Available online at www.sciencedirect.com
ScienceDirect

Journal of Memory and Language 55 (2006) 461–478

Journal of Memory and Language

www.elsevier.com/locate/jml

The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect ☆

Amy H. Criss

Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University,
115 Mellon Institute, 4400 5th St., Pittsburgh, PA 15213, USA

Received 1 March 2006; revision received 7 August 2006
Available online 18 September 2006

Abstract

When items on one list receive more encoding than items on another list, the improvement in performance usually manifests as an increase in the hit rate and a decrease in the false alarm rate (FAR). A common account of this strength based mirror effect is that participants adopt a more strict criterion following a strongly than weakly encoded list (e.g., Cary & Reder, 2003; Stretch & Wixted, 1998). Differentiation models offer an alternative: more encoding leads to a more accurate memory representation for the studied item. A more accurate representation is less confusable with an unrelated item, resulting in a decrease in the FAR (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Differentiation models make additional predictions about reversals in FARs for foils similar to a studied item as a function of the composition of the study list. These predictions were empirically tested and confirmed.

Keywords: Memory models; Recognition memory; Differentiation; Mirror effect; Criterion shifts; False memory

When some manipulation results in two different levels of performance in a recognition memory task, the different levels of performance are typically expressed as a mirror pattern (e.g., Glanzer & Adams, 1985). That is, the probability of correctly claiming that a target item was studied (i.e., hit rate, HR) mirrors the probability of erroneously claiming that a foil item was studied (i.e., false alarm rate, FAR). Mirror effects are ubiquitous and have been observed for normative word frequency, part of speech, word concreteness, rated typicality, known versus unknown scenes, and several other manipulations (e.g., Dobbins & Kroll, 2005; Glanzer & Adams, 1985, 1990; Hockley, 1994; Stretch & Wixted, 1998; Vokey & Read, 1992). The focus of this research is the strength based mirror effect where different levels of performance are obtained by manipulating encoding time. For example, suppose one group of participants studies each item once (weak list) and another group studies each item five times (strong list). The strong list tends to produce both higher HRs and lower FARs than the weak list. The strength based mirror effect has been the focus of much recent discussion and has been observed when strength is manipulated by study time or by repetition and for both single item and associative recognition (e.g., Cary & Reder, 2003; Hockley & Niewiadomski, in press; Kim & Glanzer, 1993; Stretch & Wixted, 1998).

The simple fact that participants are better to identify an item that received more encoding is not too surprising. Of more theoretical interest is why the FAR changes

☆ This research was supported by National Institutes of Mental Health Grant MH019983.

E-mail address: acriss@andrew.cmu.edu

0749-596X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2006.08.003

between the two conditions. Why should encoding conditions affect the response to items that were not on the study list? The most common answer is that the participant adopts a more stringent criterion for calling at item "studied" following a strong list than following a weak list. For example, many assume that recognition memory can be thought of as a case of signal detection theory (SDT; Benjamin & Bawa, 2004; DeCarlo, 2002; Dobbins & Kroll, 2005; Dunn, 2004; Green & Swets, 1966; Morrell, Gaitan, & Wixted, 2002; Stretch & Wixted, 1998; Verde & Rotello, in press). In this framework, the subjective response (also referred to as familiarity, strength, or global match) to targets and foils can be represented by two overlapping normal distributions as illustrated in Fig. 1. Participants select some criterion and any item evoking a subjective response greater than the criterion is called "studied" while other items are called "not studied." Additional study increases the mean of the target distribution, hence the increase in the HR. However, the foil distribution does not change as a function of encoding conditions, evident in the single foil distribution in Fig. 1. Within this framework, the only way to change the FAR as a function of the strength of the study list is to assume a change in the criterion. In the figure, the criterion for the weak list is shown as a solid line and the more stringent criterion for the strong list is shown as a dashed line.

The preceding discussion refers to a single process model where a recognition memory decision is based on the overall familiarity of the test item. Dual process models assume two different retrieval routes and the decision can be based on either route. For example, in the Source of Activation Confusion model (SAC), HRs are based on recollecting the details of the study event and FARs are based on the pre-experimental familiarity of the test item (Reder et al., 2000). Additional study time improves the ability to recollect and increases the HR. However, the pre-experimental strength of the item is not affected by the study list. In order to account for a reduced FAR in the strong list, Cary and Reder (2003) assume a criterion shift. Thus, both single process (e.g., Stretch & Wixted, 1998) and dual process models (e.g., Cary & Reder, 2003) attribute the strength based mirror effect to a change in the criterion between lists. Indeed, all models require a criterion for responding "old" and could adopt the criterion change assumption to account for the strength based mirror effect.

