Journal of Memory and Language 93 (2017) 154–168

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Relative contributions of semantic and phonological associates to over-additive false recall in hybrid DRM lists

Jason R. Finley *, Victor W. Sungkhasettee, Henry L. Roediger III, David A. Balota

Department of Psychological and Brain Sciences, Washington University in St. Louis, United States

ARTICLE INFO

Article history:
Received 26 January 2016
revision received 27 July 2016

Keywords:
False memory
DRM paradigm
Semantic processing
Phonological processing
Hybrid list

ABSTRACT

Two experiments explored false recall of unstudied critical items (e.g., *chair*) following the presentation of 16 semantic associates to the critical word (e.g., *sit*, *desk*), 16 phonological associates to the critical word (e.g., *cheer*, *hair*), and every composition of hybrid list in between (e.g., 14 semantic and 2 phonological associates). Results replicated the over-additive pattern of critical false recall from hybrid lists relative to pure lists found by Watson, Balota, and Roediger (2003) and clarified the form of the false recall function across varying degrees of hybridization. Both experiments showed that including just one or two of the other type of associate in an otherwise pure list led to a considerable increase in false recall. A within-subjects design (Experiment 1) suggested that after this initial rapid increase, false recall continued to increase gradually to an apex at the balanced hybrid list composition, whereas a between-subjects design (Experiment 2) showed that false recall plateaued after the initial rapid increase and that the overall shape of the function is a ziggurat. Furthermore, the function is roughly symmetrical; semantic and phonological associates appear to make equivalent contributions to over-additive false recall from hybrid lists. The results provide constraints on theoretical accounts of DRM false memories, and can be accommodated by a modified activation/monitoring framework.

Introduction

When people study a list of related words, they are often susceptible to falsely recalling or recognizing a critical word that was strongly semantically associated to the whole list but was not itself studied (the DRM paradigm; Deese, 1959; Roediger & McDermott, 1995). Similar false remembering has been found using lists of words that are phonologically and/or orthographically associated to the critical unstudied word (e.g., Sommers & Lewis, 1999). These findings have informed theorizing about the nature of information storage in human memory, and the nature of retrieval processes (Gallo, 2010). A question of key interest is whether the two types of associates, semantic and phonological,[1] contribute to false memory in the same way.

Watson, Balota, and Roediger (2003) explored this issue by using hybrid lists composed of both semantic and phonological associates to a critical item (see also Watson, Balota, & Sergent-Marshall, 2001). In their Experiment 1, they found that adding 1–3 phonological associates to a list that already contained 10 semantic

* Corresponding author at: Department of Behavioral Sciences, Fontbonne University, 6800 Wydown Blvd., St. Louis, MO 63105, United States.
E-mail address: jfinley@fontbonne.edu (J.R. Finley).
URL: http://jasonrfinley.com (J.R. Finley).

[1] We use the term phonological as a shorthand for representation in a lexical network, but mean to include visual/orthographic representation, too. In our experiments words were presented visually during study, so doubtless orthographic and phonological processes are involved during encoding.

http://dx.doi.org/10.1016/j.jml.2016.07.006
0749-596X/

associates increased false recall more than adding 1–3 additional semantic associates to the same list. In Experiment 2 they found that combining a semantic and phonological list (length 36, 18 words of each type) yielded greater false recall than the sum of that yielded by either list alone (length 18): over-additivity (see Balota & Paul, 1996 for discussion of additivity). In Experiment 3 they found that a balanced hybrid list of length 16 (composed of 8 semantic and 8 phonological associates) led to greater false recall than either a pure semantic or pure phonological list of length 16. Thus, when examined three different ways, hybrid lists of semantic and phonological associates lead to greater false recall than pure lists of either type.

Watson et al. (2003) discussed several possible theoretical accounts of the over-additive false recall produced by hybrid lists. One account is a simple additive spreading activation model that posits distinct semantic and phonological (lexical) associative networks, that both could contribute to total activation of a critical item,[2] and that both have their own negatively accelerated activation functions that asymptote after a certain number of associates are activated. A hypothetical example is illustrated in Fig. 1. When there are already six semantic associates studied at encoding (step 1), adding three more semantic associates (step 2a) should not produce much of an increase in false recall of the critical item because the semantic network is probably already near asymptote and thus will not contribute much to the total activation of the critical item. That is, there are diminishing returns. However adding three phonological associates (step 2b) should provide a considerable increase in false recall, because of the large increase in phonological activation driven by going from zero to three phonological associates, which involves the rapidly rising portion of the activation function. Thus, a hybrid list with even just a few of the alternative type of associate should produce higher false recall than a pure list of the same length. Note that this theoretical account focuses on encoding processes: false recall of the critical item occurs because it was sufficiently activated by one or both of the associative networks at the time of study. But what of retrieval processes?

Another account discussed by Watson et al. (2003) is an activation/monitoring framework (Gallo, 2010; Roediger, Balota, & Watson, 2001; Roediger, Watson, McDermott, & Gallo, 2001). This account begins by positing that during study of a pure semantic DRM list, the critical item is activated (either consciously or unconsciously) via spreading activation. Then during retrieval, participants generate candidate items based on their semantic activation, but may reject those items due to a lack of corresponding phonological familiarity (cf. Jacoby, Kelley, & Dywan, 1989; Watkins & Gardiner, 1979). That is, participants may reject the critical item that is only semantically related to the list items if it is not also a familiar word form or sound. For example, if a participant studied words like *sit* and *desk* at encoding, later at retrieval she might generate the critical unstudied word *chair* based on semantic activation. But she may still be able to correctly reject that word and choose not to output it because it does not feel sufficiently familiar in sound and/or appearance, because there were no words that looked or sounded like *chair* in the list. Gallo (2004) refers to this strategy as diagnostic monitoring during retrieval. When phonological associates such as *cheer* and *hair* are included in the study list along with the semantic associates (*sit* and *desk*), not only does this boost activation of the critical unstudied item (as per the spreading activation account discussed earlier) but it also disrupts diagnostic monitoring during retrieval: the participant can no longer readily reject *chair* based on phonology because the words that sounded or looked like *chair* on the studied list have boosted its phonological familiarity. Thus this theory adds a strategic retrieval component to the basic idea that false memories arise during spreading activation through lexical or semantic networks (Balota & Paul, 1996). The activation/monitoring framework account also fits nicely with modality effects in the DRM paradigm: lists presented visually create less false recall than those presented auditorily, likely due to output monitoring of the word form driving down false recall in the visual case (see Gallo, McDermott, Percer, & Roediger, 2001; Kellogg, 2001).

We pause to mention one other theory of DRM false memories: fuzzy trace theory (e.g., Brainerd & Reyna, 2002; Reyna & Brainerd, 1995). Fuzzy trace theory posits that events are coded as verbatim traces (with specific details of the events) and/or gist traces (the semantic content). False memories arise from strong gist traces that lead to "phantom recollection" (Brainerd, Payne, Wright, & Reyna, 2003). Because false memories are entirely based on gist traces (semantic content) in this theory, there is no reason to expect phonological or orthographic associates to increase false recall or false recognition. Thus, in its current conceptualization, fuzzy trace theory is incapable of handling results from experiments showing false memories from phonological associates (Sommers & Lewis, 1999) or from hybrid lists of semantic and phonological associates (Watson et al., 2003). Of course, it is quite possible that a fuzzy trace could also involve phonological associates for gist based traces, but it would then be important to further stipulate what types of information are sufficient for gist based representations.

Watson et al. (2003) found increased false memory from hybrid relative to pure lists in both free recall and recognition (Experiment 3), with healthy young adult participants. Watson et al. (2001) additionally found the same false recall effect with healthy young adults, healthy older adults, and older adults with Alzheimer's disease. Curiously, Budson, Sullivan, Daffner, and Schacter (2003) found no difference in false recognition for hybrid versus pure lists with healthy young adults, healthy older adults, or older adults with Alzheimer's disease. Nevertheless, one issue that has not been resolved by prior research is the *relative contributions* of semantic versus phonological associates to false memory in the hybrid paradigm. That is, prior research has generally used equal numbers of the two types of associates (balanced hybrid lists) and not examined other list compositions. In the one study that has systematically varied the number of associates, Watson et al.

[2] Note that such an account is comparable to Dell's interactive model of speech production (Dell, 1986), which postulates that top-down semantic activation and bottom-up phonological activation combine to converge on a critical item, yielding a speech error.

