strengths that also reflected the influence of indirect, mediated associations, as with WAS values.

Using such matrices also allowed us to remove any confounding effects of other factors on veridical and false recall. Roediger, Watson, et al. (2001) considered several potential predictors of critical word intrusions other than backward association strength and veridical-recall rates. They found that, unlike those two factors, the other factors did not explain a significant proportion of the variance. However, together, such variables could influence the critical word intrusion rate. In addition, other predictors not considered by Roediger, Watson, et al. might also exert an influence. For example, Madigan and Neuse (2004) recently reported that the length of the critical word relative to the studied words on a list influences critical word recognition. This finding contrasts with Roediger, Watson, et al.'s finding that critical word length relative to other critical words—rather than to the studied words on the list—did not exert such an influence. No such factors were able to influence our results in these simulations inasmuch as the only factors that varied were the three types of association strength that we manipulated.

Previously Reported Effects of Mean Association Strengths

Deese (1959a); Roediger, Watson, et al. (2001); and McEvoy, Nelson, and Komatsu (1999) all reported a positive relationship between backward association strength and critical word intrusions. The empirical findings are less clear as to the effects of connectivity and forward association strength on critical word intrusions. Roediger, Watson, et al. (2001) failed to find any significant correlation between connectivity and critical word intrusions, in contrast to McEvoy et al. and Deese, who found a negative association between connectivity and critical word intrusions. Roediger, Watson, et al. also did not find any significant correlation between forward association strength and critical word intrusions. Brainerd and Wright (2005), using lists with a broader range of forward association strengths than in the lists Roediger, Watson, et al. had used, reported a positive association between forward association strength and false memory for the critical word, at least when testing recognition memory. No study using free recall with a similarly broad range of forward association strengths has been reported to date.

The unsettled state of the literature on the effects of connectivity and forward association strength would usually argue against simulation of such effects inasmuch as it is unclear which of the conflicting findings should be simulated. Nevertheless, these effects, together with those of backward association strength, are central to theories of false memory, perhaps especially to our theory, given that it is based explicitly on strengths of associations between pairs of words. Accordingly, we thought it important to generate model predictions regarding these effects using the unconfounded association strengths in the abstract semantic matrices. Doing so also enabled us to compare the results of these simulations with the list effects involved in Simulation 3, particularly true–false correlations.

Simulation Method

Semantic Matrix

We created an abstract semantic matrix comprising 125 DRM lists of 15 words each by factorially varying mean forward association strength, mean backward association strength, and mean connectivity per list. There were five levels of mean values for each of these association types: .10, .30, .50, .70, and .90. Each of the 125 lists represented a unique combination of one of the five levels for each of the three association strengths. Note that these association strength values should be regarded as scaled similarly to WAS strengths rather than to raw normed probabilities because, for this simulation, we used the best fitting parameters from Simulation 1, which were determined using WAS values. Also note that the connectivity values for this simulation take into account the differing levels of strength for particular intralist associations and are therefore more finely grained than the connectivity values used by Roediger, Watson, et al. (2001) and McEvoy et al. (1999), who treated each such association the same regardless of its strength.

For each list, we created a 15 × 15 connectivity matrix with values drawn from a normal distribution with the specified mean connectivity strength for that list and a standard deviation of .025. To represent the associations to and from the critical word, we added an additional row and column to the connectivity matrix comprising values drawn from a normal distribution with the specified mean forward and backward association strengths for that list and a standard deviation of .025, thus yielding a 16 × 16 list matrix. The 125 list matrices were then placed in the full 2,000 × 2,000 semantic matrix. Association strengths between words from different list matrices were assigned residual values drawn from a normal distribution with a mean of .025 and a standard deviation of .00625.

Simulation Procedure

The fSAM multiplicative model was run with 200 simulated subjects using the parameters from the best fit for that model in Simulation 1, with each simulated subject being presented with all 125 lists. The presentation order of the lists was randomly determined for each simulated subject. To minimize serial position effects, the order of words within each list was randomized, and delayed free recall was used.

Simulation Results

Table 7 shows the correlations across the 125 list means for each combination of word type and association type.

Table 7
Correlations Between Association Strengths and Simulated Means for Studied Words and Intrusions in Simulation 4

	Type of association strength		
Word type	Forward	Backward	Connectivity
Studied words	.01	−.07	.98
Critical word intrusions	.00	.73	−.47
Extra-list intrusions	−.01	−.08	−.73
Prior-list intrusions	.00	−.09	−.74

Veridical Recall

Veridical recall was uncorrelated with mean forward association strength and only weakly and negatively correlated with backward association strength. However, veridical recall was strongly and positively correlated with connectivity.

Critical Word Intrusions

The critical word intrusion rate was uncorrelated with mean forward association strength per list. However, there was a strong positive correlation between critical word intrusions and mean backward association strength per list. There was also a moderately negative correlation between critical word intrusions and connectivity. Thus, fSAM predicted a trade-off between the influences on critical word intrusions of associations from the studied words to the critical word and of associations among studied words.

Other Intrusions

The patterns for extralist and prior-list intrusions were similar. Both were uncorrelated with forward association strength, both were only weakly and negatively correlated with backward association strength, and both were strongly and negatively correlated with connectivity.

True–False Correlation

Across the 125 lists, the correlation between veridical recall and critical word intrusions was −.52, consistent with behavioral data (Gallo & Roediger, 2002; Roediger, Watson, et al., 2001; Stadler et al., 1999).

Discussion

In this simulation, we were able to determine the model's predictions for the effects on veridical and false recall of three types of association strength—backward, forward, and intralist (connectivity) association strength—unconfounded with the influence of indirect, mediated associations and other factors that might affect veridical and false recall in the behavioral data. Unsurprisingly, backward association strength strongly predicted critical word intrusions, consistent with several empirical findings (Deese, 1959b; McEvoy et al., 1999; Roediger, Watson, et al., 2001).

The model predicted that forward association strength would have no effect on any of the dependent variables. This prediction is consistent with the finding by Roediger, Watson, et al. (2001)

that forward association strength is not correlated with either veridical recall or critical word intrusions. It is inconsistent with the finding by Brainerd and Wright (2005) that false recognition is positively associated with forward association strength when a broader range of strengths is used. However, one might plausibly expect that forward association strength would matter more in recognition than in free recall: When a critical word is presented at test as in recognition, the participant does not need to generate the critical word. Therefore, processing may be less dominated by backward association strength than may be the case in free recall, in which the participant would not produce the critical word unless it is generated through the use of backward associations. It remains to be seen whether the model's prediction is borne out by experiments testing the effects of forward association strength on free recall using a broader range of strengths.

The model predicted a strong positive correlation between connectivity and veridical recall, consistent with many findings in the literature (see, e.g., McEvoy et al., 1999; Roediger, Watson, et al., 2001). The model also predicted a moderate to strong negative correlation between connectivity and all types of intrusions—critical word, extralist, and prior-list intrusions. The tendency of studied words to cue each other thus serves to modulate the influence of backward association strength on false recall, according to fSAM.

This finding of a negative correlation between connectivity and critical word intrusions is consistent with the results reported by Deese (1959a, 1961) and McEvoy et al. (1999) but inconsistent with those reported by Roediger, Watson, et al. (2001). Roediger, Watson, et al. argued that the McEvoy et al. finding could be explained by the use of lists with only extremely high or extremely low mean connectivity and, consequently, a failure to use lists with moderate connectivity, as had Roediger, Watson, et al. Of course, it is also possible that the extensive use of lists with moderate levels of connectivity could have resulted in a restriction of the range of connectivity values and impeded the ability of Roediger, Watson, et al. to find a significant correlation.

In addition, the roughness of the measure of connectivity used by both McEvoy et al. (1999) and Roediger, Watson, et al. (2001) might have added noise that makes it difficult to detect a significant correlation. Their connectivity measure counted each pairwise connectivity association strength the same (equal to 1), ignoring differences in the normed word association probabilities. A further complication arises from the acknowledgment by Nelson et al. (2004) that the normed probabilities forming the basis of the connectivity values in McEvoy et al. and Roediger, Watson, et al. are likely to underestimate the number of connectivity associations, particularly of weak associations. This underestimation arises because the norming procedure asks participants to produce a single word as an associate of the cue word, thus biasing the norms toward measuring strong associations at the expense of weak associations. With these additional potential sources of noise, there is perhaps even greater need to use more extreme values of mean connectivity to ensure real differences in connectivity across lists and conditions, as McEvoy et al. did. Of course, this potential additional noise underscores the lack of reliable measures of association strength, a point we raised previously.

The finding of a negative true–false correlation in this simulation also points to the importance of the combined levels of connectivity and backward association strength in generating patterns of veridical recall and critical word intrusions across lists. In Simulation 5, we further explored the pattern of true–false correlations across lists with varying rates of critical word intrusions.

Simulation 5: Effects of Mean Backward Association Strength, Mean Connectivity, and Number of Studied Associates

In our model, the most likely and important contributing factor to a negative true–false correlation is connectivity, which has opposing influences on veridical recall and critical word intrusions: As shown in Simulation 4, all other things being equal, an increase in connectivity increases veridical recall and decreases critical word intrusions, which will produce a negative true–false correlation. Of course, all other things are not always equal. For example, there may be sets of lists in which connectivity influences—and correlates with—both veridical recall and critical word intrusions in the same way, either positively or negatively. For such a set of lists, we would expect a positive true–false correlation. In this simulation, we explored whether such a pattern of correlations involving connectivity might explain the shift from negative true–false correlations for sets of lists with low levels of critical word intrusions to positive true–false correlations for sets of lists with high levels of critical word intrusions. As mentioned previously, in the Roediger, Watson, et al. (2001) study, although the true–false correlation for all 55 lists was quite negative ($r = -.43$) and was also quite negative for the 19 lists with critical word intrusion rates less than or equal to .20 ($r = -.29$), the correlation was less negative for the top 36 lists ($r = -.07$) and actually positive for the top 18 lists with critical word intrusion rates above .40 ($r = .19$).

An additional goal of this simulation was to explore the effects of varying the number of the critical word's associates that appear on a studied list—that is, the number of words having more than a residual backward association strength to the critical word. Our principal reason for manipulating this variable was to assess the model's ability to simulate appropriate levels of critical word intrusions for lists in which backward association strength is either distributed evenly across the studied words or concentrated in a few strong associates of the critical word.