Differentiation models provide an alternate account, one that is not dependent on strategic criterion shifts, but is a natural consequence of the encoding process. In these models, additional experience with an item in a given context results in updating a single memory trace. The more accurate a memory trace, the less similar it is to unrelated items. Thus, the match between an unrelated foil and episodic memory is lower following a strong list than a weak list. There are two important differences between the criterion placement account and the differentiation account of the strength based mirror effect. In the former, the effect results from the *decision* process and might be influenced by external pressures such as costs and rewards, instructions given by the experimenter, age of participants, emotional valence of the stimuli, etc. In the differentiation account, the phenomenon naturally follows from the *encoding* process and should not be subject to the whims of the

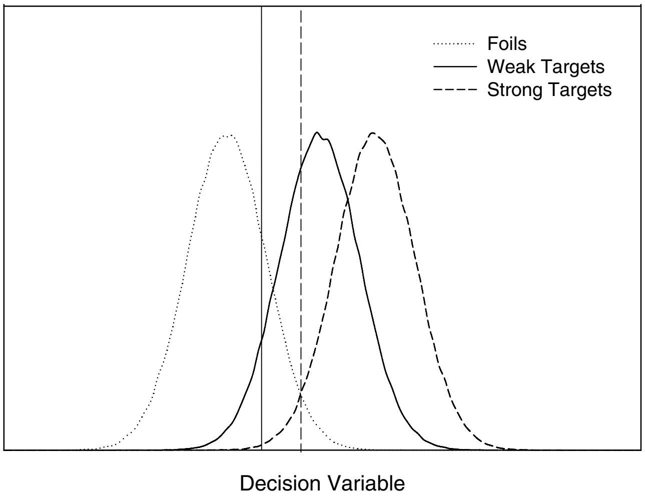

Fig. 1. An example of a signal detection theory account of the strength based mirror effect. The foil distribution is fixed regardless of the encoding conditions but the criterion changes between the two lists producing a lower FAR for the strong compared to the weak list. The dashed line is the criterion for the strong list and the solid line is the criterion for the weak list.

participant but is a necessary result of the processes underlying storage. Second, the former assumes a single familiarity distribution for foils regardless of list strength while the latter yields two different familiarity distributions for foils following a weak and strong list.

Differentiation was first introduced by Gibson (1940; 1969; Gibson & Gibson, 1955) in the realm of perceptual learning and later incorporated into models of episodic memory as an explanation for the null list strength effect in recognition memory (Murnane & Shiffrin, 1991; Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990). I use the label differentiation models to refer to both the Subjective Likelihood Model (SLiM; McClelland & Chappell, 1998) and the Retrieving Effectively from Memory model (REM; Shiffrin & Steyvers, 1997). These models do differ in critical aspects that allow them to make differential predictions (cf. Criss & McClelland, in press), but they share the assumption of differentiation and make similar qualitative predictions for issues considered here. For clarity and brevity, I present only the REM model.

The goal of this paper is to present differentiation models as an alternative to the criterion change assumption and to test additional predictions generated by differentiation models. First, I describe REM in general, then how it naturally predicts a strength based mirror effect, and finally I present novel predictions generated by the model and empirical tests of the predictions.

REM

The REM model has been extensively described in other papers and readers are referred to the original source for more details and discussion (e.g., Shiffrin & Steyvers, 1997). Here I briefly describe the model with a focus on the specific simulations reported in this paper. REM assumes two types of memory traces. Knowledge is stored in lexical/semantic traces. Lexical/semantic traces are the lifelong accumulation of episodes involving the stimulus and are complete, accurate, and decontextualized relative to episodic traces. Episodic traces are formed during study and are updated with both item and context features during each successive study presentation. A recognition memory test involves probing with the reinstated context features to restrict (more or less) the comparison to the relevant episodic traces (i.e., those from the study list) and probing with the item features to allow a decision about whether the specific test item was presented. Each of these steps is described in detail below.