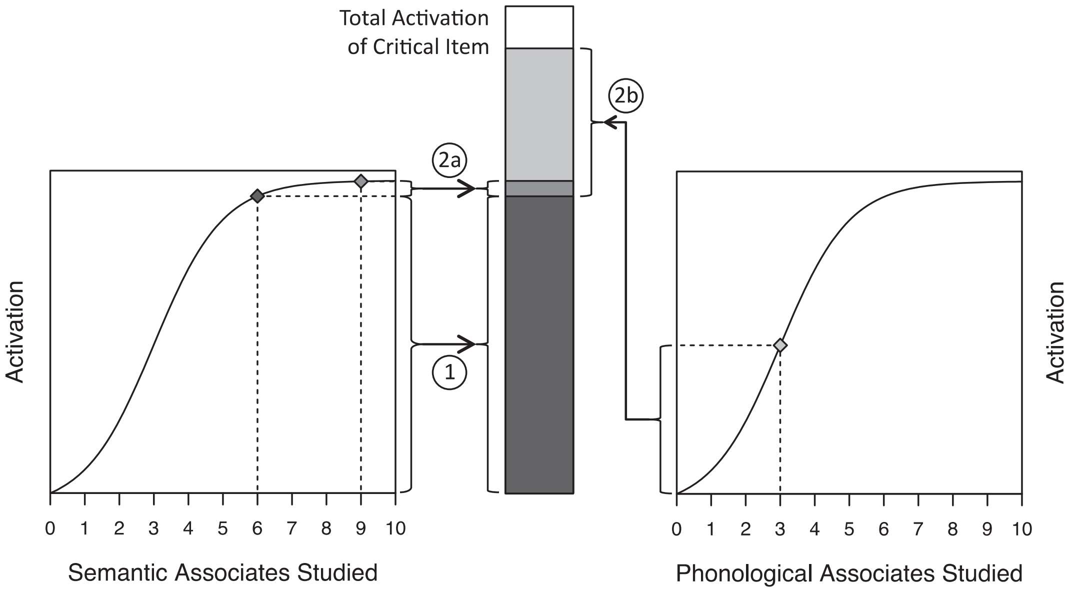

Fig. 1. Hypothetical example of activation functions corresponding to distinct semantic and phonological (lexical) networks that both contribute to total activation of a critical item. A list with 6 semantic associates provides some activation (1); adding three more semantic associates (2a) increases total activation less than adding three phonological associates (2b).

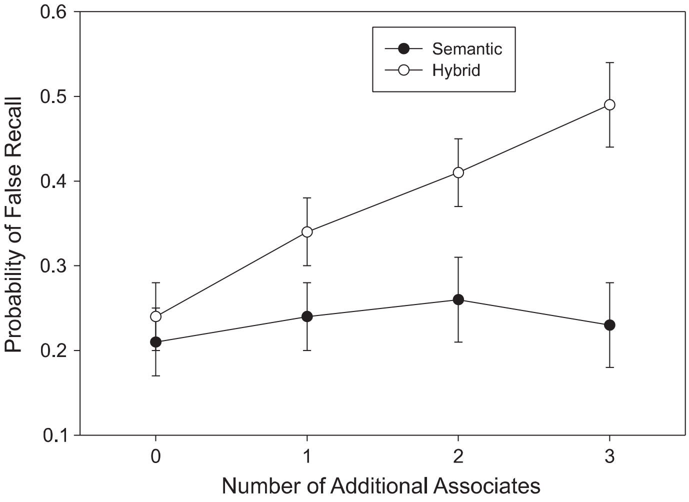

Fig. 2. From Watson et al. (2003), Experiment 1. © 2003 Elsevier Science. Mean probability of false recall as a function of increasing numbers of associates and list type. Error bars represent standard error of the means.

(2003, Experiment 1) replaced zero, one, two, or three semantic associates in a list with phonological associates. The results are reprinted in Fig. 2. Here one can see a linear increase in false recall as a function of the number of phonological associates. However, there remain at least two important questions. First, what is the shape of the function as even more phonological associates are included in the list, approaching the balanced composition? Second, what is the shape of the function when semantic associates are introduced to a phonological list, approaching the balanced composition?

In the current study we sought to answer these questions by measuring the levels of false recall yielded by the entire range of list compositions for a given list length, from purely semantic to purely phonological and everything in between. That is, we sought to determine the form of the false recall function across varying degrees of list hybridization. The shape of that function bears on the theoretical accounts we have described. The spreading activation account predicts not an endlessly linear increase in false recall with increasing hybridization, but rather a negatively accelerated increase, based on the underlying activation functions of the two networks (Fig. 1). The activation/monitoring framework predicts an asymmetrical function: including even just a few phonological associates in the study list should disrupt the strategic rejection of candidates based on phonology, thus rapidly driving up false recall on one side of the function; but including semantic associates should have no such strategy disruption effect, and thus the function should be shallower on that side. Fuzzy trace theory, by our understanding, predicts a steady decrease in false recall as list composition shifts from purely semantic to purely phonological, with no central increase for hybrid lists.

Experiment 1

This experiment was modeled after Experiment 3 by Watson et al. (2003). We expanded the study conditions to include all possible combinations of semantic and phonological associates in a list of 16 words.

Method

Participants studied and completed free recall tests for 18 lists of words. The experiment was conducted over the Internet, and was programmed using Adobe Flash ActionScript 3. There were two versions of the experiment with slight variations in materials, which we will call Experiments 1a and 1b. We will describe where they differed, but will report the results together.

Design

The independent variable was list composition. A given study list consisted of 16 words that were each related, either semantically or phonologically, to a single critical item that was not itself in the study list. Each list could be composed of anywhere from 0 to 16 semantic associates to the critical item, with the remaining list items being phonological associates to the critical item. For example, a given study list could be composed of 16 semantic associates and 0 phonological associates (s16p0 for shorthand), 15 semantic and 1 phonological (s15p1), 14 semantic and 2 phonological (s14p2), and so on, all the way to 0 semantic and 16 phonological (s0p16). Thus there were 17 possible list composition conditions.

Each participant studied and was tested on 18 lists, and experienced 9 of the 17 possible list composition conditions. Participants were randomly assigned to one of two groups that determined which conditions they experienced. Participants in the "ends" group studied and were tested on two lists in each of the following 9 conditions: s16p0, s15p1, s14p2, s13p3, s8p8, s3p13, s2p14, s1p15, s0p16. Participants in the "middle" group studied and were tests on two lists in each of the following 9 conditions: s12p4, s11p5, s10p6, s9p7, s8p8, s7p9, s6p10, s5p11, s4p12. Note that both groups received the balanced condition composed of equal numbers of semantic and phonological associates (s8p8). This design was used in order to gather data on the whole range of conditions while also compromising between two practical goals: on the one hand limiting the total duration of the experiment to minimize fatigue and attrition, and on the other hand obtaining more than one observation per condition per participant.

The primary dependent measure was the proportion of critical items falsely recalled. We also examined output order for that item, as well as veridical recall of studied items.

Participants

Participants were 381 people recruited from Amazon's Mechanical Turk and paid $3 each. There were 171 men and 210 women, and the mean age was 35.2 years (*SD* = 11.9). All participants had completed at least 500 previous HITs (human intelligence tasks) on Mechanical Turk, had at least a 95% approval rate on those HITs, and were located in the United States. There were 168 participants in the ends condition, and 213 in the middle condition. Data were collected from an additional 32 participants, but were excluded from analysis for any of three reasons: they self-reported that English was not their first language, they gave zero responses on one or more lists, or they self-reported that they took notes during the experiment (in response to a post-experimental questionnaire).

Materials

Materials were 36 pools of words developed by Watson et al. (2003) for their Experiment 3 and provided in their Appendix, pages 114–117. Each pool consisted of 32 words: 16 semantic associates and 16 phonological associates to a critical word (a common noun, adjective, or verb). We used these pools of words to construct the 18 lists of 16 words that each participant studied and was tested on.

There were three steps to construction of study lists. First, because each participant would only study and be tested on 18 lists total, we needed to determine which 18 of the 36 possible pools of words would be used to construct lists. In Experiment 1a, 18 pools were selected at random for each participant. In Experiment 1b, the same 18 pools were used to construct study lists for all participants (pools corresponding to the following critical words: chair, cold, dog, face, fat, glass, hand, kill, right, sick, sleep, slow, smoke, sweet, test, top, trash, wet); these were chosen as the 18 pools that yielded the highest levels of false recall of the critical word in the s8p8 list composition condition in Experiment 1a.

