Categorical *and* associative relations increase false memory relative to purely associative relations

Jennifer H. Coane¹ · Dawn M. McBride² · Miia-Liisa Termonen¹ · J. Cooper Cutting²

Published online: 7 August 2015
© Psychonomic Society, Inc. 2015

Abstract The goal of the present study was to examine the contributions of associative strength and similarity in terms of shared features to the production of false memories in the Deese/Roediger–McDermott list-learning paradigm. Whereas the activation/monitoring account suggests that false memories are driven by automatic associative activation from list items to nonpresented lures, combined with errors in source monitoring, other accounts (e.g., fuzzy trace theory, global-matching models) emphasize the importance of semantic-level similarity, and thus predict that shared features between list and lure items will increase false memory. Participants studied lists of nine items related to a nonpresented lure. Half of the lists consisted of items that were associated but did not share features with the lure, and the other half included items that were equally associated but also shared features with the lure (in many cases, these were taxonomically related items). The two types of lists were carefully matched in terms of a variety of lexical and semantic factors, and the same lures were used across list types. In two experiments, false recognition of the critical lures was greater following the study of lists that shared features with the critical lure, suggesting that similarity at a categorical or taxonomic level contributes to false memory above and beyond associative strength. We refer to this phenomenon as a “feature

boost” that reflects additive effects of shared meaning and association strength and is generally consistent with accounts of false memory that have emphasized thematic or feature-level similarity among studied and nonstudied representations.

Keywords False memory · Association strength · Categorical similarity · Feature overlap

False memories for nonstudied words can be reliably elicited using an experimental task known as the *Deese/Roediger–McDermott (DRM) paradigm*. In this paradigm, originally developed by Deese (1959) to examine the role of interitem relatedness in free recall, and revived by Roediger and McDermott (1995), participants study lists of words (e.g., *bed, rest, tired*) related to a single nonpresented word, hereafter referred to as the critical lure (CL; e.g., *sleep*). Participants falsely recall or recognize the CL at high rates; indeed, the levels of false recall and false recognition are often comparable to veridical recall and recognition rates (see Gallo, 2006, for a review). In addition to its validity as a measure of the malleability of memory, the DRM paradigm can also enhance theoretical understanding of the organization of the memory systems supporting semantic processes (e.g., Buchanan, Brown, Cabeza, & Maitson, 1999; Huff, Coane, Hutchison, Grasser, & Blais, 2012; Huff & Hutchison, 2011; Hutchison & Balota, 2005).

Although hundreds of studies have been published since Roediger and McDermott’s (1995) article, relatively few studies have investigated factors directly related to the DRM lists themselves. A better understanding of the types of relations, broadly defined, between list items and CLs that increase or decrease false memory is critical in terms of theory development and to predict when an intrusion error or false alarm is most likely to occur. In other words, what is the nature of the

Electronic supplementary material The online version of this article (doi:10.3758/s13421-015-0543-1) contains supplementary material, which is available to authorized users.

✉ Jennifer H. Coane
jhcoane@colby.edu

¹ Department of Psychology, Colby College, 5550 Mayflower Hill Drive, Waterville, ME 04901, USA

² Illinois State University, Normal, IL, USA

mental representations most likely to elicit a false memory? To examine this, Roediger, Watson, McDermott, and Gallo (2001) performed a multiple regression analysis on a set of variables presumed to influence false memory. The main predictor of false memory was backward associative strength (BAS), which is a measure of the probability with which a list item will elicit the CL on a free association task. Lists with higher mean BAS resulted in higher levels of false recall than did lists with lower mean BAS. In addition, veridical recall was negatively associated with false recall. Gallo and Roediger (2002) developed lists with low average BAS (i.e., weak lists), which resulted in lower rates of false recall and false recognition than did lists with higher BAS, whereas veridical recall and recognition did not differ across the strong and weak lists.

The evidence that BAS predicts false memories is consistent with spreading activation network accounts of semantic processing (Anderson, 1983; Collins & Loftus, 1975; Steyvers & Tenenbaum, 2005). According to the dual-process activation/monitoring theory (AMT; Roediger, Balota, & Watson, 2001), false memories in the DRM paradigm are due to activation spreading from the list items to the CL through semantic and associative networks. Closely related items (i.e., strong associates) send more activation than weak associates. The activation converging on the CL increases its accessibility or familiarity, and source-monitoring errors result in incorrect “old” responses, or intrusions.

Although the effect of BAS on false memory is well-established (Hutchison & Balota, 2005; McEvoy, Nelson, & Komatsu, 1999; Roediger, Watson, et al., 2001), there is still a question as to why some lists are more likely to elicit false memory than others. For example, in Roediger, Watson, et al.’s study, the *king* list, with a mean BAS of .23 resulted in a false-recall rate of .10, whereas the *smoke* list, with a mean BAS of .17, yielded a false-recall rate of .54. Clearly, factors other than BAS are involved. The question that we addressed here was whether the type of relationship between list items and CLs affects false memory. Specifically, we examined the role of shared features between list items and CLs. In many cases, items are both semantically and associatively related (e.g., *cat* and *dog* are related both by feature overlap and by associative norms); however, some items are “purely” semantically (e.g., *dog* and *goat*) or “purely” associatively (e.g., *dog* and *leash*) related. The broader theoretical question is the extent to which false memories depend on the extraction of shared meaning at the semantic level or on lexical-level associations between list items and the CL. This issue has been extensively debated in the field of semantic memory and semantic priming (e.g., Hutchison, 2003; Lucas, 2000; McNamara, 2005), and it pertains to important issues regarding the organization of knowledge structures that support semantic and episodic memory.

One of the most influential models of semantic memory, the spreading activation framework described by Collins and Loftus (1975), assumed that activation between concepts spread as a result of associative and taxonomic relatedness and that the number of shared features between two nodes in the network determined their proximity and thus the activation. The model also incorporated lexical-level information in which activation can spread along pathways determined by factors other than semantic similarity (e.g., phonological information) suggesting potential additive effects as a result of multiple sources of activation converging on a single node.

Along these lines, Watson, Balota, and Roediger (2003) examined contributions to false memory from lexical and semantic factors by creating hybrid lists of semantic and phonological/orthographic associates. For example, for the CL *dog*, a hybrid list included items such as *puppy* and *hound* as well as *log* and *dodge*. Phonological/orthographic similarity reflects activation in lexical-level networks, whereas semantic/associative similarity reflects activation in both lexical and semantic networks. Compared to pure lists of semantic or phonological/orthographic associates, hybrid lists yielded overadditive false memory, suggesting that lexical-level similarity combines with conceptual-level relatedness to increase the accessibility of information in semantic memory (see also Rubin & Wallace, 1989).

A key assumption of AMT (Roediger, Balota, & Watson, 2001), which posits that BAS is the determining factor in eliciting false memories, is that the activation is directional, spreading from the list items to the CL (see Arndt, 2012). Furthermore, BAS as a metric does not assume any similarity at the level of semantic representations, but is merely a reflection of the strength of associations in memory, with some strong associates also being highly similar (e.g., *cat* and *dog* have an association strength of .51 according to the University of South Florida Free Association Norms; Nelson, McEvoy, & Schreiber, 1998) and other strong associates reflecting different types of relations (e.g., *bark* and *dog* have an association strength of .56). Thus, examining BAS without a consideration of how the shared features or semantic similarity might vary across lists might be masking some independent effects of shared semantic similarity. Another issue is the difficulty inherent in isolating “pure” association from semantic similarity.