The lists used by Gallo and Roediger (2002) provide examples of lists with distributed association strengths to the critical word. In fact, a number of those lists include several words with associations sufficiently weak that they are not listed in the Nelson et al. (2004) norms. By contrast, Robinson and Roediger (1997) included lists in their study that had only a few very strong associates. They manipulated the number of critical word associates appearing on the studied list and found that critical word intrusions increased with the number of studied associates. However, the interpretation of their results is ambiguous because of the confounding of several variables, including number of associates, mean backward association strength of the list, and serial position of the associates (which were always presented first). Our simulation sought to determine which of these interpretations would be consistent with the model's predictions.

Simulation Method

Semantic Matrices

We again used abstract semantic matrices in this simulation. We factorially manipulated the number of studied critical word associates, the mean backward association strength of those associates to the critical word, and the mean connectivity among the studied words. The mean backward association strength per list was manipulated indirectly through the combination of the number of critical word associates and the mean backward association strength of those associates. We also used a finer grain for the manipulation of mean backward association strength and mean connectivity per list than we used in Simulation 4: Backward association strength and connectivity each ranged from .05 to 1.00 in steps of .05. The number of critical word associates in the lists ranged from 3 to 15 in steps of 3, as in Robinson and Roediger (1997). This arrangement yielded a total of 2,000 lists, each with a unique combination of values for the three variables.

To keep the semantic matrices to a computationally tractable size for each run, we split connectivity and backward association strength into a low range (.05–.50) and a high range (.55–1.0). For each of the four factorial combinations of these high and low ranges, we created semantic matrices with either 3, 6, 9, 12, or 15 associates to the critical word in each list of 15 words. This yielded 20 matrices of 100 lists each. Each of these matrices was created in the same manner as was the semantic matrix in Simulation 4, except that all forward association strengths were assigned residual values and only the designated number of critical word associates on each list were assigned nonresidual backward association strength values.

Simulation Procedure

The fSAM multiplicative model was run with 200 simulated subjects for each of the 20 semantic matrices using the parameters that provided the best fit in Simulation 1. Each simulated subject was presented with all 100 lists in a given matrix, with the order of the lists being randomly determined anew for each simulated subject. As in Simulation 4, the order of words within each list was randomized, and delayed free recall was used. The results of these runs were combined in our analysis to show the pattern of veridical and false recall over the complete ranges of backward association strength, connectivity, and number of associates.

Simulation Results

Critical Word Intrusions

Figure 3 shows fSAM's predicted critical word intrusion rates as a function of the mean backward association strength per list and the number of associates per list. That the functions for different numbers of associates lay on top of one another shows that the model's predictions for critical word intrusions were determined almost completely by the mean backward associative strength per list, without regard to the distribution of strengths across words in the list.

True–False Correlations

Table 8 shows the simulated true–false correlations for lists in the low, medium, and high ranges of critical word intrusion rates. The model qualitatively predicted two key results from Roediger, Watson, et al. (2001)—both the overall negative true–false correlation and the shift from negative to positive true–false correlations as the level of critical word intrusions increased. The table also shows that there was a corresponding shift in the correlation between critical word intrusions and connectivity, which also changed from negative to positive as critical word intrusion rates increased. By contrast, the correlation between connectivity and veridical recall remained strongly positive across the levels of critical word intrusions. Thus, for the lists with high critical word intrusion rates, connectivity was positively correlated with both true and false recall, yielding a positive true–false correlation.

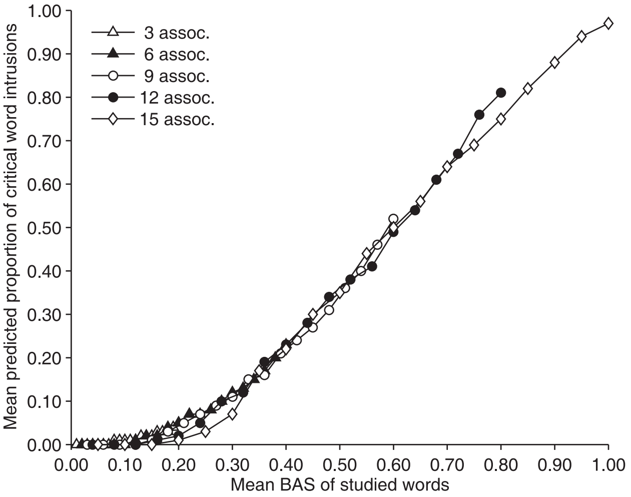

Figure 3. Predicted critical word intrusions as a function of the mean backward association strength (BAS) per list and the number of associates (assoc.) per list in Simulation 5.

Discussion

This simulation clarified that fSAM predicts that the rate of critical word intrusions is related to the mean backward association strength of the words appearing in a list, regardless of the distribution of strengths among those words. The model's predictions are consistent with the low levels of critical word intrusions for the lists used by Gallo and Roediger (2002), which consisted primarily of weak associates of the critical word. In the simulation, even if all the studied words were associates of the critical word, critical word intrusion rates remained low if the associates were all weak associates.

The model also predicted the qualitative pattern observed by Robinson and Roediger (1997), in which critical word intrusions increase as a function of the number of studied associates. The model explains that pattern as attributable to increasing mean backward association strength per list rather than to other factors. Increasing the number of studied associates per se did not increase critical word intrusions unless there was also an increase in mean backward association strength for the list. Because we randomized the order of words in the list, the serial positions of the studied associates cannot account for our simulated pattern. Thus, of the three variables we mentioned before that were confounded by Robinson and Roediger, fSAM predicts that critical word intrusions are related only to mean backward association strength for a list and not to the number or serial position of the associates.

The other finding of interest in Simulation 5 is that the true–false correlation and the correlations between critical word intru-

Table 8
Correlations Across Lists in Simulation 5, Conditionalized on Critical Word Intrusion Rates (CI)

Range of critical word intrusions	Correlation (*r*)		
	True–false	Critical word–connectivity	Studied words–connectivity
$0 < \text{CI} \leq .70$	−.29	−.31	.94
$0 < \text{CI} \leq .20$	−.16	−.19	.91
$.20 < \text{CI} \leq .40$	−.04	−.10	.93
$.40 < \text{CI} \leq .70$	.07	.04	.96

sions and connectivity both shifted from negative to positive as the level of critical word intrusions increased. We explain this pattern by noting that, for lists with high mean backward association strengths—those tending to produce high levels of critical word intrusions—recalling more studied words can facilitate intrusion of the critical word because the recalled words will tend to have strong associations to the critical word. Use of those associations multiplicatively in a retrieval cue would favor intrusion of the critical word, according to fSAM. Thus, increasing connectivity can facilitate intrusions of the critical word by increasing recall of its strong associates, which then become strong retrieval cues for the critical word. By contrast, increasing the connectivity of lists that have low mean backward association strengths will produce recall of more studied words that are, of course, not likely to serve as strong retrieval cues for the critical word because of their low backward association strength.

General Discussion

Our simulation results support a new theory of false memory based on the use of a multiplicative combination of semantic associations at both encoding and retrieval. This theory is instantiated in fSAM, a computational model within the SAM framework. The model performed well in simulations of several key findings reported in the DRM literature, as well as those in a new experiment. In Simulation 1, using a single parameter set, the model succeeded in simulating a rather intricate pattern of recall for DRM and non-DRM lists, including not just critical word intrusions but also veridical recall, extralist intrusions, and prior-list intrusions. In Simulation 2, the model succeeded in fitting developmental changes in these dependent variables, with the parameter values from Simulation 1 being allowed to vary across children's age groups in theoretically plausible ways. In Simulation 3, the model produced a good qualitative fit to the pattern of true and false recall across a total of 55 specific DRM lists. In Simulations 4 and 5, we generated model predictions for the effects on true and false recall of differences in three types of association strength—associations from critical words to studied words and vice versa, as well as among studied words—and in the number of critical word associates appearing in the studied list. These predictions corresponded well with effects reported in the literature.

Implications for Theories of False Memory

Neither of the prevailing theories of false memory—fuzzy trace theory and the activation-and-monitoring theory—have been incorporated in a quantitative model that specifies processes operating at both encoding and retrieval. A quantitative model of false recall based on fuzzy trace theory (Brainerd et al., 2003) covers only the theory's test-phase decision processes, and it has been applied to the DRM paradigm only in connection with critical word intrusions and veridical recall. In particular, that model does not quantitatively specify encoding processes such as those involved in the creation of gist and verbatim traces during study. By contrast, our model quantitatively specifies processes operating at both encoding and retrieval. Thus, to our knowledge, fSAM stands alone as a quantitative model of false recall that is fully specified at encoding and retrieval.

It remains to be seen whether either of those other theories will be able to simulate findings in the DRM literature as well as fSAM has when their assumptions are more rigorously specified in a quantitative model. Of particular interest is the ability of those theories to simulate not just appropriately high levels of critical word intrusions but also appropriately low levels of extralist and prior-list intrusions—an important constraint on false-memory theories. Also of interest is the ability of those theories to predict the pattern of true–false correlations across different levels of critical word intrusions. In Simulation 5, we showed that fSAM can simulate not only the overall negative true–false correlation but also the shift in the correlation from negative to positive as the level of critical word intrusions increases. We also offered an explanation for that pattern: Lists with high critical word intrusion rates tend to have many words with high backward association strengths; as connectivity increases across such lists, more studied words are recalled and are then used as strong retrieval cues for critical word intrusions, resulting in correlated increases in both true and false recall. Fuzzy trace theory has been described as predicting a negative true–false correlation (see, e.g., Brainerd et al., 2006) but would not seem to predict the negative-to-positive shift in the correlation in any straightforward manner, at least without some further modification of the theory.

Relative Effectiveness of Semantic Encoding and Retrieval Mechanisms

In addition to the best fitting model that incorporated multiplicative encoding and retrieval mechanisms, we tested other models in Simulation 1 that included single-association and additive versions of the encoding and retrieval mechanisms. These other mechanisms, operating alone, failed to produce as good a fit as the multiplicative version of the same mechanism. The fit improved for both the single-association and additive mechanisms when the encoding and retrieval versions of those mechanisms were used in combination, but those combined fits were still not as good as the fit for the combined multiplicative mechanisms. This pattern of results has implications for spreading activation and compound cue theories.

Comparison With Spreading Activation

Broadly speaking, our single-association and additive semantic mechanisms operate in a manner similar to spreading activation (see, e.g., Anderson, 1983; Collins & Loftus, 1975; for a review, see McNamara, 2005). As with spreading activation, words become more accessible during encoding to the extent that they are semantically related to each studied word, with the incrementing in accessibility occurring either once upon the studied word's presentation (single-item encoding) or multiple times as the studied word is repeatedly rehearsed (additive encoding). In the latter case, a word's increment in accessibility summates across multiple rehearsals, a feature that is consistent with spreading activation. In a parallel to spreading activation at retrieval, a recalled item cues retrieval of semantically related words either on the next recall attempt (single-cue retrieval) or on the next several attempts (additive retrieval). In the latter case, semantic association strength summates across multiple recalls, again consistent with spreading activation.