Second, once we had determined which pools of words would be used to construct study lists, we needed to assign those pools to the list composition conditions that a given participant would experience. In Experiment 1a this was done randomly for each participant. In Experiment 1b the 18 pools of words were counterbalanced using a Latin square such that across participants each pool was assigned equally often to each list composition condition.

Third, the particular words used to construct a study list (16 words) from a given pool (32 words) were selected randomly for each participant, within the confines of the particular list composition assigned to that pool. For example, if the "sleep" pool was assigned to the condition s9p7 for a participant, the study list would consist of 9 semantic associates to sleep selected at random from the 16 possible semantic associates in the pool, and 7 phonological associates to sleep selected at random from the 16 possible phonological associates in the pool.

Procedure

Participants studied a list of words and then completed a free recall test on that list, and this process occurred for a total of 18 lists. The order in which the list composition conditions occurred was randomized for each participant, as was the order in which particular words within a list were studied.

At the start of the experiment, participants were instructed that they would study and be tested on a series of word lists. For each list, participants first saw a study instruction screen which stated the list number (e.g., "Study List 1 of 18") and gave the following instructions:

> Now you will study a new list of words. The words will be presented individually on the screen for 1.5 seconds each. After studying this list of words you will be given a test where you will have to type in as many of the words as you can remember. Do not write down any notes; only use your own memory. Do not switch to any other windows or tabs on your computer. Press Continue when you are ready to start.

After participants pressed the Continue button, they studied 16 words as those words appeared individually on the screen for 1.5 s each with an inter-stimulus interval of 500 ms. After the sixteenth word disappeared, a test instruction screen appeared which stated the list number (e.g., "Test on List 1 of 18") and gave the following instructions:

> Now you will take a memory test on the list of words you just studied. You will have 1 minute to type as many of the words as you can remember from the study phase. Do not just type in lots of words in the hopes that some of them were in the study phase. ONLY TYPE WORDS THAT YOU REMEMBER FROM THE STUDY PHASE. Do not switch to any other windows or tabs on your computer.

This instruction screen was shown for exactly 30 s before the test automatically began. There was a 30 s countdown timer on the bottom of the instruction screen, and participants could neither skip nor pause the countdown; this enforced the same retention interval for all lists for all participants.

Then the free recall test screen appeared, with the following instruction at the top: "Type words you remember from the list you just studied." Participants typed individual word responses into a box at the bottom of the screen and pressed the enter/return key after each response. Each response entered this way was shown on the upper portion of the screen, so participants could see the responses they had already typed for this test, though they could not alter or delete them. The test lasted for 60 s, with a countdown timer shown at the top-right of the screen. After the test was done, the program moved on to the study instruction screen for the next list.

At the end of the experiment, participants were asked the series of questions given in Appendix A. Question 3 asked participants to estimate the number of lists on which they made a false recall. Question 4 concerned encoding strategies. Question 5 was a yes/no question about whether they had already known about false recall from DRM lists. Question 7 asked if they took notes during the experiment, and data from participants answering yes were excluded from analysis.

Results and discussion

An alpha level of .05 was used for all tests of statistical significance unless otherwise noted. Effect sizes for comparisons of means are reported as Cohen's *d*, calculated using the pooled standard deviation of the groups being compared. Effect sizes for one-way ANOVAs are reported as omega squared, $\widehat{\omega}^2$. Standard deviations (*SD*s) are reported raw (i.e., calculated using *N*, not $N-1$), on the grounds that the *SD* is a descriptive statistic, and the $N-1$ adjustment should be reserved for use in inferential statistics. Mauchly's test was used to detect violations of sphericity for within-subjects factors in ANOVAs; no such violations were detected.

False recall

The main dependent measure of interest was the proportion of lists on which participants falsely recalled the critical word to which all the studied words were related. First, we checked to see if there was an effect of the minor differences in materials across the two versions of the experiment (1a vs. 1b, as described in the Materials section). Remember that both versions 1a and 1b had a group of participants who received the middle set of list composition conditions and a group of participants who received the end set of list composition conditions; we label these groups middle and ends, respectively. We performed separate two-way mixed ANOVAs for the middle and ends groups and these analyses showed that experiment version (1a vs. 1b) had no main effect on false recall (middle group: $M_{1a} = .35$, $SD_{1a} = .21$, $M_{1b} = .41$, $SD_{1b} = .24$, $F(1,211) = 3.37$, $MSE = .044$, $p = .068$; ends group: $M_{1a} = .33$, $SD_{1a} = .19$, $M_{1b} = .30$, $SD_{1b} = .19$, $F(1,166) = 0.85$, $MSE = .334$, $p = .357$) and more importantly that there was no interaction with list composition (middle group: $F(8,1688) = 0.89$, $MSE = .095$, $p = .524$; ends group: $F(8,1328) = 0.45$, $MSE = .100$, $p = .891$). Thus, we combined the data from the two versions of the experiment for all subsequent analyses.

Second, we checked whether participants' self-reported prior knowledge of false memories in the DRM paradigm (Question 5, Appendix A) made a difference in their level of false recall. The number of participants who claimed that they previously knew about the DRM effect was 67 out of 213 in the middle group and 59 out of 168 in the ends group. In the middle group, the mean false recall of critical items was not significantly different for participants who claimed prior knowledge of the DRM effect ($M = .35$, $SD = .22$) versus those who did not ($M = .37$, $SD = .22$), $t(211) = 0.44$, $p = .660$. In the ends group, the mean false recall of critical items was not significantly different for participants who claimed prior knowledge of the

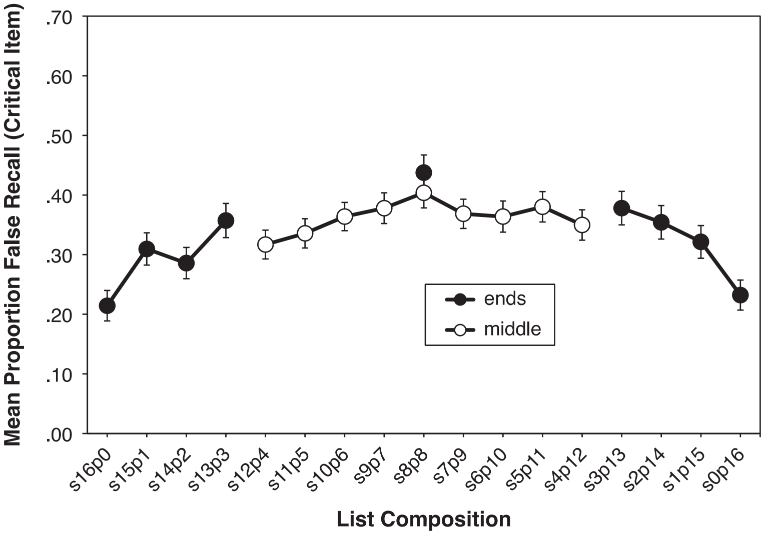

Fig. 3. Mean proportion of false recall of critical item as a function of list composition (number of semantic and phonological items per list) in Experiment 1. Note: "ends" and "middle" were two separate groups of participants. Error bars represent the standard error of each cell.

DRM effect ($M = .32$, $SD = .20$) versus those who did not ($M = .32$, $SD = .19$), $t(166) = 0.09$, $p = .926$. Thus we did not exclude data from participants claiming prior knowledge of the effect.

Fig. 3 shows mean proportion of false recall of the critical item as a function of list composition. As a reminder, "s16p0" indicates that the studied list consisted of 16 words semantically associated to the critical item and 0 words phonologically associated to the critical item. A "s3p13" list consisted of 3 semantic associated and 13 phonological associates. A table with the values of the means and *SD*s is provided in Appendix B.

The overall pattern is clear: all list compositions yielded substantial amounts of false recall, and the amount of false recall steadily increased as list composition became more hybridized. For the ends group, a one-way within-subjects ANOVA confirmed that there was a significant overall effect of list composition on false recall, $F(8,1336) = 8.39$, $MSE = .100$, $p < .001$, $\widehat{\omega}^2 = .030$, and also confirmed the visually apparent quadratic trend, $F(1,167) = 45.96$, $MSE = .118$, $p < .001$, $\widehat{\omega}^2 = .065$. For the middle group, the overall effect was curiously not significant, $F(8,1696) = 1.47$, $MSE = .095$, $p = .163$, $\widehat{\omega}^2 = .001$, but the quadratic trend was significant, $F(1,212) = 6.15$, $MSE = .106$, $p = .014$, $\widehat{\omega}^2 = .005$.