Although a review of the semantic priming literature is outside of the scope of this article, it is important to note that there is evidence for “pure” associative priming between items that do not share any features (Balota & Lorch, 1986; Hutchison, 2003). Interestingly, when category coordinates or items related through shared features (e.g., *goat–dog*) are used in semantic-priming paradigms, prime–target pairs that are also associatively related (e.g., *cat–dog*) result in larger priming effects, a phenomenon referred to as the “associative boost” (see Hutchison, 2003). Thus, converging evidence

from semantic priming paradigms suggests that in both priming and false memory paradigms, associative activation is a critical process and that multiple sources of activation, be they associative and semantic or conceptual and phonological/orthographic (e.g., Watson et al., 2003), yield additive effects in memory and priming tasks.

Clearly, shared meaning, regardless of associative strength, plays an important role in many episodic memory phenomena. For example, recall output for word lists often reflects clustering at the level of shared category membership, with participants recalling items from the same category at levels greater than chance (e.g., Bousfield, 1953). In the classic level-of-processing paradigm, attending to the meaning of an item, relative to attending to surface characteristics, promotes better retention (e.g., Craik & Lockhart, 1972; Craik & Tulving, 1975). According to an alternative explanation of false memories, namely fuzzy trace theory (FTT; Brainerd & Reyna, 2001, 2002), meaning extraction is also critical for false remembering. According to FTT, memory assessments are based on both verbatim representations, which include information such as perceptual details, and gist representations, which depend on the meaning of the item or list. Veridical retrieval of studied items can be supported by both verbatim and gist traces, whereas false memory for lures depends on the gist trace alone, because no verbatim trace is available for these items (but see Lampinen, Meier, Arnal, & Leding, 2005). The gist trace is assumed to be dependent on a shared theme or meaning, and, as a result, when the lists have a strong convergence on a shared theme, false memories are expected to be greater (e.g., Arndt, 2012).

However, Hutchison and Balota (2005) provided evidence that associative strength is a better predictor of false memory than is thematic coherence or gist. They compared lists that converged on a single theme (i.e., typical DRM lists) to homographic lists converging on two themes (e.g., a list that contained items related to both meanings of the CL *fall*). According to accounts that assign a significant role to thematic coherence, the homographic lists should have resulted in reduced false memory; however, in recall and recognition, false memory rates were equivalent across list types, suggesting that BAS, which was matched across list types, not shared meaning, was the critical determinant of false memory. Furthermore, DRM-type lists that consist of items only indirectly related to the CL through nonpresented mediators also result in reliable false memory—a compelling finding, given these lists have no apparent gist or thematic coherence (Huff & Hutchison, 2011; Huff et al., 2012). In these studies, the mediated list items were directly related to the original DRM list items, but unrelated to the CL. For example, for the CL *river*, the list included such items as *faucet* (related to *water*) and *paddle* (related to *canoe*). These results suggest that meaning extraction may be less critical for false memory than the simple spread of activation along associative links, an

automatic and relatively “passive” process (cf. Roediger, Balota, & Watson, 2001).

This conclusion, that associative links are driving false memory, with less involvement of shared meaning, suggests that similarity between list items and CLs at the level of meaning may be less important than associative strength. Although the majority of lists used in most studies have contained a combination of the two types of associates, it is possible to manipulate the type of items appearing in a list such that they do or do not share features (i.e., are semantically or associatively related to the CL). The question that we address here is the role of shared features, which taps into semantic or meaning-based relations, between list items and the CL.

In a similar study, Buchanan et al. (1999) presented participants with lists of categorically or associatively related items. For example, for the critical lure *apple*, the categorical list included *orange*, *banana*, and *pear*, and the associative list included *pie*, *tree*, and *grandma*. Associative lists resulted in higher rates of false recognition. Smith, Gerkens, Pierce, and Choi (2002) also used category-based and DRM lists to examine indirect priming effects as a measure of associative responses and only obtained priming for the DRM lists, suggesting underlying differences between the list types. Conversely, Dewhurst and colleagues (e.g., Dewhurst, Barry, Swannell, Holmes, & Bathurst, 2007; Dewhurst, Bould, Knott, & Thorley, 2009; Knott & Dewhurst, 2007) have consistently found that manipulations that affect activation processes (e.g., divided attention, blocking vs. randomized presentation) during encoding exert parallel effects on explicit memory tasks with both types of lists. In all of these studies, the associative lists had higher BAS than the categorical lists and, not surprisingly, resulted in overall higher rates of false memory and priming, consistent with AMT (Roediger, Balota, & Watson, 2001).

In a recent study, Knott, Dewhurst, and Howe (2012) developed associative and categorical lists that were matched on BAS. They orthogonally manipulated BAS (high vs. low) and connectivity (the strength of interitem associations in the list, which is negatively correlated with false recall; high vs. low). False recall and recognition did not differ across list type and were highest when BAS was high and connectivity was low for both the categorical and associative lists (see also McEvoy et al., 1999). The equivalent false memory rate across list types, when they were matched on BAS, further underscores the importance of this variable. One limitation in Knott et al.’s study, however, was that different CLs were used across the two types of lists, thus raising the question of whether item-specific differences between CLs might have affected the results (see Neely & Tse, 2007). A second limitation of Knott et al.’s study regards the list composition. Their categorical lists consisted primarily, but not exclusively, of category coordinates (e.g., the *chair* list consisted of such items as *table*, *sofa*, and *recliner*, but also included *furniture*, which could be

considered the category superordinate). Importantly, these lists did have a high degree of feature overlap and clearly came from well-defined categories. However, their associative lists included a mixture of associates (in the *chair* list, items such as *sit* and *wood*) and category coordinates (e.g., *table*, *sofa*). Thus, these lists were not “pure” associative lists, but had a high degree of feature overlap, and several list items were included in both types of lists, making it difficult to isolate the role of association from that of shared features.

In sum, the results of previous studies in this area have suggested that (1) both categorically related and associatively related lists do elicit reliable false-memory rates; (2) both types of lists respond similarly to experimental manipulations, suggesting a common locus of the effect; and (3) association strength is a powerful predictor of false memory. However, it has not been clear from these studies whether shared meaning, as defined by sharing features and/or category membership, contributes to false memory above and beyond association strength. Evidence from different paradigms has suggested that one should obtain additive effects from BAS and semantic similarity, resulting in higher error rates to CLs related both associatively and categorically/semantically to the list items. As was noted above, specifying the contribution of meaning extraction or relatedness in terms of underlying meaning in the DRM and other episodic memory tasks is critical for theory development and for determining the types of mental representations most likely to elicit errors.

In the present study, we held associative strength constant and varied the amount of feature overlap. Thus, we developed two types of lists for each CL: Categorical + associative (C+A) lists consisted of items that shared features *and* were generated on a free association task, whereas noncategorical associatively related (NC-A) lists included items that were generated on free association tasks but did not share obvious features. The lists were developed such that mean BAS was equated across the two types of lists; thus, any differences in false memory across the lists would not be due to differences in associative strength, but to differences in the types of relations between list items and CLs. Importantly, because we used the same CLs across both list types, we could rule out idiosyncratic item-level effects (see Neely & Tse, 2007).