Representation

The presented stimulus (at either study or test) is assumed to be an accurate copy of the lexical/semantic representation. Items are represented as a vector of features ($M = 20$) each independently drawn from the geometric distribution with some parameter ($g = 0.35$). Features are abstract and might include orthography, phonology, semantics, reference to personal history, and other information. The probability that a feature takes the value v is

$$P(v) = g(1-g)^{v-1} \qquad (v > 0). \tag{1}$$

Storage

During study, some of the lexical/semantic features are stored in an episodic memory trace along with the current context features. All episodic memory features begin as zeros indicating a lack of information. During a study presentation, each zero is replaced by a feature value with some probability ($u = 0.335167$).[1] The correct value corresponding to the lexical/semantic feature of the study item is stored with some probability ($c = 0.70$). Otherwise, a random feature value is selected from the geometric distribution and stored. Thus, episodic memory is incomplete (i.e., some zeros remain following study), prone to error (i.e., an incorrect feature value may be stored), and context-bound (i.e., contains a set of features representing the context). Once a feature is stored, its value is fixed and will not change during the course of the experiment. Additional study results in the storage of more features but not the correction of previously stored features.

Retrieval

For the present purposes, I adopt the simplification that context features perfectly isolate the study list and do not consider them further (cf., REM.1 in Shiffrin & Steyvers, 1997). At test, the lexical/semantic vector corresponding to test item j is compared to each trace stored in memory, indexed by i, and a likelihood ratio is computed as follows:

$$\lambda_{(i,j,k)} = (1-c)^{nq_{(i,j,k)}} \prod_{v=1}^{\infty} \left[\frac{c + (1-c)g(1-g)^{v-1}}{g(1-g)^{v-1}} \right]^{nm(v,i,j,k)}. \tag{2}$$

This is the REM equation for the likelihood that the test stimulus j matches memory trace i for simulated participant k. The number of non-zero features that mismatch is nq and the number of non-zero features that match

[1] The original paper (Shiffrin & Steyvers, 1997) contained two redundant parameters which I report as one parameter. The original manuscript contained a parameter for the probability of storing a feature (u^*) in each time step (t). These reduce to $u = 1 - (1 - u^*)^t$. Shiffrin and Steyvers (1997) used values of $u^* = 0.04$ and $t = 10$ or $u = .335167$.

and have the value v is nm. Features with a value of zero do not contribute to the decision. The decision about whether test stimulus j was studied or not is based on the odds which is simply the average of the likelihood ratios

$$\Phi_{j,k} = \frac{1}{N}\sum_{i=1}^{N}\lambda_{(i,j,k)}, \tag{3}$$

where N is the number of episodic memory traces. If the odds is greater than some criterion ($criterion = 1$), test stimulus j is called "old" otherwise it is called "new."

The strength based mirror effect in REM

Critically, when an item is repeated within a given context, such as a study list, differentiation models assume that the same episodic memory trace is updated. Each presentation results in a more complete and more accurate representation of the studied item in episodic memory. This is in sharp contrast to many other models that assume each study event results in the storage of an additional memory trace (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988; Humphreys, Pike, Bain, & Tehan, 1989; Metcalfe-Eich, 1985; Murdock, 1997; Murdock, Smith, & Bai, 2001; Nosofsky, 1988). For example, suppose the word coat was studied five times during a single study list. The latter class of models assumes that five noisy copies of the concept coat are stored. In contrast, the differentiation models assume that a single trace will be stored and information will be added to that trace with each repetition. It is this assumption that produces differentiation and the strength based mirror effect as I shall now illustrate by way of an example.

The example in Fig. 2 shows the resulting episodic traces of six items studied once (left side of the figure) and those same six items studied multiple times (right side of the figure). There are two test items: j_1 is a target corresponding to memory trace i_1 and j_7 is a foil that is similar to the studied item stored in memory trace i_1. Each item has 10 features displayed vertically. The bottom two rows show the likelihood ratio for the match between the test item and the memory trace in that column. The likelihood ratio is computed using Eq. (2) with $g = .35$ and $c = .7$. A higher likelihood ratio indicates a greater match between the memory trace and the test probe. As outlined above, some features were not stored in episodic memory indicated with a zero, some features were stored with the correct value, and some were stored with the incorrect value.