Purely semantic lists (s0p16) and purely phonological lists (s16p0) yielded equivalent levels of false recall, $t(167) = 0.58$, $p = .562$, $d = 0.05$, consistent with the findings of Watson et al. (2001, 2003, Experiment 3). Balanced hybrid lists (s8p8) yielded the most false recall. Within-subjects *t*-tests showed that false recall in the balanced hybrid condition was significantly higher than false recall averaged across all other conditions for the ends group, $M_{all\_other} = .31$, $SD_{all\_other} = .19$, $t(167) = 5.22$, $p < .001$, $d = 0.43$, and for the middle group, $M_{all\_other} = .36$, $SD_{all\_other} = .22$, $t(212) = 2.09$, $p = .038$, $d = 0.15$. As Watson et al. (2003) pointed out, the total number of associates to the critical item is the same across all lists (16), so the fact that false recall increases with hybridization indicates an *over-additive* influence of the two types of associates.

Another interesting point is that the largest increase in false recall comes from including just one of the other type of associate as compared to the pure list. Including a single phonological associate raised false recall proportion by an average of .10 ($SD = .42$) compared to the pure semantic list. Including a single semantic associate raised false recall proportion by an average of .09 ($SD = .44$) compared to the pure phonological list. This result suggests that the over-additivity of false recall from hybrid lists may be driven by the mere inclusion of both types of associates, in any proportions.

The finding that false recall increased with the inclusion of just one phonological associate is compatible with the activation/monitoring framework whereby the monitoring process uses information about phonological or orthographic information to reject a generated candidate item as not being on the list. Including even one phonological associate disrupts this monitoring process. However, the activation/monitoring framework requires some adjustment to account for the finding that false recall also increased with the inclusion of just one semantic associate; we will return to this point in the General Discussion.

The increase in false recall due to hybridization appears to be nearly symmetrical for the semantic and phonological sides of the function. In the ends group, there was no significant difference in the absolute value of participants' linear regression slopes for the semantic side on the left of Fig. 3 ($M = .04$, $SD = .04$) versus the phonological side on the right ($M = .04$, $SD = .04$), $t(167) = 0.21$, $p = .834$, $d = 0.02$. In the middle group, there was a significant difference ($M_{sem} = .08$, $SD_{sem} = .07$, $M_{phon} = .07$, $SD_{phon} = .06$), $t(212) = 2.23$, $p = .027$, $d = 0.18$, but the effect size was small. The overall near-symmetry suggests equivalent contributions of additional semantic and phonological associates to the overall level of false recall. This feature of the data is consistent with the spreading activation aspect of activation/monitoring theory. If semantic associates activate a meaning-based associative network and phonological associates activate a lexical associative network, then increasing the numbers of associates of each type increases associative activation within both networks (Watson et al., 2001, 2003).

The overall shape of the function in Fig. 3 could roughly be described as a pyramid. However, much like the gambrel roof often observed on barns, the slope appears to be steeper on the ends and shallower in the middle. Between-subjects *t*-tests confirmed this pattern by comparing the slopes between the two groups of participants (ends vs. middle) on the semantic side, $t(379) = 7.26$, $p < .001$, $d = 0.70$, and on the phonological side, $t(379) = 5.48$, $p < .001$, $d = 0.53$. This gambrel shape is somewhere between a pure pyramid and a ziggurat (i.e., a truncated pyramid). But do the inflection points (from s13p3 to s12p4, and s3p13 to s4p12) represent the true function of list composition on false recall, or do they perhaps reflect differences between the two groups of participants induced by our design (which sought to obtain multiple observations per participant for all compositions while keeping the experiment shorter than one hour)? We will address this question with Experiment 2.

Fig. 4 shows the mean output order of falsely recalled critical items. The overall trend is that when critical items were falsely recalled, they were output sooner (lower output order) for more phonological lists than for more semantic lists. The many missing cells (i.e., list conditions in which a particular participant made no false recall) as well as the two-group within-subjects design made traditional ANOVA and linear regression inappropriate. Thus we ran a linear mixed model analysis with critical output order as the outcome variable, fixed effects of number of semantic associates (0–16) and of group (ends vs. middle), and random intercepts for participant and critical item. This analysis yielded a significant *t* value of 6.72 for the effect of number of semantic associates on output order of false recall, indicating more semantic associates and fewer phonological associates yielded later output position of the critical item. Thus, although semantic and phonological lists appear to yield the same amount of overall false recall of critical items, such recall of the critical item tended to happen earlier in the test for more phonological lists.

Veridical recall

What about veridical recall of list items? Those results tell a straightforward story: the more semantic a list, the better it was recalled. See Fig. 5. For the ends group, there was a significant overall effect of list composition, $F(8,1336) = 36.62$, $MSE = .007$, $p < .001$, $\widehat{\omega}^2 = .063$, as well as a linear trend, $F(1,167) = 173.71$, $MSE = .009$, $p < .001$, $\widehat{\omega}^2 = .069$. For the middle group also, there was a significant overall effect of list composition, $F(8,1696) = 8.47$, $MSE = .006$, $p < .001$, $\widehat{\omega}^2 = .011$, as well as a linear trend, $F(1,212) = 55.29$, $MSE = .007$, $p < .001$, $\widehat{\omega}^2 = .015$. These results are consistent with those of Watson et al. (2001, 2003, Experiment 3).

Strategy and other final questions

As discussed earlier, participants' self-reported prior knowledge of false memories in the DRM paradigm (Question 5, Appendix A) made no difference in their overall level of false recall. In response to Question 3, participants in the middle group estimated they had falsely recalled on a mean of 4.8 lists ($SD = 3.8$; actual $M = 6.5$, $SD = 4.0$, only counting critical false recall), and participants in the ends group estimated they had falsely recalled on a mean of 5.3 lists ($SD = 4.7$; actual $M = 5.8$, $SD = 3.5$). The correlation between estimated and actual number of lists on which false recall occurred for the middle group was $r = .180$, $t(165) = 5.77$, $p < .001$, and for the ends group was $r = .410$, $t(211) = 2.66$, $p = .008$. The significant positive correlations indicate that participants had some insight (metamemory) into how much they had falsely recalled; furthermore, the difference in the correlations indicates that participants in the ends group had more such insight, $z = 2.43$, $p = .015$.

With regard to encoding strategies (Question 4, Appendix A), combining both groups of participants, the ratings for the questions listed in the Appendix were: (a) $M = 2.95$, $SD = 1.32$, (b) $M = 4.08$, $SD = 0.96$, (c) $M = 3.70$, $SD = 1.25$, (d) $M = 2.55$, $SD = 1.35$, (e) $M = 2.95$, $SD = 1.45$, (f) $M = 2.81$, $SD = 1.47$. Between-subjects *t*-tests revealed a significant difference between the two groups of participants only in the case of the last strategy ("I noticed that some words that were similar in both meaning and sound were missing so I tried to figure out the critical missing word."), where a higher rating was given by the middle group ($M = 2.95$, $SD = 1.47$) versus the ends group ($M = 2.63$, $SD = 1.45$), $t(379) = 2.13$, $p = .034$. Essentially, participants using that strategy were attempting to generate the critical word during encoding, using both semantic and phonological cues, at the same time that they were presumably rehearsing the study words being presented to them. Thus the critical word potentially became encoded along with the studied words. If participants forgot that the word had been self-generated, this process would increase the possibility of false recall (and also later underestimation of false recall). That this strategy was used more in the middle group may have been due to that group having experienced lists with greater hybridization, and the ends group having experienced more pure or nearly-pure lists.

The differences between the middle and ends groups in metamemory and strategy use contributed to our curiosity as to whether the design of Experiment 1 may have influenced the overall pattern of false recall, a concern we address in Experiment 2. Despite this concern, the results of Experiment 1 are quite clear in replicating the over-additive levels of false recall from hybrid lists. In addition, the highest level of false recall was when the lists were composed of half semantic associates and half phonological associates.