If false memories in the DRM paradigm are due to activation and if activation spreads along associative networks independently of the types of relationships between items, we would expect to find no difference between the lists (cf. Hutchison & Balota, 2005), because BAS was matched. However, if feature overlap contributes an independent amount of activation, then we would expect to find higher rates of false memories when the lists were not only associatively related but also shared features (cf. Watson et al., 2003). According to FTT (Brainerd & Reyna, 2002), C+A lists should result in higher error rates than NC-A lists because of stronger similarity, which should facilitate gist extraction.

Experiment 1

In Experiment 1, veridical and false recall and recognition rates were compared for C+A (categorically and associatively related) and NC-A (associatively related, but without shared features) lists to test the predictions described above. Participants completed a free recall test after the presentation of each list and then completed a final recognition test after all lists had been presented and recalled.

Method

Participants Participants were recruited from the psychology department participant pools at Illinois State University ($n = 40$) and Colby College ($n = 40$). All were native speakers of English and had normal or corrected-to-normal vision. Participants received \$5 or course credit. An additional 1,079 participants participated in the norming session conducted online (see the Materials section).

Materials The lists were developed using the Nelson et al. (1998) free association norms. The initial step involved identifying potential CLs that had a large number of associates. For the C+A lists, list items were selected that belonged to the same semantic category as the CL (e.g., *horse–donkey*), shared perceptual features (e.g., *road–highway*), or were synonyms or near synonyms of the CL (e.g., *cut–chop*). For the NC-A lists, the items were selected such that they were associatively related but did not share obvious features and were not synonymous with the CL (e.g., *horse–stable*, *road–map*, *cut–grass*).

After identifying 100 potential CLs, a further screening was performed. First, only lists with a mean BAS of .10 or more were selected. Next, items that appeared in more than one list were eliminated. Finally, 20 lists with nine items of each type were selected, such that the mean BASs of both list types across all lists were equivalent. The mean BAS for C+A lists was .239 ($SEM = .023$) and that for NC-A lists was .245 ($SEM = .019$). Across all lists, the list items were matched on several lexical characteristics, including word length, word frequency, two measures of orthographic neighborhood (a measure of item distinctiveness), and lexical decision reaction times and accuracy from the English Lexicon Project (Balota et al., 2007). These variables are predictive of word recognition times (and hence of processing time; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), and several of the factors—in particular, frequency and distinctiveness—also affect memory performance directly (e.g., Coane, Balota, Dolan, & Jacoby, 2011; Glanzer & Adams, 1985; Hunt, 1995; Hunt & McDaniel, 1993). In addition, the two types of lists were matched on two measures of semantic similarity between the list items and CLs (i.e., latent semantic analysis [LSA]

cosines, Landauer & Dumais, 1997; and pointwise mutual information [PMI], Recchia & Jones, 2009). These metrics capitalize on large-scale computational analyses of extensive linguistic corpora and provide measures of the broader linguistic context in which words occur. Briefly, LSA captures the intercorrelations between words from a large text database, such that the meaning of a word is influenced by the contexts (i.e., neighbors) in which that word occurs, as well as by the contexts and experiences of the neighbors. Semantic similarity, as we noted, can influence list memory; thus, it was important to match the lists on this variable. PMI (Recchia & Jones, 2009) is a metric that calculates the probability of two items occurring together (in a single document), relative to the probability of them occurring separately in the entire Wikipedia corpus. Thus, this measure provides a way of quantifying how likely it is that two items co-occur, given their independent frequencies in the language. In both metrics, higher values reflect more co-occurrence or similarity. To calculate the LSA and PMI values, individual pairwise comparisons between each list item and the respective CL were calculated, and these values were then averaged for each type of list.

To further ensure that the word pairs differed only in their relationship type (C+A or NC-A), a norming study was conducted using Amazon's Mechanical Turk (MTurk; Amazon.com, Inc., <https://www.mturk.com/mturk/welcome>) worker pool. MTurk has been established as a participant pool providing data comparable to those collected in a laboratory setting (Mason & Suri, 2012). Participants were compensated \$1.05–\$1.25 for completing a rating task that took on average about 5 min ($M = 295$ s). The stimuli were divided into sets of 42 pairs, such that each participant saw 21 CLs (including *rain*, which was later dropped from the experimental set) paired with two different list items. In each set, each CL appeared once with a C+A list item (e.g., *wolf-dog*) and once with an NC-A list item (e.g., *leash-dog*). The pairs were presented in a pseudorandom order, such that the two presentations of each CL were not contiguous. We modified four different rating scales from Jones and Golonka (2012): categorical relatedness, thematic relatedness, feature similarity, and familiarity. The instructions for the categorical-relatedness task required participants to rate the items in each pair on the basis of the extent to which they came from the same category. In the thematic-relatedness task, participants rated the extent to which the items occurred together in a scenario or event. The feature similarity rating task required participants to rate the items in terms of similarity across features. Finally, the familiarity rating involved a judgment of how familiar each pair of words was. Examples were provided with all instructional sets. All ratings were made using a 7-point Likert scale, with 1 being *not at all categorically related/thematically related/similar/familiar*, and 7 denoting *definitely share a category/theme or very similar/familiar*.

Of the 1,079 respondents in the rating study, 144 were rejected for a number of reasons: not meeting the age limit of 18–28 years, being too fast to be reasonably able to perform the study (e.g., finishing in 93 s when the group's mean completion time was 289 s), or giving the same response for all pairs (e.g., rating everything a 4). After omitting these data sets, we had 935 rating sets, with between 25 and 31 ratings for each word pair on each of the four rating scales.¹

Rating data were then analyzed as a function of list type. The C+A lists did not differ significantly in thematic similarity ($p = .20$) or in familiarity ($p = .10$), but C+A pairs were rated significantly higher than NC-A pairs on both feature similarity ($p < .001$) and categorical similarity ($p < .001$). Thus, both word pair types shared contexts and co-occurred in language often enough to be familiar as a pair, but C+A word pairs shared more features and were rated higher on belonging to the same category than were NC-A pairs, thus confirming that the main dimension along which these items differed was their feature overlap and shared category membership with the CL.

See Table 1 for the full descriptive characteristics of the lists, and the [supplemental materials](#) for a list of all stimuli and the item-level ratings.

Procedure Participants were tested individually or in small groups (at individual computer stations). They were instructed to study the words for a memory test. Each participant studied ten lists (five C+A and five NC-A) of nine words each, presented one at a time for 1,000 ms, with a 500-ms interstimulus interval. The lists were presented in a randomized order. After each list, the participant worked on an arithmetic problem filler task for 30 s. A tone indicated the end of the filler task, and participants were asked to write down on a sheet of paper all of the words that they could recall from that list. They were given 1 min for free recall, and then they pressed a key on the keyboard to begin the next list. After all ten lists had been presented and recalled, a surprise final recognition task followed. The test included ten CLs from the studied lists and 20 of the list items that participants had seen (two items from each list, from Serial Positions 3 and 7). In addition, ten control CLs and 20 control list items from the ten unstudied lists were included in the recognition test. The participants were asked to press the “Y” key for “yes,” if they remembered seeing the word in the study phase, and the “N” key for “no,” if they had not seen it. The word remained on the screen until the participant had made a response. The lists were counterbalanced across participants such that each list and

¹ A more conservative screening criterion, in which we omitted participants who gave ratings of high similarity to NC-A items on the measures of categorical and feature relatedness ($ns = 9$ and 26, respectively), yielded virtually identical results.