An obvious difference between our mechanisms and spreading activation is that the representations of the words themselves do not vary in strength in our models as they do in spreading activation. Instead, all of the processes in our models involve incrementing and decrementing of association strengths. However, this difference would seem to have little practical effect in the free-recall tasks simulated in this article, in contrast to other tasks such as recognition, naming, and lexical decision.

It bears keeping in mind that, although the semantic component of the single-cue and additive retrieval mechanisms can be viewed as similar to spreading activation, the basic retrieval mechanism in the models remains multiplicative in that the semantic strength is multiplied together with the contextual and episodic strengths. In that sense, even the models incorporating single-cue and additive retrieval mechanisms are in essence compound cue models, although not as pure an example of a compound cue model as the multiplicative retrieval mechanism, as we discuss next.

Comparison With Compound Cuing

The multiplicative versions of the semantic encoding and retrieval mechanisms may be classified as configural or compound cue mechanisms (Dosher & Rosedale, 1989; Ratcliff & McKoon, 1988; for a review, see McNamara, 2005). Compound cue theory was offered as an alternative to spreading activation as an explanation of semantic priming in such tasks as recognition, naming, and lexical decision. As applied to such tasks, the theory assumes that an earlier presented prime is used jointly with a target to form a compound cue, facilitating processing of the target and thus yielding priming effects.

The crucial feature of the compound cue theory is the targeting of words that are associated to all of the multiple words constituting the compound cue—the intersection or multiplicative cuing principle (Dosher & Rosedale, 1997; Humphreys, Wiles, & Bain, 1993). Multiplying association strengths is one means of increasing the ratio of strengths between those words that are and are not strongly associated with each of the compound cue words. Ratcliff and McKoon (1988, 1995) used the multiplied strengths incorporated in the SAM recognition model described by Gillund and Shiffrin (1984) to implement compound cue theory in modeling semantic priming effects in recognition and lexical decision.

Dosher and Rosedale (1997) have suggested the possibility of using previously recalled words as cue sets in free recall, but to our knowledge, our multiplicative retrieval mechanism is the first actual implementation that uses multiple previously recalled words as joint semantic retrieval cues in free recall. Our multiplicative encoding mechanism also appears to be the first of its kind in preferentially incrementing strengths of words in proportion to their strength of association to all the studied words in a particular set.

Failure of the Single-Cue Retrieval Mechanism

The difficulty for the single-cue retrieval mechanism in simulating enough critical word intrusions in the DRM conditions has ramifications for theories that assume, either explicitly or implicitly, that simple pairwise semantic associations are sufficient to generate the high rates of critical word intrusions observed in DRM list recall. Our single-item retrieval model is based on the eSAM model described by Sirotin et al. (2005) and arguably involves the least extrapolation beyond the basic SAM model. Although the results Sirotin et al. reported for the eSAM model show that it is successful in simulating certain aspects of extralist intrusion and prior-list intrusion data, our results show that it is too simple to simulate intrusions more globally, particularly critical word intrusions. Single semantic associations do not appear to discriminate sufficiently between the critical word and other related words, presumably because the aggregate semantic association strength to the other words in the lexicon surpasses the single semantic association strength to the critical word.

Variations in Global Semantic Association Strength

As we discussed in connection with Simulation 2, an additional explanation of the developmental pattern of results is that semantic associations grow in number and in strength as children mature. A future avenue of exploration with fSAM will be to simulate such developmental changes by varying average semantic association strength across the semantic matrix. Changes in global semantic association strength might also be useful in simulating developmental patterns on the other end of the life span as well, although these changes may also be consistent with parameter value changes. Another set of phenomena to explore with this approach would be decrements in memory functioning due to brain damage.

One issue with such an approach is that it would involve making a number of additional assumptions. A prime example of such an assumption concerns the extent to which such changes in association strength occur more intensively within clusters of semantic associations, such as categories or DRM lists, as opposed to more diffusely throughout the entire matrix. A related assumption concerns the extent to which connectivity and backward association strength would change, both separately and together. The patterns of veridical recall, critical word intrusions, and extralist intrusions would likely be quite sensitive to such assumptions. On the other hand, it seems likely that any pattern of increasing global association strength would be unable to simulate a decline in prior-list intrusions with development or the related dissociation between prior-list intrusions and other intrusions. Such a decline would seem to require a theoretically plausible parameter change such as we included in Simulation 2.

Immediate Versus Delayed Free Recall in the SAM Framework

Our results call into question the assumption of the SAM framework that the only consequence attributable to testing recall immediately after study rather than at a delay is access to the contents of STM as of the end of study. Others have challenged the sufficiency of SAM's explanation of the recency effect, which relies on this differential access to STM contents, particularly in connection with the continuous distractor paradigm (Bjork & Whitten, 1974). Our results do not directly challenge SAM's account of the recency effect but rather the collateral assumption that all subjects always initiate recall by emptying STM contents. Such a strategy may well evolve over numerous study–recall trials (e.g., 80 trials for each subject in Murdock, 1962) or may even be explicitly suggested by the experimenter (e.g., Roediger & McDermott, 1995). A number of studies of immediate free recall in which participants studied a relatively small number of lists and were not given specific instructions regarding output order have shown recency effects that are much less pronounced than SAM would predict (e.g., Glanzer & Cunitz, 1966; Haarmann & Usher, 2001; Postman & Phillips, 1965; Tan & Ward, 2000).

This reduced recency effect might be explained by the fact that without an explicit instruction to output the last few studied items first, many subjects initiate immediate free recall with nonrecency items. For example, in Kimball and Bjork (2002), subjects initiated recall with one of the last three studied items on only 37% of the trials, comparable to the 36% of trials on which recall began with one of the first three items studied. As a consequence, the recency effect for Kimball and Bjork was smaller than predicted by our model. In addition, the output position of the critical word was earlier than predicted by our model, presumably in part because not all trials began with the output of several studied items from STM, which would tend to push the critical word later in the output queue. At a minimum, then, our results argue for the incorporation of recall initiation strategy into the SAM model (see, e.g., Metcalfe & Murdock, 1981).

Forgetting

Another likely consequence of immediate testing is reflected in the lower behavioral rates of extralist intrusions and prior-list intrusions for Kimball and Bjork (2002) than for the new experiment. It seems likely that these differences are attributable to different effects of forgetting across shorter versus longer delays between study and test of each list. In our model, the only forgetting that takes place is between lists, through contextual decay using the ρ parameter (which was quite low in the best fits, so there was substantial forgetting across the interlist interval in our simulations). To produce the different levels of extralist intrusions and prior-list intrusions for immediate versus delayed free recall, what seems to be needed is a mechanism that allows for forgetting across the retention interval.

A good candidate for such a mechanism is provided by the stimulus fluctuation and sampling theory (Estes, 1955a, 1955b), which was incorporated into SAM by Mensink and Raaijmakers (1988). Mensink and Raaijmakers represented context as a vector of elements (see also Howard & Kahana, 2002; Shiffrin & Steyvers, 1997). Each contextual element is in either an active or inactive state at any given time. The identity of the active contextual elements changes over time—that is, context drifts—with some active elements becoming inactive and some inactive elements becoming active at each time step. At a given time step, associations between active contextual elements and items then in STM are strengthened. Memory is probed using the contextual elements active at the time of test. Therefore, the probability that an item will be sampled and recovered is proportional to the number of contextual elements that are active at both the time of encoding the item and the time of test. By incorporating contextual drift in this way, Mensink and Raaijmakers were able to use SAM to simulate various interference and forgetting effects observed in paired-associates experiments.

The mechanism developed by Mensink and Raaijmakers (1988, 1989) could produce differences in extralist intrusion and prior-list intrusion rates as a function of retention interval by increasingly favoring retrieval of words with strong semantic associations to studied words as the retention interval increases. With longer retention intervals, a smaller proportion of the contextual elements active at encoding would remain active at test. As a result, unstudied words would be less distinguishable from studied words on the basis of contextual strength. In such a case, words—whether studied or unstudied—that are strongly associated to previously recalled words will be more likely to be retrieved. Thus, semantically induced extralist intrusions and prior-list intrusions would be more likely at longer retention intervals. Clearly, a future avenue of exploration will be to evaluate the incorporation of the Mensink and Raaijmakers contextual drift mechanism into our models, to determine whether it can provide an even better fit of these aspects of the data.

Postretrieval Decision Processes

Although fSAM includes fully specified processes at encoding and retrieval, it does not at present include mechanisms for postretrieval decision processes. Such processes are likely to be involved in some of the findings in the DRM literature, such as those involved in conjoint recall (Brainerd et al., 2003) and in rejection of critical words that are longer or shorter than the studied words on a list (Madigan & Neuse, 2004). Postretrieval decision processes can be added to fSAM in the future, such as by adding a response criterion for a relevant dimension. For example, to simulate conjoint recall, one could add a response criterion for semantic association strength to filter retrieved words that are or are not judged as sufficiently similar in meaning to other studied words from a list. Similarly, one could add a response criterion based on word length of a retrieved word relative to the lengths of other words from the episode. A goal of the present article was to determine the degree to which a number of core findings in the literature could be simulated without adding the complexity attendant to such postretrieval processes.

Recognition Memory

We have focused on simulating false recall in this article, but extending fSAM to recognition memory is a logical future step. For one thing, SAM has previously been extended to cover recognition (Gillund & Shiffrin, 1984; Ratcliff, Van Zandt, & McKoon, 1995; Shiffrin et al., 1990). In the general version of the recognition model proposed by Gillund and Shiffrin (1984), familiarity and recollection both play a role in recognition, and there is a potential role for other decision processes as well. Familiarity involves using the test item and context as cues, calculating for each item in the lexicon a product of the item's contextual association strength and its strength of association to the test item, then summing those products across all items in the lexicon to determine the familiarity of the test item. This familiarity value is the same as the denominator in the sampling rule for recall, thus uniting recognition and recall theoretically. If familiarity is sufficiently high or low, the test item can be accepted or rejected on the basis of familiarity alone, but otherwise, a recollection process is used to search LTM in a manner similar to recall. Strategic decision processes can also adjust the relative roles of familiarity and recollection. The SAM recognition model has been used to simulate a number of effects in the recognition literature including list-length effects for lists of unrelated words (Gillund & Shiffrin, 1984; but see also Gronlund & Elam, 1994) and categorized lists (Shiffrin, Huber, & Marinelli, 1995), list-strength effects for lists of unrelated words (Shiffrin et al., 1990) and categorized lists (Shiffrin et al., 1995), the generation effect (Clark, 1995), effects of presentation speed and retention interval (Gillund & Shiffrin, 1984), and multiple-choice recognition (Mensink & Raaijmakers, 1989).