Experiment 2

The purpose of Experiment 2 was to replicate and clarify the overall picture provided by the results of Experiment 1. The use of two groups in Experiment 1 (ends and

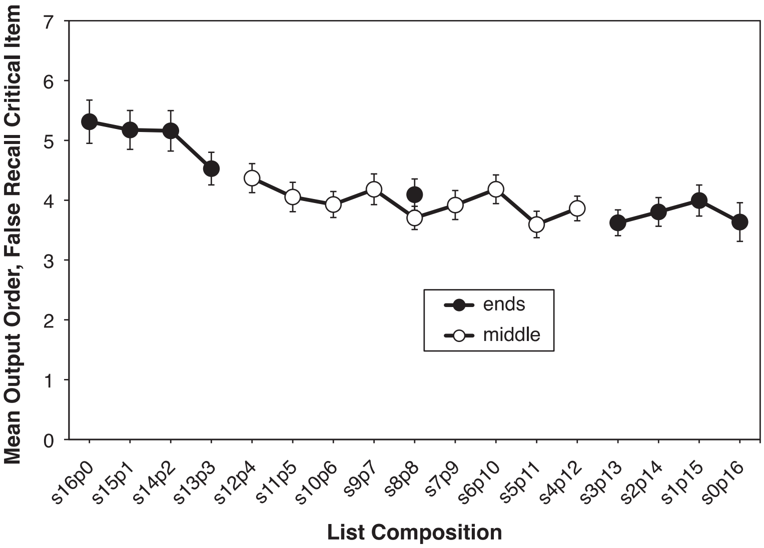

Fig. 4. Mean output order of falsely recalled critical items as a function of list composition (number of semantic and phonological items per list) in Experiment 1. Note: "ends" and "middle" were two separate groups of participants. Error bars represent the standard error of each cell.

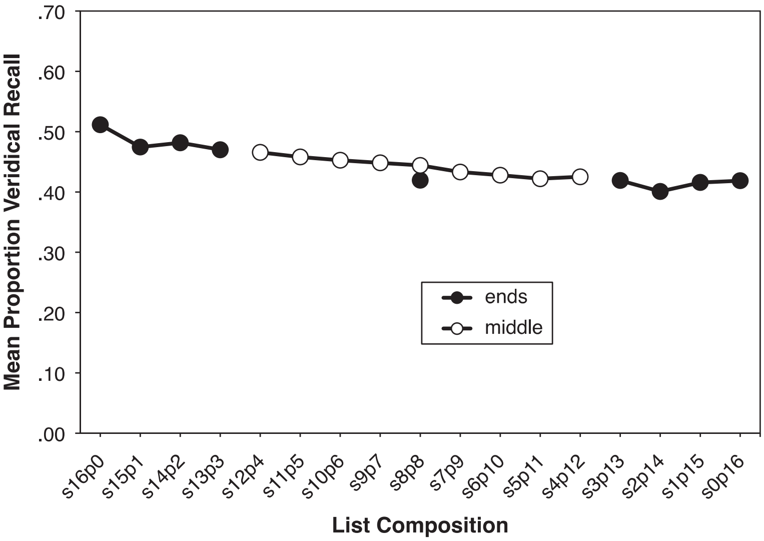

Fig. 5. Mean proportion of veridical recall as a function of list composition (number of semantic and phonological items per list) in Experiment 1. Note: "ends" and "middle" were two separate groups of participants. Standard error bars are too small to be visible.

middle) yielded discontinuities in the data series (from 3 to 4 and 12 to 13 semantic associates). It was unclear to what extent the change in slope at these points (yielding the gambrel shape) may have been due to differences in the experience of participants in the ends versus the middle group. Indeed, it seems that participants in the middle group, who experienced lists with greater hybridization, were more likely to report that during encoding they noticed a related word was missing based on both meaning and sound, perhaps leading to their higher levels of false recall and poorer estimates of false recall.

Furthermore, Experiment 1 used a within-subjects design within each group, and several memory effects are obtained more readily in such designs than in between-subjects designs (McDaniel & Bugg, 2008). Perhaps the act of completing study and free recall for numerous lists influenced the effect of list composition on false recall, for example by enabling all of the encoding strategies that involved noticing a missing related word (based on meaning, sound, or both) and then trying to figure out that word. This is also a potential concern in the experiments by Watson et al. (2003), in which participants studied multiple lists.

Looking at Fig. 3, it is clear that there is an overall quadratic trend: the ends are lower than the middle. But what is the true shape? The data from Experiment 1 suggest a gambrel pyramid, such that increasing hybridization of list composition steadily increases false recall up to the pinnacle at the balanced hybrid list (s8p8), but does so more rapidly at the edges. But we wondered about the possible influences of Experiment 1's design on this overall shape. If the changes in slope were due to the design, the function's shape as assessed in a different way could in fact be a smooth pyramid. Thus, we used a completely between-subjects design in Experiment 2: each participant studied and then recalled only one list in one condition. The use of an internet sample (Mechanical Turk) was ideal for the large sample size required in this kind of experiment, which would be impractical in most laboratory settings.

Method

Participants studied and completed a free recall test for one list of 16 words. The experiment was conducted over the Internet, and was programmed using Adobe Flash ActionScript 3.

Design

The independent variable was the same as in Experiment 1: list composition, ranging from pure semantic (s16p0) to balanced hybrid (s8p8) to pure phonological (s0p16). Thus there were again 17 conditions. Participants were randomly assigned to one of these conditions, such that each condition occurred roughly equally often. The dependent measures were the proportion of critical items falsely recalled. the output order of those items, and veridical recall of studied items.

Participants

Participants were 1199 people recruited from Amazon's Mechanical Turk and paid 30 cents each. There were 678 men and 521 women, and the mean age was 31.7 years ($SD = 10.3$). All participants had completed at least 100 previous HITs, had at least an 85% approval rate, and were located in the United States. Data were collected from an additional 59 participants but were excluded from analysis because they self-reported that English was not their first language, gave zero responses on the test, typed multiple words in the response box instead of pressing enter after each one, or they self-reported that they took notes during the experiment. The number of participants in each list composition condition ranged from 65 to 79 across the 17 list compositions.

Materials

Materials were the same 36 sets of words used in Experiment 1. Sets of words were counterbalanced such that across participants each set was used equally often in each list composition condition. The particular words used to form the study list (16 words) from a given set (32 words) were selected randomly for each participant, within the confines of the particular list composition assigned to that set.

Procedure

The procedure was the same as in Experiment 1 except that each participant studied and was tested on only one list. To keep the procedure short, only final Questions 5–7 (Appendix A) were asked.

Results and discussion

False recall

For false recall, we have only one binary observation per participant: whether or not s/he falsely recalled the critical item (0, 1). Thus for statistical analysis of false recall we could not use ANOVA and *t*-tests as in Experiment 1, but instead use logistic regression and chi-square tests. We report effect sizes as the odds ratio (*OR*) and Cramer's *V*, respectively.

The number of participants who claimed that they previously knew about the DRM effect was 289 out of 1,199. The mean false recall of critical items was not significantly different for participants who claimed prior knowledge of the DRM effect ($M = .44$, $SD = .50$) versus those who did not ($M = .46$, $SD = .50$), $\chi^2(1) = 0.34$, $p = .561$, $V = .02$. Thus we did not exclude data from participants claiming prior knowledge of the effect.

Fig. 6 shows mean proportion of false recall of the critical item as a function of list composition (see also Appendix B). A polynomial logistic regression revealed no significant linear effect, $\beta = 0.96$, $SE = 2.04$, Wald's $z = 0.473$, $p = .636$, $OR = 2.62$, but a significant quadratic effect, $\beta = -7.35$, $SE = 2.04$, Wald's $z = -3.601$, $p < .001$, $OR = .0006$. The overall shape appears in fact to be a ziggurat rather than a gambrel or smooth pyramid. Purely semantic and purely phonological lists again yielded equivalent levels of false recall, $\chi^2(1) = 0.17$, $p = .677$, $V = .04$. Hybrid lists yielded higher false recall than pure lists, $\chi^2(1) = 14.63$, $p < .001$, $V = .11$. Yet increasing hybridization did not steadily increase false recall; there was no apex at the balanced hybrid condition (s8p8) as suggested by Experiment 1. False recall in this condition did not significantly differ from that in all the other conditions combined, $M_{all\_other} = .46$, $SD_{all\_other} = .50$, $\chi^2(1) = 0.003$, $p = .960$, $V = .001$. Indeed, the effect of hybridization appears to plateau after just one or two inclusions of the other type of associate. This outcome again suggests that activation of two different associative networks (semantic and phonological/lexical) provides a boost to false recall even with just one or two associates, in line with a modified activation/monitoring framework which we will propose in the General Discussion.