Table 1 Listwise lexical characteristics of the Deese/Roediger–McDermott and categorical lists used in Experiments 1 and 2

Measure	List Type		<i>p</i> Value
	C+A	NC-A	
BAS ^a	.24 (.02)	.24 (.02)	.84
Length	5.57 (.22)	5.86 (.20)	.34
HAL log frequency ^b	8.82 (.18)	8.72 (.16)	.68
SUBTLEX log frequency ^b	2.83 (.08)	2.80 (.06)	.78
SUBTLEX log contextual diversity ^b	2.57 (.07)	2.57 (.06)	.99
Orthographic N ^b	5.54 (.58)	5.24 (.62)	.73
OLD ^b	2.07 (.08)	2.03 (.08)	.76
OLD frequency ^b	7.60 (.08)	7.47 (.08)	.27
Lexical decision RT ^b	542.50 (5.32)	549.33 (6.63)	.43
Lexical decision accuracy ^b	.96 (.003)	.96 (.003)	.80
LSA cosine ^c	.37 (.02)	.31 (.03)	.07
PMI ^d	8.65 (1.51)	7.96 (1.22)	.12
Thematic similarity ^c	4.37 (.10)	4.54 (.09)	.20
Feature similarity ^c	4.97 (.09)	3.15 (.07)	<.001
Pair familiarity ^c	5.23 (.09)	5.40 (.05)	.10
Categorical similarity ^c	5.22 (.08)	3.63 (.07)	<.001

C+A = categorically and associatively related lists; NC-A = noncategorical associatively related lists; HAL = hyperspace analog to language, Burgess & Lund, 1997; SUBTLEX = SUBTLEX_{US}, Brysbaert & New, 2009; Orthographic N = orthographic neighborhood size, Coltheart, Davelaar, Jonasson, & Besner, 1977; OLD = orthographic Levenshtein distance, Yarkoni, Balota, & Yap, 2008; OLDF = frequency of OLD neighbors, Yarkoni et al., 2008; LSA = latent semantic analysis, Landauer & Dumais, 1997; PMI = pointwise mutual information, Recchia & Jones, 2009. ^a Values from Nelson et al. (1998). ^b Values obtained from the English Lexicon Project (<http://lexicon.wustl.edu/>; Balota et al., 2007). ^c Values obtained from latent semantic analysis at CU Boulder (<http://lsa.colorado.edu/>). ^d Values courtesy of Michael Jones. ^e Values obtained from the norming session (see the Materials section)

the associated CL appeared equal numbers of times in all conditions (studied vs. control and C+A vs. NC-A).

After the recognition task, the participants were debriefed about the purposes of the study, thanked, and compensated.

Results

Recall data The veridical and false recall rates were analyzed in a 2×2 analysis of variance (ANOVA) with List Type (C+A or NC-A) and Item Type (list items or CLs) as within-subjects factors. Table 2 displays the mean proportion recall rates by factor. The list type main effect was significant, $F(1, 79) = 8.12$, $p = .006$, $\eta_p^2 = .093$, with higher recall for C+A lists ($M = .41$, $SE = .01$) than for NC-A lists ($M = .38$, $SE = .01$). The item type by list type interaction was not significant, $F < 1.0$, indicating that the list type recall difference was present for both veridical and false recall. The item type main effect was significant, $F(1, 79) = 521.20$, $p < .001$, $\eta_p^2 = .87$, with higher veridical recall ($M = .64$, $SE = .01$) than false recall ($M = .15$, $SE = .02$).

Recognition data Recognition rates were calculated from the proportions of “old” responses for list items and CLs from the studied lists minus the proportions of “old” responses for list

items and CLs from the nonstudied control lists. Thus, the recognition rates for studied lists were corrected by false alarm rates for the control lists. Table 2 displays the proportions of “old” responses for each item and list type for the studied and control lists. A 2×2 ANOVA was conducted for these corrected recognition rates with List Type (C+A or NC-A)

Table 2 Recall and recognition rates by item type and list type in Experiments 1 and 2 (standard errors are in parentheses)

	Studied List	Studied CL	Control List	Control CL
Experiment 1				
Recall				
C+A lists	.66 (.01)	.16 (.02)	–	–
NC-A lists	.62 (.01)	.14 (.02)	–	–
Recognition				
C+A lists	.84 (.02)	.53 (.03)	.04 (.01)	.04 (.01)
NC-A lists	.78 (.02)	.39 (.04)	.04 (.01)	.03 (.01)
Experiment 2				
Recognition				
C+A lists	.71 (.02)	.61 (.03)	.14 (.02)	.13 (.02)
NC-A lists	.63 (.02)	.51 (.03)	.10 (.01)	.12 (.02)

CL = critical lure; C+A = categorically and associatively related lists; NC-A = noncategorical associatively related lists

and Item Type (list items or CLs) as factors. The list type main effect was significant, $F(1, 79) = 21.38$, $p < .001$, $\eta_p^2 = .21$, with higher recognition rates for C+A lists ($M = .65$, $SE = .02$) than for NC-A lists ($M = .55$, $SE = .02$). The item type main effect was also significant, $F(1, 79) = 118.31$, $p < .001$, $\eta_p^2 = .60$, with higher veridical recognition ($M = .77$, $SE = .02$) than false recognition ($M = .43$, $SE = .03$). The item type by list type interaction was marginally significant, $F(1, 79) = 3.77$, $p = .056$, $\eta_p^2 = .05$. To confirm that the list type effect was significant for both list items and CLs, follow-up comparisons showed that the C+A lists resulted in higher recognition rates than NC-A lists for both veridical recognition, $p = .005$, and false recognition, $p < .001$.²

Although the lists were matched on a number of factors, we were unable to match them on forward associative strength (FAS). FAS can be interpreted as a measure of the association strength from the lure to list items and is assumed to reflect semantic similarity (Arndt, 2012). Although FAS has not consistently influenced false memory (e.g., Roediger, Balota, & Watson, 2001), recently Arndt did report independent contributions of FAS to false recognition, such that lists with higher FAS resulted in more errors than did lists with lower FAS when BAS was controlled. One interpretation of Arndt's results is that any factor that increases the similarity between memory traces of lures and list items can affect error rates. Arndt concluded that AMT could not account for these findings because FAS should not increase false recognition (i.e., the activation of list items from the CL should not influence CL errors). We note, however, that Arndt's stimuli were not matched on other dimensions and that different CLs were used in the high versus low BAS and FAS conditions in his study. It is possible that some item-level differences might have influenced his results. Importantly, however, his findings suggest that the semantic similarity of the lures to the list items can also contribute to false-memory creation. Because we

were attempting to empirically manipulate semantic similarity in terms of shared features, examining FAS might provide insights into the processes underlying the heightened errors with C+A lists. An examination of the mean FAS revealed that C+A lists had higher FAS ($M = .04$, $SEM = .01$) than NC-A lists ($M = .02$, $SEM = .003$), $t(19) = 2.69$, $p = .02$. To address this issue, we reanalyzed the data after removing seven lists, such that FAS values were equal ($M = .02$ for both types of lists, $p = .98$). The analyses with this subset of items confirmed greater false recognition for C+A lists ($M = .47$, $SEM = .04$) than for NC-A lists ($M = .36$, $SEM = .04$), $t(79) = 2.47$, $p = .02$. Thus, although FAS might be capturing some element of similarity, the effect of the shared features does seem to have contributed to errors independently.