Accrual of features in a single trace following multiple study opportunities has two consequences: (1) The match between the target test probe (j_1) and the corresponding memory trace stored during study of that same item (i_1) will be a better match following multiple presentations (a higher likelihood ratio on the right side of the figure than the left) and (2) The match between the target test probe (j_1) and the memory traces stored during study of other items (i_2, i_3, i_4, i_5, or i_6) will be a poorer match following multiple encoding opportunities (a lower likelihood ratio on the right side of the figure than the left). This is differentiation. When the test probe is a target these two factors are in competition: the increased match to the corresponding memory trace with additional study would result in a higher HR, but the decreased match to the remaining memory traces would result in a lower HR. Because the match between

Target	Similar Foil	Episodic Memory Following a Single Study Presentation						Episodic Memory Following Multiple Study Presentations					
j_1	j_7	i_1	i_2	i_3	i_4	i_5	i_6	i_1	i_2	i_3	i_4	i_5	i_6
2	2	0	0	2	0	4	0	3	0	2	1	4	1
3	3	3	2	0	0	0	7	3	2	19	1	1	7
4	3	0	0	0	0	0	2	0	2	0	9	1	2
3	1	0	0	0	6	0	0	3	2	0	6	8	5
2	1	0	3	0	2	0	0	1	3	1	2	0	1
1	3	1	3	0	3	0	0	1	1	2	3	0	2
1	1	1	0	0	0	1	1	1	0	9	2	1	1
3	3	0	1	1	0	0	0	5	1	1	3	1	9
7	7	0	0	0	0	0	0	7	1	6	1	2	0
4	4	5	0	3	0	0	0	5	2	3	7	1	0
Match to Target λ_{j1}		7.9885	0.0081	0.3039	0.3039	0.69	0.207	29.1176	0.0005	0.0007	0.0011	0.0005	0.0005
Match to Similar Foil λ_{j7}		1.042	0.1359	0.3039	0.453	0.69	0.207	1.7353	0.0001	0.0057	0. 0017	0.0005	0.0039

Fig. 2. A numerical example of the features stored in REM following study of six weakly encoded items (left side) and those same items studied many times (right side). The bottom two rows are the resulting likelihood ratios (e.g., match to memory) for a comparison between each stored memory trace and test probe. Memory trace i_1 was stored following study of target test probe j_1 and is similar to foil test probe j_7.

a target and its corresponding memory trace tends to be very large relative to the others, this factor dominates the odds and the HR increases with additional study. When the test probe is a foil, there is no corresponding memory trace that matches well. Instead, all of the studied items mismatch the test probe with the degree of mismatch growing and the FAR decreasing as the items receive additional study.

The example in Fig. 2 and the above verbal description are supported by simulations. Fig. 3 shows simulated distributions from REM with the parameter values specified above. A weak list contains 60 items each studied once and a strong list contains 60 items each studied five times. Note that the list composition is pure with respect to strength, each list contains all weakly encoded items or all strongly encoded items. The plotted distributions are based on 2500 simulated participants. The distributions are highly skewed, especially the upper tail of the target distribution, so I plot the natural log of the odds for ease of observation. Note that the decision is based on the untransformed value (as shown in Eq. (3)). The figure clearly shows that the odds for a target (and resulting HR) is greater for a strong than weak list and the odds for a foil (and resulting FAR) is lower for a strong than weak list. It is important to note that the strength based mirror effect in REM is not dependent on the parameters used in these simulations. However, it is completely dependent on the assumption that additional study results in the updating of a memory trace rather than storage of a new trace. As long as repetitions result in the accrual of features in a single trace, nearly any set of parameter values will produce a strength based mirror effect for unrelated foils.

Similarity and differentiation

The focus of this paper is an empirical test of differentiation models by considering the effect of testing a foil that is similar to a single studied item. As illustrated in Fig. 3, the subjective match between the stored memory traces and an unrelated foil decreases as the strength of studied items increase. However, what happens when the foils are not randomly chosen, but selected to be similar to a studied item? A foil that is similar to a studied item has more features in common with that item than does an unrelated foil. Not surprisingly, many studies have demonstrated higher FARs for foils that are similar to studied items (e.g., Criss & Shiffrin, 2004b; Roediger & McDermott, 1995; Shiffrin, Huber, & Marinelli, 1995; Zaki & Nosofsky, 2001). I now consider the effect of differentiation on the subjective match to a foil that is similar to a single studied item following study of a weak or a strong list.