What about symmetry? The absolute values of the estimated coefficients of the linear effect from logistic regression (similar to the slopes in linear regression) are very similar for the semantic side ($|\beta| = 0.056$, $SE = 0.031$) and the phonological side ($|\beta| = 0.069$, $SE = 0.032$) of Fig. 6, though there is no current consensus on an acceptable inferential test to compare such coefficients (Mood, 2010). However, the overall picture is consistent with equivalent contributions of semantic and phonological associates to false recall.

Fig. 7 shows the mean output order of falsely recalled critical items. The overall trend replicates that found in

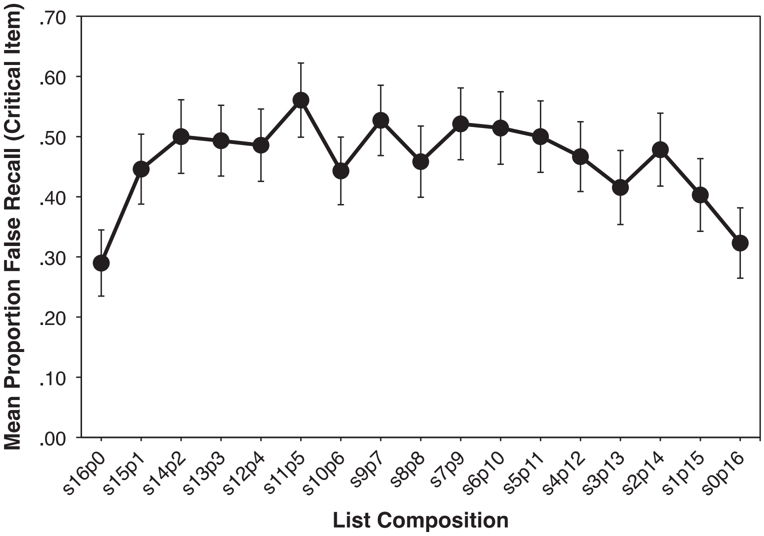

Fig. 6. Mean proportion of false recall of critical item as a function of list composition (number of semantic and phonological items per list) in Experiment 2. Note: manipulation was entirely between-subjects. Error bars represent the standard error of each cell.

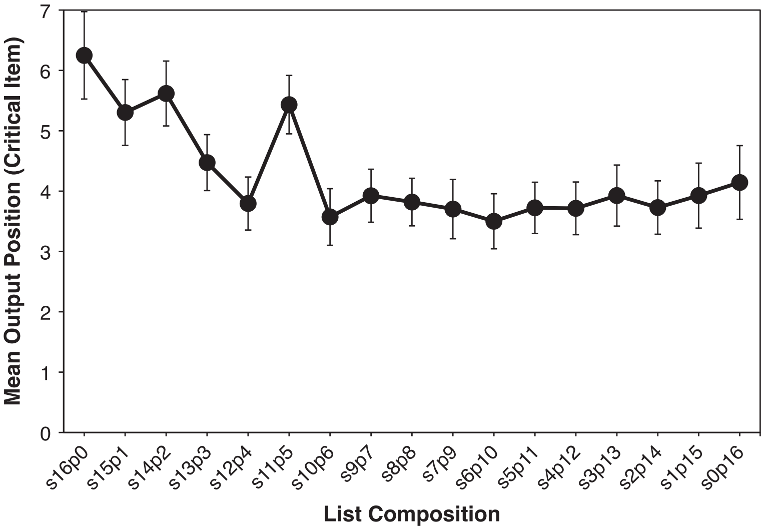

Fig. 7. Mean output order of falsely recalled critical items as a function of list composition (number of semantic and phonological items per list) in Experiment 2. Note: manipulation was entirely between-subjects. Error bars represent the standard error of each cell.

Experiment 1. When critical items were falsely recalled, they were generally output earlier as a function of the number of phonological items within the list. (One data point, s11p5, seems a clear outlier for which we have no explanation.) A one-way between-subjects ANOVA revealed a significant overall effect of list composition, $F(16,536) = 2.70$, $MSE = 7.717$, $p < .001$, $\widehat{\omega}^2 = .047$, a significant linear trend, $F(1,536) = 18.57$, $MSE = 7.717$, $p < .001$, $\widehat{\omega}^2 = .030$, and a significant quadratic trend, $F(1,536) = 11.57$, $MSE = 7.717$, $p = .001$, $\widehat{\omega}^2 = .018$. The overall trend of false recall output order found in both experiments is interesting because although pure phonological and pure semantic lists yielded false recall at equivalent rates, the phonological false recalls were more likely to occur earlier in recall than the semantic false recalls. This pattern may reflect distinct phonological/lexical and semantic associative networks with somewhat different properties operating during recall (e.g., phonological information is more available for early output from a short-term buffer than semantic information which is more available for long-term memory; see Crowder, 1976; Kintsch & Buschke, 1969).

Veridical recall

Fig. 8 shows veridical recall as a function of list composition. The figure is much more regular (with smaller standard errors) than the false recall data in Fig. 6. This is due to the fact that potentially 16 studied items are included in

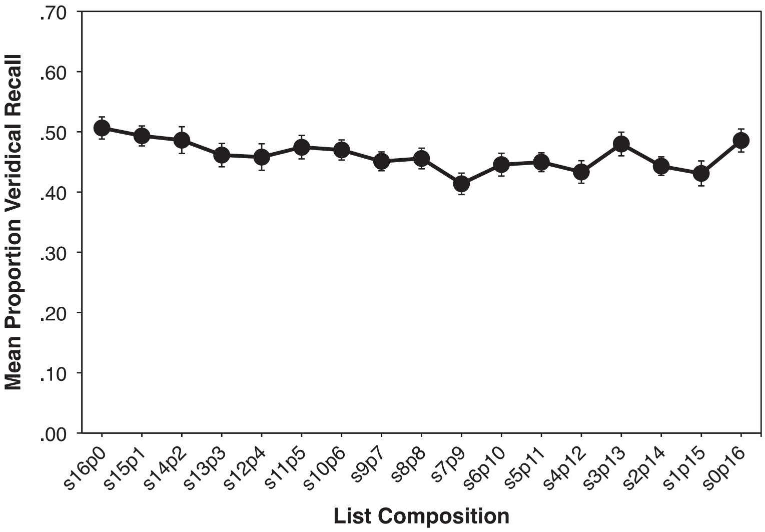

Fig. 8. Mean proportion of veridical recall as a function of list composition (number of semantic and phonological items per list) in Experiment 2. Note: manipulation was entirely between-subjects. Error bars represent the standard error of each cell.

veridical recall whereas only one unstudied item is scored for false recall; hence the latter data will always be noisier. A one-way between-subjects ANOVA revealed a significant overall effect of list composition, $F(16,1182) = 1.76$, $MSE = .024$, $p = .032$, $\widehat{\omega}^2 = .047$, a significant linear trend, $F(1,1182) = 7.42$, $MSE = .024$, $p = .007$, $\widehat{\omega}^2 = .005$, and a significant quadratic trend, $F(1,1182) = 8.38$, $MSE = .024$, $p = .004$, $\widehat{\omega}^2 = .006$. Overall there was a slight decline in veridical recall as list composition became less semantic, but with an uptick for the pure phonological condition.

General discussion

The current study investigated false recall produced by word lists composed of varying proportions of semantic and phonological associates to an unstudied critical word. In two experiments we systematically manipulated list composition to explore the entire range of hybrid lists as well as both types of pure lists. Lists were always 16 words in length, and ranged from pure semantic (s16p0) to balanced hybrid (s8p8) to pure phonological (s0p16) and everything in between. Using this design, we replicated past findings and added significantly to them. In particular, we replicated two basic findings of Watson et al. (2001, 2003): first that pure semantic and pure phonological lists yielded equivalent levels of false recall, and second that hybrid lists in general yielded more false recall than the pure lists.