We also note that the subsets of lists were also more closely matched on LSA and PMI, two other metrics that tended to have higher values for C+A than for NC-A lists. Additional analyses at the item level were conducted using FAS and LSA as covariates. Importantly, the difference in false recognition remained significant following these analyses, $F(1, 17) = 5.91$, $p = .03$, $\eta_p^2 = .26$. Thus, although lists that include feature overlap tend to be more strongly associated with the CL in some measures of semantic similarity, this difference does not seem to have driven the effects.³

Experiment 2

In Experiment 1, C+A lists resulted in higher false recognition than did NC-A lists, suggesting that feature overlap increased false memory above and beyond the contribution of BAS. Because veridical recall was also higher for C+A lists, the increased false memory effect might have been due to the higher initial recall resulting in a stronger memory trace. If the initial recall created more persistent memory for the C+A lists, this could have been a sort of testing effect (e.g., Roediger & Karpicke, 2006) whereby prior retrieval attempts modulate later memory (see also Huff et al., 2012). In Experiment 2, participants did not do free recall after each list. Thus, the performance on the final recognition task was uncontaminated by potential differences in initial recall.

² In addition to the high-threshold correction, because corrected recognition scores do not distinguish between old/new item sensitivity and response bias, we also analyzed d' , a commonly used metric to assess the strength of memory, as a measure of sensitivity, along with c , a measure of criterion. For the sake of brevity, we do not report those analyses here, because they were consistent with the analyses of corrected recognition. In the d' analyses, the interaction between list type and item type reached conventional significance levels. The criterion analyses revealed more conservative responses to NC-A than to C+A lists in both experiments, as well as more liberal responses to list items than to CLs in Experiment 1 and an interaction between item type and list type in Experiment 1, indicative of the fact that the type of relations between the list items affected criterion placement more for CLs than for list items.

³ As was noted by a reviewer, the difference in PMI approached significance, with C+A lists having slightly higher values than NC-A lists ($p = .12$). To address this potential confound, we entered PMI as a covariate at the item level. The magnitude of the difference remained the same; however, the effect of list type on false recognition was only marginally significant, $F(1, 18) = 3.75$, $p = .07$, $\eta_p^2 = .17$, in Experiment 1, and was nonsignificant in Experiment 2, suggesting that PMI as a metric might capture something similar to categorical relations.

Method

Participants Nineteen participants were tested at Illinois State University and 52 at Colby College. Three participants' data (two from Colby, one from Illinois State) were omitted from the analyses because their false alarms to control list items exceeded their hit rates to studied list items, suggesting they either misread the instructions or were guessing. The following analyses included the data from 68 participants.

Materials and procedure The same materials were used as in Experiment 1. The only difference was that participants did not do free recall after each list; instead, they completed math problems for 60 s. The final recognition test was identical.

Results

As in Experiment 1, the proportions of “old” responses from Experiment 2 were corrected by subtracting the proportions of “old” responses for the control lists from the proportions of “old” responses for studied lists for both the list items and CLs. The proportions of “old” responses of each type by condition and sample are displayed in Table 2. A 2×2 ANOVA was conducted for the corrected recognition scores with List Type (C+A or NC-A) and Item Type (list items or CLs) as factors. The list type main effect was significant, $F(1, 67) = 5.77, p = .019, \eta_p^2 = .08$, with higher recognition rates for C+A lists ($M = .52, SE = .02$) than for NC-A lists ($M = .45, SE = .02$). The item type main effect was also significant, $F(1, 67) = 13.33, p = .001, \eta_p^2 = .17$, with higher veridical recognition ($M = .54, SE = .02$) than false recognition ($M = .43, SE = .03$). The item type by list type interaction was not significant, $F < 1.0$, suggesting that the C+A advantage was present for both list and lure items. Although the interaction was not significant, because the main goal of Experiment 2 was to replicate the finding that false memories were higher following study of C+A lists than of NC-A lists, we conducted subsequent analyses to confirm this effect. In fact, follow-up t tests indicated that the list type effect (C+A > NC-A) was only reliable for CLs, $t(67) = 2.00, p = .049$, but the numerical difference for list items was not significant, $t(67) = 1.56, p = .12$. Thus, even when veridical recognition did not differ as a function of list type, false recognition still showed sensitivity to the main manipulation. As in Experiment 1, we examined false recognition for FAS-matched lists. Although numerically the C+A lists ($M = .46, SEM = .04$) resulted in more false alarms than did the NC-A lists ($M = .39, SEM = .04$), the difference was not reliable, $p = .18$. At the item level, when FAS and LSA were entered as covariates, the numerical advantage of C+A (corrected mean = .46) over NC-A (corrected mean = .40) lists was still not reliable, $p = .39$. Thus, it appears that prior recall might have enhanced the effect of list type. To examine this further, we conducted an analysis on the FAS-

matched lists across experiments (at the participant level). The effect of list type was reliable in this analysis, $F(1, 146) = 6.93, p = .009, \eta_p^2 = .04$; however, neither the effect of experiment nor the interaction was significant (both F s < 1), suggesting that the overall effect of list type was consistent.

General discussion

The primary question addressed in this study was the extent to which false memories in the DRM paradigm are influenced by shared meaning, defined here as feature-level similarity between list items and nonpresented CLs. In other words, in order to falsely recall or recognize *dog*, do the perceptual and semantic features associated with “dogness” (e.g., four legs, fur, sharp teeth) need to be activated or accessed through the presentation of items that share those features? Alternatively, activation converging from semantic networks through purely associative pathways may be sufficient to drive subsequent false memory. Short-term semantic priming studies have demonstrated that facilitation can emerge for targets that are preceded by associatively related items in the absence of shared features (Hutchison, 2003). To examine whether shared features exert additive effects over and above association strength, two types of lists were developed: C+A lists, in which list items and CLs were associated in free association norms *and* were related through shared features and/or category membership, and NC-A lists, in which the relations between list items and CLs were associative in nature and not due to shared features. Because list items were matched on overall associative strength, as well as in overall similarity (LSA and PMI), familiarity, and thematic relatedness, and only differed in terms of shared features and categorical similarity, the present design allowed us to test for independent contributions of this factor above and beyond those of other factors known to affect false memory and semantic and lexical processing.

In the first experiment, false recall rates were equivalent across list types, with a slight but nonsignificant increase in critical intrusions for C+A lists relative to NC-A lists. Significantly higher rates of false recognition were observed with C+A lists relative to NC-A lists. Veridical recall and recognition were also higher for C+A than for NC-A lists. To determine whether the effect in false recognition was driven by the initial recall advantage for C+A lists, in Experiment 2, the recognition test was administered without prior recall. Once again, C+A lists yielded higher false-recognition rates than did NC-A lists. Importantly, this occurred even though no difference was apparent between C+A and NC-A lists in veridical recognition, suggesting that the effect was not driven by higher veridical recall or recognition. Thus, when lists were carefully matched on multiple dimensions and only differed in whether or not the CL shared features with the list items, it

appears that similarity in terms of shared perceptual and semantic features exerts additive effects with associative strength. Additional analyses with lists matched on FAS at the participant level confirmed the difference between C+A and NC-A lists in Experiment 1; however, the difference was reduced in Experiment 2. Item analyses using FAS, LSA, and PMI as covariates further supported the general effect in Experiment 1, but the list type effect was not reliable in Experiment 2. One possibility is that prior recall, by strengthening the item-level representations of studied items (where a C+A advantage was present), contributes to the list type effect. In other words, the more robust effect in Experiment 1 might have benefitted from the prior recall task. An overall analysis combining the data from both experiments did, however, confirm the list type difference. Thus, although other factors, such as prior recall or similarity as captured by metrics such as LSA and PMI, might contribute to the effect, the overall pattern of results does suggest that categorical or feature-level similarity contribute to false recognition beyond these variables and associative strength. Another possibility is that FAS affects false memory by increasing semantic similarity, and this similarity is more sensitive to shared features than is BAS. However, whereas FAS is a global measure of similarity and, much like BAS, includes associations at both a semantic and a lexical level, our empirical manipulation of similarity was specific to shared features. Although controlling for FAS decreased the magnitude of the effect in Experiment 2, it appears that the feature-based similarity employed here contributed independent sources of information that increased false recognition. Thus, we suggest that neither BAS nor FAS can fully explain the observed errors, but that some level of semantic similarity above and beyond associative strength is involved in the process.