The fSAM model can similarly be extended to deal with recognition by adding familiarity and other decision processes as contemplated by Gillund and Shiffrin (1984). This approach is similar to other dual-process theories of recognition that have been advanced to explain the false-recognition effect in the DRM paradigm (for reviews, see Brainerd & Reyna, 2005; Gallo, 2006). There are a number of core findings reported in the literature regarding false recognition that can serve as good tests of a quantitative model such as fSAM, including the basic DRM false-recognition effect (Roediger & McDermott, 1995), list-length effects (Robinson & Roediger, 1997), levels of processing effects (McCabe, Presmanes, Robertson, & Smith, 2004; Soraci, Carlin, Toglia, Chechile, & Neuschatz, 2003; Toglia, Neuschatz, & Goodwin, 1999), speeded recognition (Benjamin, 2001), effects of recall on subsequent recognition tests (Gallo, McDermott, Percer, & Roediger, 2001), conjoint recognition (e.g., Brainerd, Wright, Reyna, & Mojardin, 2001), and dissociations between false recall and false recognition (e.g., differences in the effects of connectivity on false recall; McEvoy et al., 1999). Of course, as with any findings, fSAM's ability to simulate false-recognition effects can only be tested with actual model implementation and simulation.

The SAM approach to recognition that we have outlined here is rather different from that adopted in a number of other quantitative models of recognition memory. These models represent words as vectors of features, and the recognition process involves comparing the features of a test probe word with corresponding features of all other words and summing the similarities across all features and all words in the lexicon (see, e.g., Hintzman, 1988; Kahana & Sekuler, 2002; Nosofsky, 1992; Shiffrin & Steyvers, 1997). Indeed, Arndt and Hirshman (1998) have already applied Hintzman's (1988) MINERVA 2 model to the basic false-recognition effect in the DRM paradigm. It remains to be seen how such models will perform in simulating a broader array of false-recognition results and how that performance will compare with that of the fSAM recognition model.

Methodological Points

Simultaneous Satisfaction of Multiple Constraints

The success of our model is all the more striking because of the multiple constraints we imposed. Of course, the overarching constraint that we imposed was to specify all of our assumptions quantitatively by implementing our theoretical mechanisms in a computational model. In addition, in Simulation 1, we constrained our model to use a single set of parameters to fit simultaneously not just the data from conditions that used DRM lists but also the data from conditions that used lists comprising one word from each DRM list and lists comprising words that were not systematically related. Moreover, the data that were fit included not just the levels of critical word intrusions but also the output percentile of the critical word, the serial position curve, and the levels of veridical recall, extralist intrusions, and prior-list intrusions in all four simulated conditions, for which the behavioral means had exhibited an intricate pattern. Other important constraints that we imposed but that tests of memory models historically have not imposed are the use of actual behavioral norming data as the basis for semantic association strengths among words and the inclusion of both studied and unstudied words in our lexicons. Notwithstanding all of these constraints, our model was able to simulate the data well both qualitatively and quantitatively in Simulation 1.

It is important to use multiple constraints as we have because doing so allows the modeler to distinguish between special-purpose models and more general models. For example, many of our single-mechanism models were able to simulate appropriate levels of critical word intrusions in Simulation 1A when we weighted the fit of those intrusions 100 times as much as the fit of either veridical recall or other intrusions. However, the fits of the other dependent variables were extremely poor in most models (although a notable exception was the model incorporating only the multiplicative retrieval mechanism). If we only sought to fit the critical word intrusions in DRM lists, many of our mechanisms would suffice. Only by demanding that a model also fit the other dependent variables and that it do so in non-DRM conditions as well as DRM conditions were we able to identify mechanisms that were plausible as more general explanations of memory processes. This is a test to which all theories of false memory should be subjected.

Simultaneously fitting multiple conditions and multiple dependent variables with a single parameter set is also important because it allows us to evaluate the extent to which the model's mechanisms, rather than variations in parameter values, are responsible for fitting the data. Because our fit in Simulation 1 cannot be explained by any differences in parameter values across conditions, the mechanisms themselves must be responsible for the fits, and that allows a clearer evaluation of the theory implemented in the mechanisms.

Of course, the general SAM theory contemplates variations in parameter values to accommodate differences in conditions, tasks, and strategies across experimental conditions, so the single-parameter-set constraint is over and above those imposed by the general SAM theory. Indeed, we allowed parameters to vary in theoretically plausible ways in simulating the developmental data in Simulation 2. We also allowed parameters to vary across different experiments from different labs using different sets of lists in simulating specific list effects in Simulation 3. However, in Simulations 4 and 5, we used the parameter set from Simulation 1 to generate model predictions that were quite consistent with reported patterns of effects on true and false recall of forward, backward, and intralist association strengths and number of studied associates. Obtaining such good qualitative predictions using a parameter set from a fit of different data indicates the robustness and generalizability of the mechanisms incorporated in our model.

Semantic Association Strength Metric

We used WAS as our semantic association strength metric in Simulations 1–3 because it is based on word production norms, provides a value for each pair of words, and has performed well in categorized list recall in Sirotin et al. (2005). There are other semantic association strength metrics available, for example, latent semantic analysis (Landauer & Dumais, 1997) and Wordnet (Miller, 1996). These metrics are not based on word production norms, nor do they provide a complete set of association strengths for all pairs of words. Latent semantic analysis also discriminates more poorly than WAS between related and unrelated words (see Sirotin et al., 2005). Accordingly, these did not seem to be viable metrics for our purposes.

WAS thus seems the best metric that is currently available for our purposes. At a global level, the mean WAS values mapped well onto our subjective estimation of the relative levels of semantic association strength between words on a DRM list and the corresponding critical word, among words on a DRM list, and among words selected unsystematically. However, as we noted in connection with Simulation 3, WAS values correlate only moderately with normed backward association strengths for particular words and lists, as opposed to classes of words as a whole. One factor contributing to this problem is the incorporation of indirect, mediated associations into WAS values. Incorporating such associations is necessary to allow for the computation of WAS values for pairs of words with no normed direct association strengths, but the incorporation of such indirect associations dilutes the effects of direct associations, which are more important for simulating free recall (see Steyvers et al., 2005).

One way that WAS might be improved as a semantic metric for word list recall would be to disaggregate WAS values into separate values for forward and backward association strengths given that those strengths are often different for a given pair of words (e.g., *dog* as a cue may elicit *house* in a free-association task more often than would the reverse) and that the forward and backward associations may have different effects on false memory (cf. Brainerd & Wright, 2005; Roediger, Watson, et al., 2001). Taking this step did not seem fruitful for the present article given that there would still be a substantial problem in simulating free recall because of incorporation of indirect association strengths, but such a step may prove beneficial for simulation of results based on recognition or cued recall.

Conclusions

Our results support a new theory of false recall that assumes that people use conjunctions of semantic associations to process information at encoding and retrieval. A quantitative model implementing this theory, fSAM, simulated a number of findings in the DRM literature. The success of the model presents a challenge to the leading theories offered to date as explanations of false recall inasmuch as those theories have not been fully specified in a quantitative model and thus have not been as rigorously tested as our theory. Of course, our model also faces challenges, in particular those involving further tests of its viability and generalizability, including application and extension of the model to other paradigms such as categorized list recall and recognition memory. The results thus far bode well for such tests.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22,* 261–295.

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

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

Bäuml, K., & Kuhbandner, C. (2003). Retrieval-induced forgetting and part-list cuing in associatively structured lists. *Memory & Cognition, 31,* 1188–1197.

Benjamin, A. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 941–947.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6,* 173–189.

Bjorklund, D. F. (1987). How age changes in knowledge base contribute to the development of children's memory: An interpretive review. *Developmental Review, 7,* 93–130.

Bjorklund, D. F., & Jacobs, J. W. (1985). Associative and categorical processes in children's memory: The role of automaticity in the development of organization in free recall. *Journal of Experimental Child Psychology, 39,* 599–617.

Brainerd, C. J., Forrest, T. J., Karibian, D., & Reyna, V. F. (2006). Development of the false-memory illusion. *Developmental Psychology, 42,* 962–979.

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. (1998). Fuzzy-trace theory and children's false memories. *Journal of Experimental Child Psychology, 71,* 81–129.

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review, 24,* 396–439.

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

Brainerd, C. J., Reyna, V. F., & Forrest, T. J. (2002). Are young children susceptible to the false-memory illusion? *Child Development, 73,* 1363–1377.

Brainerd, C. J., & Wright, R. (2005). Forward association, backward association, and the false-memory illusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 554–567.

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.

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

Cole, M., Frankel, F., & Sharp, D. (1971). Development of free recall learning in children. *Developmental Psychology, 4,* 109–123.

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

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 33*(A), 497–505.

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

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

Deese, J. (1961). Associative structure and the serial reproduction experiment. *Journal of Abnormal and Social Psychology, 63,* 95–100.

Dempster, E. N. (1981). Memory span: Sources of individual and developmental differences. *Psychological Bulletin, 89,* 63–100.

Dewhurst, S. A., & Robinson, C. A. (2004). False memories in children: Evidence for a shift from phonological to semantic associations. *Psychological Science, 15,* 782–786.

Dosher, B. A., & Rosedale, G. S. (1989). Integrated retrieval cues as a mechanism for priming in retrieval from memory. *Journal of Experimental Psychology: General, 118,* 191–211.

Dosher, B. A., & Rosedale, G. S. (1997). Configural processing in memory retrieval: Multiple cues and ensemble representations. *Cognitive Psychology, 33,* 209–265.

Estes, W. K. (1955a). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369–377.

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

Gallo, D. (2006). *Associative illusions of memory: False memory research in DRM and related tasks.* New York: Psychology Press.

Gallo, D., McDermott, K., Percer, J., & Roediger, H. L. (2001). Modality effects in false recall and false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 339–353.

Gallo, D., & Roediger, H. L. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language, 47,* 469–497.

Gillund, G., & Shiffrin, R. M. (1981). Free recall of complex pictures and abstract words. *Journal of Verbal Learning and Verbal Behavior, 20,* 575–592.

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

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5,* 351–360.

Goodwin, K. A., Meissner, C. A., & Ericsson, K. A. (2001). Toward a model of false recall: Experimental manipulation of encoding context and the collection of verbal reports. *Memory & Cognition, 29,* 806–819.