An important contribution of our current results was to fill in and clarify the overall shape of the false recall function for all of the compositions of hybrid semantic and phonological lists. The results suggest that outcomes may differ for within- versus between-subject designs. Using a partially within-subjects design (with “ends” and “middle” of the list compositions represented between-subjects), the resulting graph of false recall as a function of list composition assumed what we termed a gambrel pyramid shape. That is, as seen in Fig. 3, the pyramid shape was marked by elevated false recall on both sides by including only one or two items of the other type in a list. Experiment 2 assessed whether we could replicate this pattern with an entirely between-subjects design in which participants studied and recalled a list using only one of the 17 possible compositions. The results from Experiment 2 confirmed the powerful nature of hybrid lists relative to pure lists in yielding false recall, but the observed function yielded a roughly symmetrical ziggurat shape (Fig. 6). Once again, we replicated the result from Experiment 1 that simply including one or two phonological (or semantic) words in an otherwise pure list of semantic (or phonological) associates greatly increased false recall.

The false recall data from Experiment 2 are perforce somewhat noisy, because each participant contributed only one datum to his or her respective condition. Thus, to further assess the overall pattern of data, we combined the data from Experiment 2 with the data from just the first list that each participant completed in Experiment 1, which gave us a total sample size of 1580 (with a mean of 93 participants per condition for the 17 conditions). The results are shown in Fig. 9 and confirm the data in Fig. 6 from Experiment 2. Adding just one or two phonological associates in an otherwise semantic list, or one or two semantic associates in an otherwise phonological list, greatly increases false recall that seems to plateau at about 50%. Overall, semantic and phonological associates appear to contribute equally to the over-additive levels of false recall.

What about possible explanations for the over-additivity? Our findings are consistent with the spreading activation explanations given by Watson et al. (2003) that we described in the Introduction. To speak in terms of converging associative networks, it appears that activation saturates rapidly from both lexical (phonological/orthographic) and semantic networks, and that hybrid lists acti-

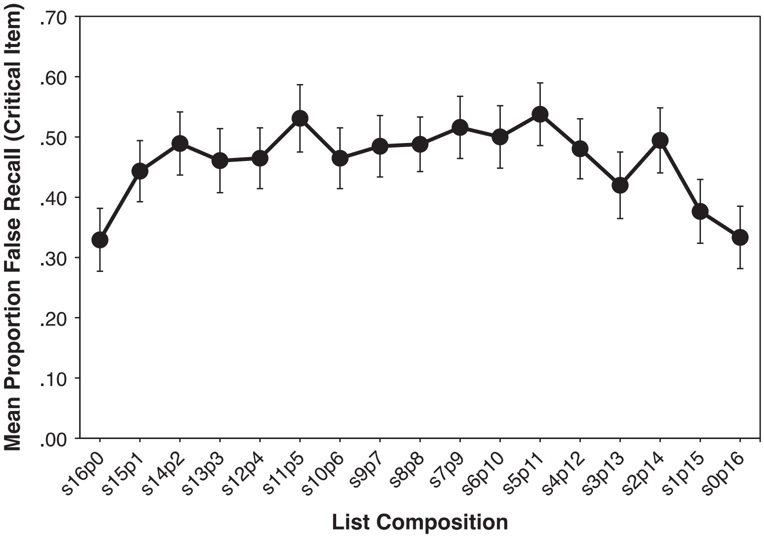

Fig. 9. Mean proportion of false recall of critical item as a function of list composition (number of semantic and phonological items per list) combined across Experiments 1 and 2. Only data from a participant's first list are included from Experiment 1. Error bars represent the standard error of each cell. $N = 1580$ (mean 93 per condition).

vate representations in both networks. As illustrated in the hypothetical example of Fig. 1, steep activation functions mean that a small initial number of associates from the alternative network can contribute substantially to activation of the critical item at encoding, and that past a certain point additional associates of either type will yield diminishing returns in how much they increase false recall. That is precisely the pattern of results we obtained: negatively accelerated slopes on both sides of the function.

The activation/monitoring framework outlined in the Introduction requires one significant adjustment to provide a reasonable account of our results (specifically, the symmetry of the function and its steepness on both ends). Recall that a basic idea of the framework is that when a candidate item is generated from conceptual (semantic) activation during retrieval, if it lacks surface level familiarity in terms of phonological features, that basis can be used to reject the item as not studied; but once phonological associates are included in the study list, that diagnostic monitoring strategy is impaired. This framework works for the increasing false recall on the semantic side of the function shown in Figs. 3, 6 and 9 (left side of function). However, the same kind of increase occurs on the phonological side with the inclusion of semantic items, an outcome not anticipated in the original framework. The activation/monitoring framework could be expanded such that, for pure phonological lists, semantic unfamiliarity could be used to reject candidates generated from phonological activation, and including semantic associates in the study list likewise impairs this strategy. For example, if a participant studied words like *cheer* and *hair*, later at retrieval she might generate the critical unstudied word *chair* based on phonological activation, but she may still be able to correctly reject that word and choose not to output it because the concept of *chair* does not feel sufficiently familiar. Including semantic associates (e.g., *sit* and *desk*) in the list may disrupt such semantic diagnostic monitoring, just as phonological associates disrupt phonological diagnostic monitoring. We admit that this account is post hoc. Further research is needed to investigate the extent to which participants may engage in different kinds of diagnostic monitoring strategies based on list composition. Nonetheless, the data are certainly consistent with the possibility that semantic monitoring strategies are used and would thus be disrupted by semantic associates. This possibility constitutes a significant and intriguing modification to the activation/monitoring framework. Alternatively, it may be that diagnostic monitoring plays a greater role on one side of the function (adding phonological associates to semantic lists) while automatic activation plays a greater role on the other side of the function (adding semantic associates to phonological lists).

We note again, as in the Introduction, that fuzzy trace theory, another prominent theory of DRM false memories, seems mute on the issue of why hybrid lists produce such powerful false memory phenomena. Within the theory, gist representations, which are purely semantic, are hypothesized to give rise to false memories. However, our current research and that of many others shows that false memories arise from phonological lists as readily as from semantic lists. In our study and in Watson et al. (2003), false recall from pure lists of phonological associates equaled that from pure semantic associates. Furthermore, there is no readily apparent way for fuzzy trace theory to account for the observed shape of the function, with false recall being highest for hybrid relative to pure lists. Thus, the current findings support activation-based theoretical accounts over fuzzy trace theory.

Although the current experiments provide a start in understanding the power of hybrid lists in producing false recall, much remains to be discovered. Just because the contribution to overall levels of false recall in hybrid lists appears to be equivalent for the two types of associates does not mean that both types operate via the same underlying processes. Indeed, intriguing data from several studies suggest that they do not. In the current study, false recall of critical items tended to occur earlier in lists with mostly phonological associates relative to lists with more semantic associates. In addition, Watson et al. (2003, Experiment 3) found that false recognition was more often accompanied by "remember" than "know" judgments for semantic lists, and that the opposite occurred for phonological lists. Studies by McDermott and Watson (2001) and by Ballardini, Yamashita, and Wallace (2008) found that presentation rate (ranging from 20 ms to 5000 ms) has a different effect on false recall for semantic versus phonological lists: for semantic lists, false recall starts low at 20 ms, peaks at 250 ms, and declines from there; for phonological lists, false recall starts high at 20 ms and declines monotonically. Tse, Li, and Neill (2011) examined *d′* (recognition discriminability) for critical items by creating some lists in which the critical items were studied and some in which they were not. They found that critical item *d′* was lower than the yoked associate *d′* for semantic lists, but higher than the yoked associate *d′* for phonological lists. They argued for distinct activation mechanisms, as did Ballou and Sommers (2008) who found little correlation between participants' level of false memory from semantic versus phonological lists. Garoff-Eaton, Kensinger, and Schacter (2007) found that the frontal cortex showed more activation during false recognition from conceptually-related word triplets versus from perceptually-related word triplets. Finally, there is a suggestion of different developmental trajectories for false memory from the two types of lists, such that false memory increases with age for semantic lists and decreases with age for phonological lists (Dewhurst & Robinson, 2004; Holliday & Weekes, 2006). However, Swannell and Dewhurst (2012) did show an increase in false recall with age for phonological lists that properly converge on a single critical lure (as with the phonological lists in the current study, Watson et al., 2003, and Sommers & Lewis, 1999).