Because the same pattern was observed in both experiments, the effect was not entirely dependent on the initial recall task (cf. Huff et al., 2012). As we noted, however, veridical recognition was also higher for C+A than for NC-A lists (significantly in Exp. 1, numerically in Exp. 2). An analysis that included corrected veridical recognition as a covariate showed that the effect of list type was still reliable ($p < .001$ and $p = .046$, in Exps. 1 and 2, respectively). Thus, the higher false recognition for C+A lists seems to occur even when veridical recognition is taken into account. In addition, the recognition performance in Experiment 1 was conditionalized on prior recall in order to examine whether the slight increase in recall for C+A relative to NC-A lists contributed to the subsequent recognition difference. False alarms to CLs from studied lists were scored as being “previously recalled” or “not previously recalled” and submitted to a two-way ANOVA with List Type and Prior Recall Status as factors. Because, overall, false recall was low, only 20 participants’ data were included in this analysis. Previously recalled CLs ($M = .85$, $SEM = .05$) were recognized more than CLs that had

not been recalled ($M = .48$, $SEM = .08$), $F(1, 19) = 31.36$, $p < .001$, $\eta_p^2 = .62$, and C+A lists ($M = .74$, $SEM = .05$) yielded more false alarms than did NC-A lists ($M = .59$, $SEM = .07$), $F(1, 19) = 4.83$, $p = .04$, $\eta_p^2 = .20$. The interaction was not significant, $F < 1$. Additional analyses with the full sample compared the effects of list type only on nonrecalled CLs; once again, the effect of list type was significant, $t(79) = 3.64$, $p < .001$, with CLs from C+A lists ($M = .47$, $SEM = .04$) being falsely recognized more than CLs from NC-A lists ($M = .34$, $SEM = .04$). Thus, the conditional analyses confirmed that the differential false-recognition rates as a function of list type were not dependent on prior recall. One caveat to the present study is that the effect only seemed to emerge robustly in recognition; because the recognition test was administered after a brief delay, it is unclear whether the effect of shared features was due to the delay or to the type of test (recall vs. recognition). An experiment in which an immediate recognition or delayed recall test was administered would clarify this issue.

To our knowledge, this study has been the first to report greater false recognition of C+A related items relative to purely associative related items. Knott et al. (2012) did have matched lists, yet they reported equivalent false memory rates across list types. One possible explanation for the discrepancy between the present results and Knott et al.’s is that the associative lists used by Knott et al. included some items that shared features with the CL (e.g., *table* in the *chair* list). If semantic similarity does indeed contribute to forming the mental representations that support false recognition, it is possible that even a few items in each list might have provided enough activation of the features shared with the CL to boost error rates. Such an account would be consistent with proposals that false memories are supported, at least in part, by content-borrowing mechanisms, in which the episodic and/or perceptual details from encoded events are attributed to nonstudied events such as the CL. For example, Lampinen et al. (2005) found that content borrowing accounted for a large proportion of false alarms in the DRM paradigm. Converging evidence has come from source-monitoring paradigms, in which imagining the perceptual details of one item (e.g., a lollipop) increased false alarms to perceptually similar items (e.g., a magnifying glass; Henkel & Franklin, 1998), suggesting that shared perceptual features across object representations can influence subsequent memory for nonstudied but similar items. Thus, because the C+A lists activated more of the features present in the CL than did the NC-A lists, the representation of the CL would have been more readily accessed.

The results presented here also have some parallels to the “more is less” idea put forth by Toglia, Neuschatz, and Goodwin (1999), who reported that processes that increase veridical memory, such as deeper processing or presenting the items in blocked rather than random order, also increase

false memory. Similarly, we noted higher veridical recognition for C+A than for NC-A lists. Higher recognition of C+A list items is consistent with reports of category clustering (e.g., Bousfield, 1953), because C+A lists were more likely to include items from the same taxonomic category than were NC-A lists.

Regarding the role of associations, in the present study we found clear evidence supporting the role of associative strength in the absence of feature overlap, because NC-A lists yielded reliable false recall and false recognition. Importantly, the data indicate that conceptual similarity might make independent contributions related to different shared features. These results are consistent with the conclusions reached by Hutchison (2003), who concluded that automatic semantic priming effects are largely due to contributions from both associative relations and feature overlap, with some evidence suggesting independence of the two (e.g., mediated priming, priming from antonyms and synonyms). As we noted in the introduction, the “associative boost” in semantic priming (Hutchison, 2003; Lucas, 2000) refers to the fact that items that are related through shared category membership *and* associated according to free association norms tend to result in larger priming effects than do purely categorically related items (i.e., those that do not occur in free association norms). Here, we report a complementary phenomenon we refer to as a “feature boost”: Lists that shared features or category membership with the CL resulted in increased rates of false memory, relative to NC-A lists. Such a result suggests that these two factors—associative strength in lexical networks *and* similarity in semantic networks—exert additive effects in episodic memory tasks, increasing the accessibility of the CL. In terms of the organization of the semantic networks supporting priming and false-memory effects, it would appear, therefore, that lexical-level and semantic-level effects are dissociable. As was noted by Hutchison, it is not clear whether the associative boost observed is due to the combination of semantic/conceptual and lexical-level information (as was outlined in Collins & Loftus, 1975) or, as was suggested by McRae and Boisvert (1998), occurs because associated items also tend to be more similar. Because our lists were carefully matched in associative strength, we hypothesize that the feature boost is in fact due to additional similarity at the semantic/conceptual level, above and beyond the lexical-level associations.

Prior evidence supporting the additive roles of multiple sources of activation was presented by Watson et al. (2003), who found overadditive effects of semantic and orthographic/phonological information (see also Rubin & Wallace, 1989). Such findings are consistent with models of speech production, such as the interactive-activation model proposed by Dell, Schwartz, Martin, Saffran, and Gagnon (1997; see also Dell & O’Searghda, 1992). In such models, the mental lexicon consists of networks of form (i.e., lexemes), syntactic (i.e.,

lemmas), and semantic information that are distinct from conceptual representations. Thus, different levels or types of information, as well as the relationships between them, can be stored in distinct parts of the system. In the present context, the C+A lists would have provided additional conceptual-level information, above and beyond that given by lexical-level associations, thus resulting in the increased false recognition observed (i.e., the feature boost).