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

Gronlund, S. D., & Shiffrin, R. M. (1986). Retrieval strategies in recall of natural categories and categorized lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 550–561.

Haarmann, H., & Usher, M. (2001). Maintenance of semantic information in capacity-limited item short-term memory. *Psychonomic Bulletin & Review, 8,* 568–578.

Hall, J. W., & Tinzmann, M. B. (1985). Presentation-rate effects and age differences in children's free recall. *Bulletin of the Psychonomic Society, 23,* 227–229.

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

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

Howe, M. L. (2006). Developmentally invariant dissociations in children's true and false memories: Not all relatedness is created equal. *Child Development, 77,* 1112–1123.

Humphreys, M. S., Wiles, J., & Bain, J. D. (1993). Memory retrieval with two cues: Think of intersecting sets. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 489–507). Cambridge, MA: MIT Press.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3–28.

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

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research, 42,* 2177–2192.

Kimball, D. R., & Bjork, R. A. (2002). Influences of intentional and unintentional forgetting on false memories. *Journal of Experimental Psychology: General, 131,* 116–130.

Kobasigawa, A. (1977). Retrieval strategies in the development of memory. In R. V. Kail, Jr., & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 177–201). Hillsdale, NJ: Erlbaum.

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

Madigan, S., & Neuse, J. (2004). False recognition and word length: A reanalysis of Roediger, Watson, McDermott, and Gallo (2001) and some new data. *Psychonomic Bulletin & Review, 11,* 567–573.

Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition, 25,* 826–837.

McCabe, D. P., Presmanes, A. G., Robertson, C. L., & Smith, A. (2004). Item-specific processing reduces false memories. *Psychonomic Bulletin & Review, 11,* 1074–1079.

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

McDermott, K. B. (1997). Priming on perceptual implicit memory tests can be achieved through presentation of associates. *Psychonomic Bulletin & Review, 4,* 582–586.

McEvoy, C. L., Nelson, D. L., & Komatsu, T. (1999). What is the connection between true and false memories? The differential roles of interitem associations in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1177–1194.

McKone, E., & Murphy, B. (2000). Implicit false memory: Effects of modality and multiple study presentations on long-lived semantic priming. *Journal of Memory and Language, 43,* 89–109.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition.* New York: Psychology Press.

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

Mensink, G.-J. M., & Raaijmakers, J. G. W. (1989). A model for contextual fluctuation. *Journal of Mathematical Psychology, 33,* 172–186.

Metcalfe, J., & Murdock, B. B. (1981). An encoding and retrieval model of single-trial free recall. *Journal of Verbal Learning and Verbal Behavior, 20,* 161–189.

Miller, G. (1996). Meaning matters: Problems in sense resolution. In D. Steier & T. M. Mitchell (Eds.), *Mind matters: A tribute to Allen Newell* (pp. 119–132). Hillsdale, NJ: Erlbaum.

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

Moely, B. E. (1977). Organizational factors in the development of memory. In R. V. Kail, Jr., & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 203–236). Hillsdale, NJ: Erlbaum.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64,* 482–488.

Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology, 86,* 263–267.

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

Nosofsky, R. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes: Vol. 1. From learning theory to connectionist theory* (pp. 149–167). Hillsdale, NJ: Erlbaum.

Ornstein, P. A., Naus, M. J., & Liberty, C. (1975). Rehearsal and organizational processes in children's memory. *Child Development, 46,* 818–830.

Phillips, J. L., Shiffrin, R. J., & Atkinson, R. C. (1967). The effects of list length on short-term memory. *Journal of Verbal Learning and Verbal Behavior, 6,* 303–311.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology, 17,* 132–138.

Price, J. L., Metzger, R. L., Williams, D., Phelps, N. Z., & Phelps, A. M. (2001, April). *Children produce as many false memories as adults (sometimes!)*. Poster presented at the biennial meeting of the Society for Research in Child Development, Minneapolis, MN.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing* (pp. 216–260). Cambridge, MA: MIT Press.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science, 27,* 431–452.

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

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review, 95,* 385–408.

Ratcliff, R., & McKoon, G. (1995). Sequential effects in lexical decision: Tests of compound-cue retrieval theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1380–1388.

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

Read, J. (1996). From a passing thought to a false memory in 2 minutes: Confusing real and illusory events. *Psychonomic Bulletin & Review, 3,* 105–111.

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

Reysen, M. B., & Nairne, J. S. (2002). Part-set cuing of false memories. *Psychonomic Bulletin & Review, 9,* 389–393.

Ritter, K., Kaprove, B. H., & Fitch, J. P. (1973). The development of retrieval strategies in young children. *Cognitive Psychology, 5,* 310–321.

Robinson, K., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science, 8,* 231–237.

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

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

Roediger, H. L., McDermott, K. B., & Robinson, K. J. (1998). The role of associative processes in creating false memories. In M. Conway, S. Gathercole, & C. Cornoldi (Eds.), *Theories of memory* (Vol. 2, pp. 187–245). New York: Psychology Press.

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

Schneider, W. (2002). Memory development in childhood. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 236–256). Oxford, England: Blackwell Publishers.

Schneider, W., & Pressley, M. (1997). *Memory development between two and twenty* (2nd ed.). Mahwah, NJ: Erlbaum.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

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

Shiffrin, R. M., & Raaijmakers, J. G. W. (1992). The SAM retrieval model: A retrospective and prospective. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (pp. 69–86). Potomac, MD: Erlbaum.

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

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

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

Smith, S. M., Gerkens, D. R., Pierce, B. H., & Choi, H. (2002). The roles of associative responses at study and semantically guided recollection at test in false memory: The Kirkpatrick and Deese hypotheses. *Journal of Memory and Language, 47,* 436–447.

Soraci, S. A., Carlin, M. T., Toglia, M. P., Chechile, R. A., & Neuschatz, J. S. (2003). Generative processing and false memories: When there is no cost. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 511–523.

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

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 237–249). Washington, DC: American Psychological Association.

Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1589–1626.

Toglia, M. P., Neuschatz, J. S., & Goodwin, K. A. (1999). Recall accuracy and illusory memories: When more is less. *Memory, 7,* 233–256.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology, 70,* 122–129.

(*Appendixes follow*)

Appendix A

Description of the New Experiment

To evaluate the contributions of our new semantic mechanisms more fully, we conducted a new experiment with the goal of producing rather complex patterns of veridical-recall and intrusion data that would impose more constraints on our model. In particular, we thought it important to simulate recall of lists comprising semantically unrelated words, matched with the Deese–Roediger–McDermott paradigm (DRM) list words on several psycholinguistic dimensions, as well as lists comprising DRM list words that were distributed across lists rather than concentrated in a single list.

The key independent variable in the new experiment was a three-level, between-subjects manipulation of materials. Participants in the standard condition studied 15 standard DRM lists, each with 15 words. Participants in the mixed condition studied the same 225 words as the first group, but each of their 15 lists was a mixture comprising one word from each of the 15 DRM lists that the standard group had studied. Participants in the control condition also studied 225 words—15 lists of 15 words—but the words were not semantically related to each other in any systematic way. Instead, the words were randomly selected subject to multiple constraints—specifically, that they match the 225 words studied by each of the first two groups on several dimensions, including word frequency, number of letters, number of syllables, and normed judgments of concreteness, imageability, and familiarity.

This design thus produced conditions in which (a) all words in a list were semantically related to a single unpresented critical word, but words on different lists were not systematically related to each other nor jointly to any unpresented word (standard condition); (b) sets of words on different lists were each semantically related to a single unpresented word, but words within each list were not systematically related to each other nor jointly to any unpresented word (mixed condition); and (c) neither the words within each list nor words on different lists were systematically related to each other or to any unpresented word (control condition).

We expected different patterns across the three conditions for veridical recall, critical word intrusions, extralist intrusions, and prior-list intrusions. Not surprisingly, we expected that the mean number of critical word intrusions would be higher in the standard condition than in the mixed and control conditions. We also expected that critical word intrusions in the mixed condition would be reliably greater than zero, based on an expectation that the effects of associations between a critical word and its strong semantic associates would accumulate across lists, but more weakly than in the standard DRM lists because of the delays between presentations of the words from a particular DRM list. We expected that critical word intrusions in the control condition would be at or close to zero.

We thought that extralist intrusions would be generated primarily because of pairwise associations between studied words and unstudied words (other than the critical word). We further thought that, except for the pairwise associations among words from a particular DRM list and between each of those words and the list's critical word, the average strength of association among pairs of words in the lexicon would be similar. We therefore expected no differences among the rates of extralist intrusions in the three conditions.

We thought that prior-list intrusions would be generated in part because of pairwise semantic associations between words from a previously studied list and a subsequently studied list. We therefore expected that there would be more prior-list intrusions in the mixed condition than in the other two conditions because of the distribution of the strongly related words from a particular DRM list across the study lists in the mixed condition.

Finally, based on the stronger associations among DRM list words than among either control words or words from different DRM lists, we expected that veridical recall would be higher in the standard condition than in the mixed and control conditions.

Method

Participants

Participants were 128 undergraduate students at the University of Texas at Arlington enrolled in the Introduction to Psychology course, participating for partial course credit, and quasi-randomly assigned to conditions. After excluding 1 participant in each of the standard and mixed conditions and 2 participants in the control condition because of equipment malfunction and failure to follow instructions, there were 43 participants in the standard condition, 41 in the mixed condition, and 40 in the control condition.

Materials

The 15 DRM lists were selected from among those that Stadler et al. (1999) reported as producing the highest rates of false recall and were based on the following critical words: *anger, chair, cold, doctor, mountain, needle, rough, sleep, slow, smell, smoke, soft, sweet, trash,* and *window*. These comprised the 12 lists used in Kimball and Bjork (2002, Experiment 2) plus the *doctor, mountain,* and *trash* lists. The average rates of critical word intrusions and veridical recall reported by Stadler et al. for these 15 lists were .52 and .60, respectively. For the words on each list, see Stadler et al.

The set of 225 words used for the control condition lists were selected from among those reported on the interactive MRC Psycholinguistic Database (Coltheart, 1981) at www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm. The words were selected to match the set of 225 DRM list words closely as to the mean, median, standard deviation, minimum, and maximum values for each of the following measures: word frequency, number of letters, number of syllables, and normed judgments of concreteness, imageability, and familiarity. In both sets, there were a number of words for which less than all of these measures were available; those words were included only in calculating averages for those measures for which values were available.