Future research will be needed to determine why these differences occur between false memories produced by semantic and phonological lists. We can but speculate about how the activation/monitoring framework could account for such results. Perhaps associates activate lexical and semantic networks somewhat differently, for example based on the content or speed of activation of the network (cf. associative activation theory, Howe, Wimmer, Gagnon, & Plumpton, 2009). And/or perhaps there are differences in the processes of using phonological versus semantic information to reject generated items (cf. Gallo, 2010, Fig. 4). Further research may untangle these issues, but must also account for the similarity found on the semantic and phonological sides of the false recall function across hybrid list compositions (Figs. 3, 6 and 9).

Hybrid lists provide a valuable and thus-far underutilized tool in the ongoing efforts to understand the processes underlying false memory. Although the standard purely semantic DRM lists produce striking levels of false recall, hybrid lists show much more. Holding list length constant as in the current experiments, false recall rates were 33% with pure lists and 48% with hybrid lists. The current study has clarified the form of the false recall function across varying degrees of list hybridization, and the shape of that function provides guidance and inspiration for further research.

Author note

This research was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind, and Behavior Collaborative Award to Henry L. Roediger, III.

Appendix A

Final questions given to participants in Experiment 1.

1. What do you think was the purpose of this experiment?
2. Did you notice anything special or unusual about the lists of words?
3. On how many of the 18 tests do you think you might have remembered a word that was not actually studied in that list?
4. How much did you use each of the following memory strategies when you studied the lists? 1 = not at all, 5 = extensively
   a. I didn't use a strategy; I just tried to remember as many words as I could.
   b. I rehearsed the words in my mind.
   c. I formed images or associations to link the words together.
   d. I noticed that some associated words were missing, and I tried to figure these out as each list was presented.
   e. I noticed that some words that sounded like the other words were missing so I tried to figure out which word was missing based on its sound.
   f. I noticed that some words that were similar in both meaning and sound were missing so I tried to figure out the critical missing word.
   g. Other:
5. When people study certain lists of words, they often remember a word that was associated with all of the studied words but was not in fact in the study list. For example, people might incorrectly remember the word "sleep" when they studied the following list: bed, rest, awake, tired, dream, wake, snooze, blanket, doze, slumber, snore, nap, peace, yawn, drowsy. Did you already know about this effect before? [yes/no]
6. (Optional) Any additional final thoughts or comments about the experiment.
7. (Optional) Did you write down any notes at any point in the experiment? [yes/no]

Appendix B

Critical false recall as a function of list composition in Experiments 1 and 2.

Exp.	Group	N	List composition																
			s16p0	s15p1	s14p2	s13p3	s12p4	s11p5	s10p6	s9p7	s8p8	s7p9	s6p10	s5p11	s4p12	s3p13	s2p14	s1p15	s0p16
1	ends	*168*	.21	.31	.29	.36					.44					.38	.35	.32	.23
			(.33)	(.35)	(.34)	(.37)					(.38)					(.36)	(.36)	(.35)	(.33)
	middle	*213*					.32	.34	.36	.38	.40	.37	.36	.38	.35				
							(.35)	(.36)	(.35)	(.37)	(.37)	(.36)	(.38)	(.37)	(.37)				
2		*1199*	.29	.45	.50	.49	.49	.56	.44	.53	.46	.52	.51	.50	.47	.42	.48	.40	.32
			(.45)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.49)	(.50)	(.49)	(.47)
		n:	*69*	*74*	*68*	*73*	*70*	*66*	*79*	*74*	*72*	*71*	*70*	*72*	*75*	*65*	*69*	*67*	*65*

Note. Data are means with standard deviations in parentheses. s and p indicate the number of semantic and phonological associates per list, respectively. Sample sizes per composition condition are given for Experiment 2 as *n*.

References

Ballardini, N., Yamashita, J. A., & Wallace, W. P. (2008). Presentation duration and false recall for semantic and phonological associates. *Consciousness and Cognition, 17*, 64–71. http://dx.doi.org/10.1016/j.concog.2007.01.008.

Ballou, M. R., & Sommers, M. S. (2008). Similar phenomena, different mechanisms: Semantic and phonological false memories are produced by independent mechanisms. *Memory & Cognition, 36*, 1450–1459. http://dx.doi.org/10.3758/MC.36.8.1450.

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. http://dx.doi.org/10.1037/0278-7393.22.4.827.

Brainerd, C. J., Payne, D. G., Wright, R., & Reyna, V. F. (2003). Phantom recall. *Journal of Memory and Language, 48*, 445–467. http://dx.doi.org/10.1016/S0749-596X(02)00501-6.

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science, 11*, 164–169. http://dx.doi.org/10.1111/1467-8721.00192.

Budson, A. E., Sullivan, A. L., Daffner, K. R., & Schacter, D. L. (2003). Semantic versus phonological false recognition in aging and Alzheimer's disease. *Brain and Cognition, 51*, 251–261. http://dx.doi.org/10.1016/S0278-2626(03)00030-7.

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

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

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*(3), 283–321. http://dx.doi.org/10.1037/0033-295X.93.3.283.

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. http://dx.doi.org/10.1111/j.0956-7976.2004.00756.x.

Gallo, D. A. (2004). Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 120–128. http://dx.doi.org/10.1037/0278-7393.30.1.120.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition, 38*, 833–848. http://dx.doi.org/10.3758/MC.38.7.833.

Gallo, D. A., McDermott, K. B., Percer, J. M., & Roediger, H. I. (2001). Modality effects in false recall and false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 339–353. http://dx.doi.org/10.1037/0278-7393.27.2.339.

Garoff-Eaton, R. J., Kensinger, E. A., & Schacter, D. L. (2007). The neural correlates of conceptual and perceptual false recognition. *Learning & Memory, 14*(10), 684–692. http://dx.doi.org/10.1101/lm.695707.

Holliday, R. E., & Weekes, B. S. (2006). Dissociated developmental trajectories for semantic and phonological false memories. *Memory, 14*, 624–636. http://dx.doi.org/10.1080/09658210600736525.

Howe, M. L., Wimmer, M. C., Gagnon, N., & Plumpton, S. (2009). An associative-activation theory of children's and adults' memory illusions. *Journal of Memory and Language, 60*(2), 229–251. http://dx.doi.org/10.1016/j.jml.2008.10.002.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 391–422). Hillsdale, NJ: Erlbaum.

Kellogg, R. T. (2001). Presentation modality and mode of recall in verbal false memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(4), 913–919. http://dx.doi.org/10.1037/0278-7393.27.4.913.

Kintsch, W., & Buschke, H. (1969). Homophones and synonyms in short-term memory. *Journal of Experimental Psychology, 80*(3), 403–407. http://dx.doi.org/10.1037/h0027477.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15*(2), 237–255. http://dx.doi.org/10.3758/PBR.15.2.237.

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

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*(1), 67–82. http://dx.doi.org/10.1093/esr/jcp006.

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

Roediger, H. L., Balota, D. A., & Watson, J. M. (2001). Spreading activation and the 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 Press.

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. http://dx.doi.org/10.1037/0278-7393.21.4.803.

Roediger, H. L., 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. http://dx.doi.org/10.3758/BF03196177.

Sommers, M. S., & Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language, 40*, 83–108. http://dx.doi.org/10.1006/jmla.1998.2614.

Swannell, E. R., & Dewhurst, S. A. (2012). Phonological false memories in children and adults: Evidence for a developmental reversal. *Journal of Memory and Language, 66*(2), 376–383. http://dx.doi.org/10.1016/j.jml.2011.11.003.

Tse, C., Li, Y., & Neill, W. T. (2011). Dissociative effects of phonological vs. semantic associates on recognition memory in the Deese/Roediger–McDermott paradigm. *Acta Psychologica, 137*(3), 269–279. http://dx.doi.org/10.1016/j.actpsy.2011.01.013.

Watkins, M. J., & Gardiner, J. M. (1979). An appreciation of generate–recognize theory of recall. *Journal of Verbal Learning and Verbal Behavior, 18*(6), 687–704. http://dx.doi.org/10.1016/S0022-5371(79)90397-9.

Watson, J. M., Balota, D. A., & Roediger, H. L. (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. http://dx.doi.org/10.1016/S0749-596X(03)00019-6.

Watson, J. M., Balota, D. A., & Sergent-Marshall, S. D. (2001). Semantic, phonological, and hybrid veridical and false memories in healthy older adults and in individuals with Dementia of the Alzheimer type. *Neuropsychology, 2*, 254–267. http://dx.doi.org/10.1037/0894-4105.15.2.254.