As we noted in the introduction, according to FTT (Brainerd & Reyna, 2002), participants can retrieve an item in one of two ways: by accessing the verbatim trace or by accessing a gist trace. Because AMT and FTT often make similar predictions regarding the effects of certain manipulations on subsequent false memory, it has been difficult to tease these frameworks apart experimentally. For example, lists with higher BAS are more likely to elicit strong gist traces than are lists with lower BAS, and so are lists with more shared features. Thus, Gallo and Roediger’s (2002) finding that weak lists indeed result in lower rates of false memory is largely consistent with this account, as well, although the authors of that study observed that gist-based theories appear to assume that gist depends largely on semantic features, whereas BAS is a measure of free association that does not necessarily correlate with overlap in terms of semantic features. In addition, typical DRM lists generally include items that are related by shared features (i.e., semantically related items) or because they are normatively or associatively related.

The results reported here are quite consistent with FTT, in that the lists that presumably give rise to the stronger gist or theme (i.e., C+A lists) resulted in higher rates of false recognition than did purely NC-A lists. Whether C+A lists do elicit a stronger gist trace is something that remains to be determined. Recently, Cann, McRae, and Katz (2011) developed DRM-type lists consisting of a specific type of semantic relation—specifically, situation features (such as function, location, or participant; see Wu & Barsalou, 2009). Although these lists were fairly low in BAS, they still elicited reliable rates of false memory. Cann et al. noted that BAS itself was predicted by specific types of semantic relations: specifically, taxonomic relations, synonyms, and situation features. Clearly, such findings underscore the importance of examining semantic and associative effects further, and support the idea that both factors contribute to false memory in the DRM paradigm.

Although we have been framing this work largely in the context of activation-based or error-inducing processes, an additional error-reducing process is involved in both AMT and FTT: Monitoring and recollection rejection are assumed to operate in order to counteract increases in accessibility or familiarity or reliance on gist traces. Thus, processes that can increase these error-editing processes might also differ as a function of list composition. For example, the extent to which

the theme word is identifiable can reduce errors (Carneiro, Fernandez, & Diaz, 2009; Huff et al., 2012). In the present context, there are two possibilities. One is that the shared features make the CL easier to identify in the context of a C+A list than in a NC-A list (we note that this is something we do not currently have evidence for or against). If that were the case, one would expect warnings to be more effective for C+A than for NC-A lists. Conversely, it is possible that participants can use a form of the recall-to-reject strategy (Gallo, 2010) and are better able to reject the CLs from NC-A lists because fewer shared features would be active. In other words, studying the list of items related to *horse* might lead one to retrieve studied items such as *stable*, *cowboy*, and *saddle* and to remember not studying any words that referred to four-legged mammals. However, the C+A list would result in the retrieval or reactivation of shared features between the CL *horse* and items such as *zebra* and *donkey*.

An alternative to both AMT (Roediger, Balota, & Watson, 2001) and FTT (Brainerd & Reyna, 2002) that can readily accommodate differential rates of false memory for items that are associatively and categorically related is a global-matching model such as MINERVA2 (Hintzman, 1986, 1988). Briefly, this model assumes that information is represented in memory as sets of features. In the DRM paradigm, a study event such as a word list will result in a trace or vector for each item on the list, with specific features of that item being probabilistically encoded (i.e., not all features will be encoded perfectly). At test, a probe is matched to all stored vectors, and a familiarity or strength signal is returned. The signal, or echo, is a result of shared features between the probes and traces being activated, and the overall echo intensity is the sum across all vectors. Thus, a high familiarity signal can indicate a high degree of similarity to one trace in memory (i.e., a hit) or a moderate degree of similarity to many traces (i.e., a false alarm to a CL, which shares fewer features with any single trace but many features across a number of stored traces).

Two characteristics of MINERVA2 make it particularly interesting with regard to the present research. First, this model can account for associative memory effects (e.g., Hintzman, 1988) by assuming that the activation of a feature in a vector will activate all features in that vector, even those not present in the probe. Thus, the model can explain the demonstrated effects of associative strength on false recognition (e.g., Arndt & Hirshman, 1998). Second, the model predicts that a greater number of shared features between probes and memory traces will increase false alarms. Therefore, whereas AMT (Roediger, Balota, & Watson, 2001) primarily assumes that BAS will drive false memory and is somewhat agnostic on the role of additional similarity between list and lure items, global-matching models are more explicit in assuming that associative strength and similarity make independent contributions to

false alarms. Similarly, FTT (Brainerd & Reyna, 2002) also assumes that factors that increase the similarity between list items and a CL will increase gist-based processes, and thus false alarms. The evidence reported by Arndt (2012) that FAS contributes to false recognition independently of BAS is consistent with global-matching models, as is the present evidence that the number of shared features affects errors.

One of the key issues addressed here was the nature of the mental representations that give rise to memory intrusions. The broader theoretical question is often framed as whether the mental lexicon is organized according to semantic or feature similarity or associative relations (e.g., Hutchison, 2003; Lucas, 2000; McRae & Boisvert, 1998). According to feature-based models (e.g., Masson, 1995; McRae, de Sa, & Seidenberg, 1997), items cluster because of feature overlap. Thus, activation spreads along feature nodes (e.g., “has four legs,” “has fur,” “has a tail”), and neighbors in a semantic network are those that share large numbers of features. Conversely, associative models (e.g., Anderson, 1983; Collins & Loftus, 1975; Steyvers & Tenenbaum, 2005) assume that the mental lexicon is organized by similarity but also by co-occurrence in the language, such that items that tend to appear close to each other in linguistic exchanges are more closely associated than items that rarely co-occur (Fodor, 1983). In these models, therefore, relatedness is not defined solely in terms of shared features or similarity, but along a variety of conceptual dimensions. Furthermore, such models assume that relatedness effects can be driven by associations at the lexical level, without necessarily implying shared meaning at a conceptual level (see, e.g., Balota & Paul, 1996). The present results can be considered largely consistent with both approaches: Shared features across item representations can increase the accessibility of a nonstudied item, although they are not strictly necessary (as was demonstrated by the robust false memory for NC-A items), at least as far as perceptual features are concerned.

In sum, the present study provides, to our knowledge, the first demonstration of higher false recognition for categorically and associatively related lists than for purely associatively related lists. Importantly, the lists were extensively matched on a variety of lexical and associative factors, and thus provide evidence for the importance of both associations in semantic networks and shared features or categorical similarity in false memory. Clearly, associations occurring at a purely lexical level, independent of meaning, and semantic or conceptual information, as provided by shared features, are important in false-memory formation and appear to exert independent effects. A simple spread of activation as a function of association strength does result in high error rates; however, feature-based similarity further boosts these errors, suggesting that gist or thematic extraction also plays a role in the effect.