In the standard condition, the presentation orders of the DRM lists and of words within lists were randomized anew for each participant. In the mixed condition, the assignments of words to lists and to serial positions within lists were randomized anew for each participant, with the constraint that exactly one word from each of the 15 DRM lists in the standard condition appeared in each of the lists in the mixed condition. In the control condition, the assignments of words to lists and to serial positions within lists were randomly determined anew for each participant.

Procedure

An eMac computer presented the stimuli, controlled the timing of tasks, and recorded participants' responses. Participants were first given oral instructions by the experimenter, in which they were told they would be memorizing word lists and solving math problems and that they should do their best on both tasks. The experimenter then provided step-by-step instructions for each task as samples of the displays corresponding to each step were presented on the computer. After the instructions, participants were left alone to complete the experiment proper.

Word lists were presented auditorily through headphones connected to the computer. Prior to the beginning of each list, there was a verbal warning to the participant to get ready for the next list. After a delay of 5 s, the 15 words in a list were presented at a rate of one word every 1.5 s.

After the last word in the list had been presented, a chime signaled the end of the list and the beginning of the next task, which was a series of 10 linked arithmetic problems that appeared in a single display on the computer screen. Participants were told to type in the answer to each problem, moving down the screen.

After 30 s had elapsed for the math task, the problems were replaced by the recall test display, and a chime sounded, signaling the participant to begin recalling the words from the most recent word list in any order. Participants had been told in the instructions that they would have 90 s to type all the words they could recall from that list, that they should keep trying for the full 90 s, that they should only type in a word if they were confident that they had heard the word on the list, and that they should not guess. At the end of the 90 s, the verbal warning to get ready for the next list was given, and the procedure began anew for the next list, and so on through the 15 lists.

Results

Data were scored using three different criteria, but the patterns of results were similar for the three criteria, so we report the results using a moderately liberal criterion. For this criterion, differences in tense, number, and spelling were ignored; homophones were scored as correct; compound words reported as two words were scored as correct (e.g., *mole hill* for *molehill*); and repetitions were ignored. In addition, for the unpresented critical words, all parts of speech (e.g., *anger* and *angry*) were scored as the critical word, and in the standard condition, critical words from prior lists (n = 4) and subsequent lists (n = 2) were scored as extralist intrusions.

Veridical Recall

Recall of studied words differed across the three groups, $F(2, 121) = 54.91$, $MSE = .01$, $p < .0001$. Planned comparisons revealed that recall in the standard condition ($M = .47$, $SE = .01$) was reliably higher than in both the mixed condition ($M = .31$, $SE = .01$), $F(1, 82) = 103.46$, $MSE = .01$, $p < .0001$, and the control condition ($M = .32$, $SE = .01$), $F(1, 81) = 61.94$, $MSE = .01$, $p < .0001$, and that recall in the latter two conditions did not reliably differ ($F < 1$).

This pattern in overall veridical recall was also evident in the serial position curves depicted in Figure 1 in the main text, with the curves for the mixed and control conditions lying substantially on top of each other and the standard curve being offset from the other two curves at a higher rate that was fairly constant except for the last one or two serial positions.

Critical Word Intrusions

A between-subjects analysis of variance (ANOVA) revealed a reliable difference in critical word intrusions across the three groups, $F(2, 121) = 113.22$, $MSE = .03$, $p < .0001$. Planned comparisons indicated that the mean proportion of critical word intrusions in the standard condition ($M = .49$, $SE = .04$) was reliably higher than in both the mixed condition ($M = .07$, $SE = .02$), $F(1, 82) = 99.99$, $MSE = .04$, $p < .0001$, and the control condition ($M = .01$, $SE = .005$), $F(1, 81) = 152.97$, $MSE = .03$, $p < .0001$, and that the mixed mean was higher than the control mean, $F(1, 79) = 9.91$, $MSE = .01$, $p = .0023$. In addition, both the mixed and the control means differed reliably from zero, $t(40) = 4.03$, $p = .0002$, and $t(39) = 2.73$, $p = .0096$, respectively. Except for the trivially nonzero mean in the control condition, the pattern of results for critical word intrusions matched our expectations.

Prior-List Intrusions

A between-subjects ANOVA revealed a reliable difference among the groups in the number of prior-list intrusions produced during recall, $F(2, 121) = 19.31$, $MSE = .04$, $p < .0001$. Planned comparisons revealed that the mean number of prior-list intrusions in the mixed condition ($M = .33$, $SE = .04$) was reliably higher than the means in both the control condition ($M = .13$, $SE = .02$), $F(1, 79) = 15.91$, $MSE = .05$, $p < .0001$, and the standard condition ($M = .08$, $SE = .01$), $F(1, 82) = 25.72$, $MSE = .05$, $p < .0001$, and that the control mean was reliably higher than the standard mean, $F(1, 81) = 4.11$, $MSE = .01$, $p = .0459$. This pattern was consistent with our expectations.

Extralist Intrusions

Also consistent with our expectations, a between-subjects ANOVA revealed no reliable differences ($F < 1$) in the mean number of extralist intrusions produced in the standard condition ($M = .55$, $SE = .07$), the mixed condition ($M = .60$, $SE = .05$), and the control condition ($M = .55$, $SE = .08$).

(*Appendixes continue*)

Discussion

The experiment succeeded in generating the intricate pattern of means for veridical and false recall that we had expected on the basis of the strengths of the semantic associations among studied words on the same and different lists and between studied words and critical words. These results, combined with those from Kimball and Bjork (2002), created a rich set of data to constrain our models in Simulation 1.

Appendix B

Bayesian Information Criterion Values for Simulation 1

In Simulation 1, each version of the fSAM model was fit to veridical- and false-recall data drawn from Kimball and Bjork (2002, Experiment 2) and the three conditions in our new experiment (see Appendix A), using a single parameter set. Bayesian information criterion values for Simulation 1 are reported by experimental condition in Table B1 and by word type in Table B2.

Table B1
Bayesian Information Criterion Values by Experimental Condition for Simulation 1

	Semantic retrieval			
Semantic encoding	None	Single item	Additive rule	Multiplicative rule
A. All conditions combined				
None	−166	−214	−265	−335
Single item	−225	−243	−334	−356
Additive rule	−222	−240	−340	−336
Multiplicative rule	−283	−314	−348	−371
B. Kimball and Bjork (2002, Experiment 2)				
None	−16	−24	−54	−68
Single item	−17	−19	−47	−49
Additive rule	−17	−20	−44	−40
Multiplicative rule	−37	−41	−52	−57
C. Standard condition				
None	−24	−32	−44	−63
Single item	−39	−45	−77	−85
Additive rule	−37	−40	−74	−78
Multiplicative rule	−45	−65	−67	−70
D. Mixed condition				
None	−35	−43	−54	−70
Single item	−76	−74	−90	−78
Additive rule	−79	−98	−86	−93
Multiplicative rule	−91	−73	−89	−95
E. Control condition				
None	−41	−70	−54	−72
Single item	−101	−86	−62	−79
Additive rule	−83	−77	−77	−81
Multiplicative rule	−94	−70	−78	−92

Table B2
Bayesian Information Criterion Values by Word Type for Simulation 1

	Semantic retrieval			
Semantic encoding	None	Single item	Additive rule	Multiplicative rule
		A. Studied words		
None	−223	−281	−315	−245
Single item	−344	−291	−290	−271
Additive rule	−306	−294	−300	−290
Multiplicative rule	−296	−265	−286	−274
		B. Critical word		
None	1.3	−0.6	−15.3	−19.7
Single item	−0.9	−1.4	−25.1	−20.9
Additive rule	−0.4	−0.5	−21.9	−16.6
Multiplicative rule	−5.3	−10.8	−16.3	−18.7
		C. Extralist intrusions		
None	2.8	−0.2	1.7	−4.6
Single item	0.0	−0.5	−0.8	−3.7
Additive rule	−0.8	−0.9	−2.1	−1.9
Multiplicative rule	−5.0	−2.7	−3.3	−6.1
		D. Prior-list intrusions		
None	−6.1	−2.8	−4.4	−13.4
Single item	−8.2	−4.9	−12.1	−13.9
Additive rule	−7.9	−7.7	−6.8	−10.9
Multiplicative rule	−13.9	−9.3	−12.7	−16.5

(*Appendixes continue*)

Appendix C

Detailed Results for Simulation 1

In Simulation 1, each model version was fit to the combination of data drawn from Kimball and Bjork (2002, Experiment 2) and the three conditions of our new experiment (see Appendix A), using a single parameter set. The behavioral means and predicted values for veridical and false recall are reported in full in this appendix. Mean veridical recall is reported in Table C1, and veridical recall by serial position is reported in Tables C2, C3, C4, and C5 for data from Kimball and Bjork and the standard, mixed, and control conditions of our new experiment, respectively. Critical word intrusion rates and output percentiles are reported in Table C6. Extralist and prior-list intrusions are reported in Table C7.

Table C1
Mean Veridical Recall in Simulation 1

		Experimental condition			
		KB2	Standard	Mixed	Control
	Behavioral means	0.51	0.47	0.31	0.32
Semantic encoding	Semantic retrieval				
None	None	0.53	0.46	0.46	0.46
	Single item	0.48	0.38	0.32	0.33
	Additive rule	0.52	0.42	0.33	0.34
	Multiplicative rule	0.55	0.46	0.41	0.41
Single item	None	0.49	0.46	0.32	0.33
	Single item	0.47	0.40	0.30	0.31
	Additive rule	0.54	0.45	0.37	0.38
	Multiplicative rule	0.55	0.46	0.39	0.40
Additive rule	None	0.53	0.50	0.37	0.39
	Single item	0.49	0.40	0.33	0.34
	Additive rule	0.49	0.41	0.32	0.33
	Multiplicative rule	0.52	0.43	0.36	0.37
Multiplicative rule	None	0.52	0.41	0.33	0.33
	Single item	0.56	0.45	0.36	0.37
	Additive rule	0.51	0.40	0.32	0.33
	Multiplicative rule	0.48	0.39	0.32	0.32

Note. KB2 = Kimball & Bjork (2002, Experiment 2).