References

- Anderson, J. R. (1983). Retrieval of information from long-term memory. *Science*, 220, 25–30. doi:10.1126/science.6828877
- Arndt, J. (2012). The influence of forward and backward associative strength on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 747–756. doi:10.1037/a0026375
- 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. doi:10.1006/jmla.1998.2581
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133, 283–316. doi:10.1037/0096-3445.133.2.283
- Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 336–345. doi:10.1037/0278-7393.12.3.336
- Balota, D. A., & Paul, S. T. (1996). Summation of activation: Evidence from multiple primes that converge and diverge within semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 827–845. doi:10.1037/0278-7393.22.4.827
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445–459. doi:10.3758/BF03193014
- Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, 49, 229–240. doi:10.1080/00221309.1953.9710088
- Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. In H. W. Reese & R. Kail (Eds.), *Advances in child development and behavior* (pp. 41–100). San Diego, CA: Academic Press. doi:10.1016/S0065-2407(02)80062-3
- Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, 11, 164–169. doi:10.1111/1467-8721.00192
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 977–990. doi:10.3758/BRM.41.4.977
- Buchanan, L., Brown, N. R., Cabeza, R., & Maitson, C. (1999). False memories and semantic lexicon arrangement. *Brain and Language*, 68, 172–177. doi:10.1006/brln.1999.2072
- Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12, 177–210. doi:10.1080/016909697.86844
- Cann, D. R., McRae, K., & Katz, A. N. (2011). False recall in the Deese–Roediger–McDermott paradigm: The roles of gist and associative strength. *Quarterly Journal of Experimental Psychology*, 64, 1515–1542. doi:10.1080/17470218.2011.560272
- Carneiro, P., Fernandez, A., & Diaz, A. R. (2009). The influence of theme identifiability on false memories: Evidence for age-dependent opposite effects. *Memory & Cognition*, 37, 115–129. doi:10.3758/MC.37.2.115
- Coane, J. H., Balota, D. A., Dolan, P. O., & Jacoby, L. L. (2011). Not all sources of familiarity are created equal: The case of word frequency and repetition in episodic recognition. *Memory & Cognition*, 39, 791–805. doi:10.3758/s13421-010-0069-5
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407–428. doi:10.1037/0033-295X.82.6.407
- Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). New York, NY: Academic Press.
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684. doi:10.1016/S0022-5371(72)80001-X
- Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104, 268–294. doi:10.1037/0096-3445.104.3.268
- Deese, J. (1959). Influence of inter-item associative strength upon immediate free recall. *Psychological Reports*, 5, 305–312. doi:10.2466/PRO.5.3.305-312
- Dell, G. S., & O’Searghda, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42, 287–314. doi:10.1016/0010-0277(92)90046-K
- Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, 104, 801–838. doi:10.1037/0033-295X.104.4.801
- Dewhurst, S. A., Barry, C., Swannell, E. R., Holmes, S. J., & Bathurst, G. L. (2007). The effect of divided attention on false memory depends on how memory is tested. *Memory & Cognition*, 35, 660–667. doi:10.3758/BF03193304
- Dewhurst, S. A., Bould, E., Knott, L. M., & Thorley, C. (2009). The roles of encoding and retrieval processes in associative and categorical memory illusions. *Journal of Memory and Language*, 60, 154–164. doi:10.1016/j.jml.2008.09.002
- Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.
- Gallo, D. A. (2006). *Associative illusions of memory: False memory research in DRM and related tasks*. New York, NY: Psychology Press.
- Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, 38, 833–848. doi:10.3758/MC.38.7.833
- Gallo, D. A., & Roediger, H. L., III. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language*, 47, 469–497. doi:10.1016/S0749-596X(02)00013-X
- Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, 13, 8–20. doi:10.3758/BF03198438
- Henkel, L. A., & Franklin, N. (1998). Reality monitoring of physically similar and conceptually related objects. *Memory & Cognition*, 26, 659–673. doi:10.3758/BF03211386
- Hintzman, D. L. (1986). ‘Schema abstraction’ in a multiple-trace memory model. *Psychological Review*, 93, 411–428. doi:10.1037/0033-295X.93.4.411
- Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95, 528–551. doi:10.1037/0033-295X.95.4.528
- Huff, M., Coane, J. H., Hutchison, K. A., Grasser, E. B., & Blais, J. E. (2012). Interpolated task effects on direct and mediated false recognition: Effects of initial recall, recognition, and the ironic effect of guessing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 1720–1730. doi:10.1037/a0028476
- Huff, M., & Hutchison, K. A. (2011). The effects of mediated word lists on false recall and recognition. *Memory & Cognition*, 39, 941–953. doi:10.3758/s13421-011-0077-0
- Hunt, R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review*, 2, 105–112. doi:10.3758/BF03214414

- Hunt, R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language*, 32, 421–445. doi:10.1006/jmla.1993.1023
- Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, 10, 785–813. doi:10.3758/BF03196544
- Hutchison, K. A., & Balota, D. A. (2005). Decoupling semantic and associative information in false memories: Explorations with semantically ambiguous and unambiguous critical lures. *Journal of Memory and Language*, 52, 1–28. doi:10.1016/j.jml.2004.08.003
- Jones, L. L., & Golonka, S. (2012). Different influences on lexical priming for integrative, thematic, and taxonomic relations. *Frontiers in Human Neuroscience*, 6, 1–17. doi:10.3389/fnhum.2012.00205
- Knott, L. M., & Dewhurst, S. A. (2007). The effects of divided attention at study and test on false recognition: A comparison of DRM and categorized lists. *Memory & Cognition*, 35, 1954–1965. doi:10.3758/BF03192928
- Knott, L. M., Dewhurst, S. A., & Howe, M. L. (2012). What factors underlie associative and categorical memory illusions? The roles of backward associative strength and interitem connectivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 229–239. doi:10.1037/a0025201
- Lampinen, J. M., Meier, C. R., Arnal, J. D., & Leding, J. K. (2005). Compelling untruths: Content borrowing and vivid false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 954–963. doi:10.1037/0278-7393.31.5.954
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240. doi:10.1037/0033-295X.104.2.211
- Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, 7, 618–630. doi:10.3758/BF03212999
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44, 1–23. doi:10.3758/s13428-011-0124-6
- Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 3–23. doi:10.1037/0278-7393.21.1.3
- 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. doi:10.1037/0278-7393.25.5.1177
- McNamara, T. P. (2005). *Semantic priming*. New York, NY: Psychology Press.
- McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 558–572. doi:10.1037/0278-7393.24.3.558
- McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126, 99–130. doi:10.1037/0096-3445.126.2.99
- Neely, J. H., & Tse, C. (2007). Semantic relatedness effects on true and false memories in episodic recognition: A methodological and empirical review. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III* (pp. 313–351). New York, NY: Psychology Press.
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from www.usf.edu/FreeAssociation/
- Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, 41, 647–656. doi:10.3758/BRM.41.3.647
- Roediger, H. L., III, Balota, D. A., & Watson, J. M. (2001). Spreading activation and arousal of false memories. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). Washington, DC: American Psychological Association. doi:10.1037/10394-006
- Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181–210. doi:10.1111/j.1745-6916.2006.00012.x
- Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 803–814. doi:10.1037/0278-7393.21.4.803
- Roediger, H. L., III, Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, 8, 385–407. doi:10.3758/BF03196177
- Rubin, D. C., & Wallace, W. T. (1989). Rhyme and reason: Analyses of dual retrieval cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 698–709. doi:10.1037/0278-7393.15.4.698
- 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. doi:10.1016/S0749-596X(02)00012-8
- Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29, 41–78. doi:10.1207/s15516709cog2901_3
- Toglia, M. P., Neuschatz, J. S., & Goodwin, K. A. (1999). Recall accuracy and illusory memories: When more is less. *Memory*, 7, 233–256. doi:10.1080/741944069
- Watson, J. M., Balota, D. A., & Roediger, H. L., III. (2003). Creating false memories with hybrid lists of semantic and phonological associates: Over-additive false memories produced by converging associative networks. *Journal of Memory and Language*, 49, 95–118. doi:10.1016/S0749-596X(03)00019-6
- Wu, L. L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, 132, 173–189. doi:10.1016/j.actpsy.2009.02.002
- Yarkoni, T., Balota, D. A., & Yap, M. J. (2008). Beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15, 971–979. doi:10.3758/PBR.15.5.971