Table C2
Mean Veridical Recall by Serial Position for Kimball and Bjork (2002, Experiment 2) in Simulation 1

		Serial position														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Behavioral means	0.82	0.64	0.65	0.46	0.45	0.40	0.44	0.44	0.46	0.35	0.47	0.45	0.59	0.60	0.78
Semantic encoding	Semantic retrieval															
None	None	0.81	0.61	0.50	0.48	0.46	0.46	0.47	0.47	0.51	0.52	0.57	0.66	0.79	0.94	1.00
	Single item	0.80	0.53	0.41	0.41	0.39	0.38	0.46	0.42	0.46	0.46	0.52	0.66	0.78	0.94	1.00
	Additive rule	0.83	0.61	0.46	0.47	0.40	0.44	0.51	0.42	0.49	0.50	0.56	0.70	0.77	0.93	1.00
	Multiplicative rule	0.86	0.64	0.52	0.49	0.46	0.50	0.51	0.45	0.52	0.51	0.57	0.69	0.77	0.92	1.00
Single item	None	0.70	0.56	0.45	0.45	0.44	0.42	0.45	0.44	0.46	0.46	0.54	0.62	0.76	0.94	1.00
	Single item	0.72	0.49	0.41	0.42	0.37	0.43	0.46	0.42	0.45	0.46	0.53	0.67	0.77	0.92	1.00
	Additive rule	0.81	0.61	0.48	0.50	0.46	0.47	0.53	0.44	0.53	0.49	0.57	0.68	0.73	0.91	1.00
	Multiplicative rule	0.82	0.64	0.51	0.50	0.47	0.50	0.52	0.48	0.53	0.54	0.59	0.69	0.77	0.92	1.00
Additive rule	None	0.71	0.60	0.51	0.52	0.52	0.49	0.48	0.49	0.50	0.50	0.52	0.62	0.73	0.88	1.00
	Single item	0.74	0.52	0.43	0.44	0.39	0.42	0.45	0.44	0.51	0.49	0.57	0.67	0.78	0.93	1.00
	Additive rule	0.76	0.57	0.41	0.45	0.41	0.42	0.46	0.43	0.47	0.47	0.52	0.65	0.75	0.91	1.00
	Multiplicative rule	0.81	0.59	0.45	0.45	0.44	0.46	0.49	0.43	0.51	0.50	0.57	0.66	0.75	0.92	1.00
Multiplicative rule	None	0.80	0.63	0.46	0.48	0.46	0.46	0.48	0.44	0.49	0.49	0.52	0.65	0.77	0.95	1.00
	Single item	0.87	0.64	0.47	0.52	0.45	0.49	0.54	0.51	0.56	0.52	0.57	0.72	0.79	0.95	1.00
	Additive rule	0.82	0.61	0.42	0.47	0.41	0.44	0.48	0.43	0.48	0.49	0.52	0.68	0.76	0.92	1.00
	Multiplicative rule	0.81	0.58	0.40	0.40	0.39	0.39	0.46	0.42	0.46	0.46	0.54	0.67	0.76	0.90	1.00

Table C3
Mean Veridical Recall by Serial Position for the Standard Condition in Simulation 1

		Serial position														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Behavioral means	0.68	0.55	0.47	0.46	0.41	0.43	0.42	0.46	0.41	0.43	0.46	0.45	0.51	0.47	0.51
Semantic encoding	Semantic retrieval															
None	None	0.80	0.57	0.44	0.40	0.40	0.40	0.39	0.38	0.39	0.38	0.41	0.43	0.45	0.48	0.61
	Single item	0.67	0.44	0.34	0.32	0.32	0.33	0.32	0.30	0.33	0.33	0.34	0.35	0.35	0.41	0.54
	Additive rule	0.68	0.50	0.38	0.35	0.36	0.35	0.36	0.36	0.34	0.37	0.37	0.38	0.43	0.44	0.56
	Multiplicative rule	0.79	0.55	0.45	0.40	0.39	0.40	0.39	0.39	0.40	0.40	0.41	0.41	0.44	0.50	0.63
Single item	None	0.65	0.52	0.45	0.43	0.42	0.42	0.41	0.40	0.40	0.42	0.42	0.44	0.44	0.49	0.58
	Single item	0.60	0.45	0.36	0.37	0.35	0.35	0.34	0.36	0.35	0.37	0.36	0.38	0.39	0.43	0.53
	Additive rule	0.70	0.52	0.44	0.41	0.39	0.39	0.41	0.41	0.41	0.41	0.41	0.42	0.44	0.46	0.57
	Multiplicative rule	0.72	0.53	0.44	0.40	0.41	0.41	0.40	0.40	0.40	0.41	0.39	0.43	0.45	0.46	0.59
Additive rule	None	0.65	0.55	0.49	0.46	0.47	0.46	0.46	0.47	0.46	0.49	0.48	0.50	0.50	0.53	0.61
	Single item	0.62	0.46	0.37	0.36	0.36	0.35	0.34	0.35	0.35	0.37	0.37	0.38	0.38	0.41	0.52
	Additive rule	0.66	0.46	0.38	0.37	0.35	0.36	0.36	0.34	0.36	0.36	0.37	0.39	0.41	0.44	0.53
	Multiplicative rule	0.68	0.48	0.39	0.39	0.39	0.37	0.37	0.38	0.39	0.38	0.37	0.39	0.42	0.45	0.58
Multiplicative rule	None	0.70	0.51	0.40	0.36	0.36	0.36	0.36	0.34	0.36	0.36	0.35	0.39	0.40	0.44	0.53
	Single item	0.74	0.52	0.43	0.39	0.38	0.39	0.38	0.39	0.41	0.40	0.42	0.40	0.43	0.47	0.57
	Additive rule	0.69	0.46	0.38	0.36	0.35	0.35	0.35	0.34	0.35	0.36	0.36	0.38	0.40	0.42	0.51
	Multiplicative rule	0.70	0.46	0.38	0.35	0.32	0.32	0.33	0.32	0.33	0.33	0.34	0.36	0.40	0.43	0.52

(*Appendixes continue*)

Table C4
Mean Veridical Recall by Serial Position for the Mixed Condition in Simulation 1

		Serial position														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Behavioral means	0.52	0.41	0.29	0.25	0.25	0.24	0.23	0.24	0.29	0.28	0.23	0.29	0.36	0.38	0.43
Semantic encoding	Semantic retrieval															
None	None	0.78	0.58	0.43	0.39	0.39	0.40	0.38	0.39	0.41	0.39	0.41	0.42	0.45	0.48	0.62
	Single item	0.63	0.39	0.29	0.27	0.25	0.25	0.24	0.25	0.26	0.27	0.27	0.29	0.30	0.36	0.47
	Additive rule	0.64	0.41	0.32	0.27	0.27	0.26	0.26	0.27	0.28	0.28	0.27	0.31	0.32	0.38	0.50
	Multiplicative rule	0.79	0.52	0.38	0.34	0.34	0.32	0.33	0.33	0.35	0.33	0.34	0.37	0.40	0.43	0.57
Single item	None	0.52	0.39	0.32	0.27	0.27	0.26	0.26	0.26	0.28	0.28	0.28	0.30	0.32	0.38	0.47
	Single item	0.56	0.35	0.28	0.25	0.24	0.25	0.24	0.24	0.24	0.24	0.27	0.27	0.30	0.33	0.44
	Additive rule	0.65	0.45	0.35	0.32	0.31	0.30	0.33	0.31	0.32	0.32	0.32	0.33	0.35	0.40	0.51
	Multiplicative rule	0.69	0.46	0.37	0.33	0.31	0.31	0.33	0.33	0.32	0.33	0.33	0.35	0.38	0.43	0.54
Additive rule	None	0.54	0.44	0.35	0.33	0.32	0.32	0.31	0.33	0.31	0.33	0.33	0.36	0.37	0.42	0.51
	Single item	0.60	0.41	0.31	0.27	0.27	0.27	0.26	0.28	0.27	0.30	0.29	0.31	0.31	0.34	0.48
	Additive rule	0.62	0.38	0.30	0.27	0.26	0.27	0.26	0.26	0.26	0.25	0.25	0.30	0.30	0.34	0.47
	Multiplicative rule	0.66	0.40	0.34	0.31	0.29	0.29	0.32	0.29	0.32	0.30	0.31	0.33	0.36	0.38	0.52
Multiplicative rule	None	0.71	0.43	0.32	0.27	0.24	0.25	0.25	0.24	0.25	0.24	0.28	0.28	0.32	0.37	0.51
	Single item	0.75	0.43	0.35	0.29	0.29	0.26	0.28	0.29	0.29	0.30	0.29	0.32	0.34	0.38	0.53
	Additive rule	0.69	0.39	0.30	0.26	0.25	0.25	0.25	0.25	0.27	0.26	0.27	0.28	0.32	0.34	0.48
	Multiplicative rule	0.71	0.39	0.27	0.24	0.24	0.25	0.24	0.25	0.26	0.25	0.25	0.29	0.31	0.36	0.49

Table C5
Mean Veridical Recall by Serial Position for the Control Condition in Simulation 1

		Serial position														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Behavioral means	0.49	0.38	0.34	0.29	0.24	0.25	0.24	0.23	0.26	0.29	0.31	0.34	0.36	0.39	0.48
Semantic encoding	Semantic retrieval															
None	None	0.77	0.55	0.43	0.39	0.41	0.40	0.40	0.38	0.39	0.39	0.40	0.40	0.45	0.50	0.63
	Single item	0.65	0.39	0.30	0.28	0.26	0.25	0.26	0.27	0.26	0.26	0.28	0.28	0.33	0.35	0.50
	Additive rule	0.63	0.41	0.32	0.29	0.29	0.28	0.28	0.27	0.28	0.30	0.30	0.31	0.32	0.40	0.49
	Multiplicative rule	0.78	0.50	0.37	0.36	0.33	0.33	0.35	0.33	0.34	0.35	0.36	0.37	0.39	0.43	0.59
Single item	None	0.54	0.39	0.30	0.29	0.29	0.28	0.28	0.27	0.28	0.29	0.27	0.31	0.36	0.38	0.50
	Single item	0.56	0.36	0.30	0.27	0.26	0.26	0.25	0.27	0.26	0.27	0.27	0.28	0.30	0.34	0.46
	Additive rule	0.68	0.46	0.37	0.34	0.32	0.29	0.33	0.31	0.31	0.32	0.34	0.34	0.38	0.41	0.53
	Multiplicative rule	0.70	0.49	0.38	0.33	0.33	0.33	0.33	0.33	0.34	0.34	0.35	0.36	0.37	0.41	0.56
Additive rule	None	0.57	0.45	0.36	0.34	0.34	0.33	0.33	0.34	0.33	0.35	0.35	0.36	0.39	0.43	0.53
	Single item	0.62	0.41	0.32	0.29	0.29	0.29	0.28	0.27	0.29	0.28	0.29	0.31	0.35	0.38	0.49
	Additive rule	0.63	0.41	0.30	0.26	0.27	0.27	0.29	0.28	0.28	0.28	0.29	0.30	0.32	0.36	0.45
	Multiplicative rule	0.68	0.44	0.32	0.32	0.29	0.30	0.31	0.31	0.32	0.33	0.32	0.33	0.37	0.40	0.53
Multiplicative rule	None	0.70	0.44	0.31	0.27	0.25	0.26	0.26	0.25	0.28	0.25	0.28	0.30	0.33	0.36	0.50
	Single item	0.77	0.45	0.33	0.32	0.28	0.30	0.29	0.29	0.28	0.29	0.31	0.32	0.34	0.39	0.54
	Additive rule	0.70	0.42	0.30	0.28	0.26	0.25	0.25	0.28	0.25	0.26	0.27	0.30	0.32	0.35	0.48
	Multiplicative rule	0.70	0.40	0.29	0.26	0.26	0.26	0.24	0.25	0.26	0.26	0.27	0.28	0.31	0.36	0.49

Table C6
Mean Number of Critical Word Intrusions per List and Mean Critical Word Output Percentile in Simulation 1

		Critical word intrusions				Critical word output percentile			
		KB2	Standard	Mixed	Control	KB2	Standard	Mixed	Control
	Behavioral means	0.54	0.49	0.07	0.01	66	63	75	80
Semantic encoding	Semantic retrieval								
None	None	0.0004	0.0003	0.0027	0.0043	89	50	72	76
	Single item	0.1538	0.1567	0.2323	0.0000	81	63	64	
	Additive rule	0.5475	0.4400	0.1877	0.0003	83	68	66	100
	Multiplicative rule	0.4917	0.4083	0.0630	0.0033	78	60	43	50
Single item	None	0.1496	0.1333	0.0420	0.0067	76	65	68	74
	Single item	0.2063	0.2173	0.1323	0.0053	76	59	65	63
	Additive rule	0.5592	0.4727	0.0853	0.0057	77	66	73	62
	Multiplicative rule	0.4992	0.4293	0.0750	0.0053	75	59	50	69
Additive rule	None	0.1371	0.1137	0.0460	0.0067	78	66	62	65
	Single item	0.1921	0.1747	0.0987	0.0043	78	61	65	62
	Additive rule	0.5371	0.4777	0.0980	0.0043	81	68	72	63
	Multiplicative rule	0.4692	0.4020	0.0723	0.0047	77	59	57	67
Multiplicative rule	None	0.3346	0.1910	0.0637	0.0063	78	63	62	44
	Single item	0.4296	0.3363	0.1127	0.0057	75	58	64	66
	Additive rule	0.5975	0.4003	0.1003	0.0060	79	66	68	55
	Multiplicative rule	0.5425	0.4030	0.0567	0.0037	78	61	61	58

Note. KB2 = Kimball & Bjork (2002, Experiment 2).

Table C7
Mean Number of Extralist and Prior-List Intrusions per List in Simulation 1

		Extralist intrusions				Prior-list intrusions			
		KB2	Standard	Mixed	Control	KB2	Standard	Mixed	Control
	Behavioral means	0.32	0.55	0.60	0.55	0.01	0.08	0.33	0.13
Semantic encoding	Semantic retrieval								
None	None	0.14	0.18	0.17	0.16	0.15	0.16	0.17	0.17
	Single item	0.18	0.22	0.69	0.37	0.03	0.05	0.62	0.11
	Additive rule	0.16	0.15	0.60	0.27	0.03	0.05	0.57	0.10
	Multiplicative rule	0.33	0.33	0.65	0.61	0.06	0.07	0.39	0.14
Single item	None	0.71	0.62	0.49	0.48	0.11	0.11	0.22	0.10
	Single item	0.63	0.49	0.60	0.44	0.14	0.12	0.46	0.13
	Additive rule	0.52	0.39	0.57	0.37	0.07	0.06	0.35	0.08
	Multiplicative rule	0.50	0.45	0.66	0.59	0.06	0.07	0.37	0.13
Additive rule	None	0.67	0.53	0.48	0.45	0.15	0.15	0.30	0.16
	Single item	0.60	0.47	0.58	0.43	0.12	0.11	0.38	0.08
	Additive rule	0.51	0.40	0.59	0.42	0.12	0.11	0.43	0.14
	Multiplicative rule	0.57	0.45	0.61	0.48	0.09	0.09	0.36	0.10
Multiplicative rule	None	0.52	0.48	0.53	0.54	0.06	0.12	0.35	0.15
	Single item	0.53	0.46	0.59	0.44	0.06	0.10	0.42	0.10
	Additive rule	0.46	0.40	0.60	0.44	0.05	0.08	0.39	0.11
	Multiplicative rule	0.43	0.43	0.57	0.55	0.05	0.09	0.33	0.14

Note. KB2 = Kimball & Bjork (2002, Experiment 2).

(*Appendixes continue*)

Appendix D

Parameter Values

This appendix reports the free parameter values for the simulations reported in this article. Table D1 reports the values for Simulation 1, Table D2 those for Simulation 1A, and Table D3 those for Simulations 2 and 3. Unless noted otherwise, the following parameters were fixed in all simulations: the increment in backward interitem strength during study ($b_2 = 0.5*b_1$) and during recall ($f_2 = 0.5*f_1$), the distribution of short-term memory buffer size across lists and subjects ($\mu_r = 4$, $\sigma_r = 1.4$), the factor biasing displacement of older items from the short-term memory buffer at encoding ($q = 0.266$), the distribution of the default values for episodic and contextual strength ($\mu = 0.001$, $\sigma = 0.0005$), and the maximum number of retrieval failures using a particular set of cues ($L_{max} = 0.1*K_{max}$). These fixed parameters and their values were all inherited from the simulations reported in Sirotin et al. (2005), which were conducted using the eSAM model, a predecessor of fSAM.

Table D1
Parameter Values for the Best Fit of Each Version of the fSAM Model in Simulation 1

		Encoding			Retrieval				Output encoding		
Semantic encoding	Semantic retrieval	a	a_s	b_1	W_c	W_e	W_s	K_{max}	e	f_1	Forgetting (ρ)
None	None	0.46		0.00	0.03	1.18		182	0.93	0.40	0.0359
	Pairwise	0.68		0.09	0.05	0.39	82.1	25	1.07	0.56	0.0025
	Additive	1.07		0.36	0.12	0.05	113.9	23	0.30	1.22	0.0225
	Multiplicative	0.62		0.09	0.07	0.18	188.5	62	0.93	1.49	0.0205
Pairwise	None	0.12	0.03	1.19	0.97	0.26		77	0.28	0.41	0.0035
	Pairwise	0.93	0.10	0.14	0.67	0.43	59.6	28	0.74	0.87	0.0184
	Additive	1.06	0.08	0.61	0.91	0.18	296.0	45	1.49	1.33	0.0001
	Multiplicative	1.04	0.06	0.56	0.17	0.02	227.1	45	1.43	0.17	0.0021
Additive	None	0.22	0.01	0.59	0.82	0.65		38	0.05	0.36	0.0303
	Pairwise	1.02	0.03	0.25	0.49	0.16	183.5	30	0.30	0.91	0.0020
	Additive	0.64	0.02	0.14	0.77	0.20	259.0	44	0.78	0.35	0.0328
	Multiplicative	0.91	0.02	1.06	0.13	0.22	265.1	39	1.32	1.46	0.0136
Multiplicative	None	0.35	0.41	0.07	0.26	1.59		36	0.07	1.07	0.0024
	Pairwise	0.75	0.46	0.06	0.61	0.16	109.4	51	0.93	0.34	0.0073
	Additive	0.75	0.47	0.99	0.20	0.73	263.5	34	1.16	1.30	0.0084
	Multiplicative	0.44	0.21	0.20	0.67	0.29	209.9	39	0.39	0.84	0.0287

Note. Each version of the fSAM model was fit using a single parameter set to the combination of data drawn from Kimball and Bjork (2002, Experiment 2) and the three conditions of our new experiment (see Appendix A).

Table D2
Parameter Values for the Best Fit of Each Version of the fSAM Model in Simulation 1A

		Encoding			Retrieval				Output encoding		
Semantic encoding	Semantic retrieval	a	a_s	b_1	W_c	W_e	W_s	K_{max}	e	f_1	Forgetting (ρ)
None	None	0.43		0.02	0.01	1.14		187	0.07	1.47	0.0448
None	Pairwise	1.15		1.24	0.03	0.01	203.1	63	0.69	0.40	0.0216
None	Additive	0.86		1.28	0.47	0.15	273.4	141	1.04	1.35	0.0064
None	Multiplicative	0.87		0.83	0.12	0.10	183.2	88	0.08	0.44	0.0349
Pairwise	None	0.69	0.08	0.83	1.00	0.08		88	0.33	0.56	0.0062
Additive	None	0.38	0.03	0.70	0.94	0.19		51	0.29	0.84	0.0120
Multiplicative	None	0.46	0.38	0.31	0.97	0.07		41	0.74	0.26	0.0307

Note. Each model was fit using a single parameter set to the combination of data drawn from Kimball and Bjork (2002, Experiment 2) and the three conditions of our new experiment (see Appendix A); critical word intrusions were weighted 100 times as much as the fit of veridical recall or other intrusions.

Table D3
Parameter Values for the Best Fit of the fSAM Multiplicative Model in Simulations 2 and 3

			Encoding			Retrieval				Output encoding			Mean STM size
Simulation	Data simulated	Age group	a	a_s	b_1	W_c	W_e	W_s	K_{max}	e	f_1	Forgetting (ρ)	(r)
2	Brainerd, Reyna, & Forrest (2002)	5-yos	0.13	0.002	0.07	0.24	0.01	139.8	14	0.05	0.32	0.3237	1.0
		7-yos	0.18	0.005	0.09	0.37	0.04	183.2	20	0.06	0.49	0.1506	1.1
		11-yos	0.18	0.029	0.18	0.50	0.06	195.7	21	0.13	0.67	0.0653	3.3
3	Gallo & Roediger (2002)		1.37	0.134	1.24	0.12	0.59	9.7	41	1.21	1.28	0.1768	4.0
	Stadler, Roediger, & McDermott (1999)		0.69	0.064	0.78	0.86	1.49	260.6	44	2.86	1.17	0.0165	4.0

Note. In Simulation 2, the fSAM multiplicative model was fit to developmental data from Brainerd, Reyna, and Forrest (2002, Experiments 2 and 3). In Simulation 3, the model was fit to list-effect data for the 19 unique lists in Gallo and Roediger (2002) and the 36 lists in Stadler, Roediger, and McDermott (1999) that were combined in the multiple regression analysis performed by Roediger, Watson, McDermott, and Gallo (2001). STM = short-term memory buffer; yos = year-olds.

Received July 20, 2006
Revision received April 5, 2007
Accepted April 6, 2007 ■