Let's return to the example in Fig. 2 and consider test item j_7 which is similar to the studied item corresponding to memory trace i_1. Consistent with the literature, the match between a foil (j_7) and the memory trace to which the foil is similar (i.e., i_1) is greater than the match between the foil and the remaining memory traces (i.e., i_2, i_3, i_4, i_5, and i_6). As before, the match between the test item j_7 and the memory traces stored during study of unrelated items is lower for strongly encoded words than weakly encoded words (i.e., the match to i_2, i_3, i_4, i_5, and i_6 is lower on right side than the left side of the figure). In contrast, the match between the test stimulus j_7 and the similar memory trace (i.e., i_1) increases as the studied item receives additional encoding. What happens when

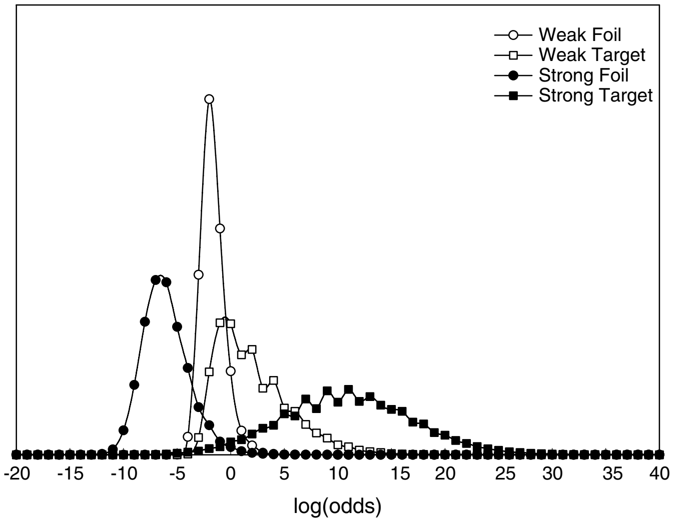

Fig. 3. Simulated distributions of the log odds in REM. Target and foil distributions following study of a pure weak list and a pure strong list are pictured. Differentiation increases the mean of the target distribution and decreases the mean of the foil distribution for strong lists.

these likelihood ratios combine into a single odds value to determine the recognition memory decision? The answer depends on the composition of the study list. Specifically, when the list is pure with respect to target strength, FARs to similar foils will decrease with target strength. But when the list is mixed, with half of the items studied once and half studied many times, the FAR for similar foils will increase with target strength.

Pure lists

Because the match between a foil j_7 and the single trace corresponding to the similar item (i.e., i_1) is only slightly greater than the match to the unrelated memory traces (i_2, i_3, i_4, i_5, and i_6), it does not strictly dominate the odds value (at least not for moderately similar foils). For the example in Fig. 2, the odds for the similar foil j_7 following a weak list is 0.472 $((1.042 + 0.1359 + 0.3039 + 0.4530 + 0.690 + 0.207)/6)$ and the odds for that same foil following study of a strong list is 0.349 $((1.7353 + 0.0001 + 0.0057 + 0.0017 + 0.0005 + 0.0039)/6)$. The final result is counter-intuitive: the odds and resulting FAR for a similar foil is lower following study of a pure strong list compared to study of a pure weak list. This is a reversal of the typical finding that FARs increase when foils are similar to studied items (e.g., Criss & Shiffrin, 2004b; Roediger & McDermott, 1995; Shiffrin et al., 1995; Zaki & Nosofsky, 2001).

Mixed lists

The match between test item j_7 and the similar memory trace (i.e., i_1) grows as that studied item receives additional encoding (i.e., 1.042 vs. 1.7353 in the example). However, when repetition is manipulated between-list, this effect is hidden by the impact of the other list items (i.e., i_2, i_3, i_4, i_5, and i_6). If these remaining traces, stored during study of unrelated items, are equated between the strong and weak conditions then the impact similarity to a studied item should be reflected in the data. That is, the FAR should increase slightly. One possibility is to mix repetitions within a single list so that half of the items are studied once and half are studied multiple times. Returning to the example in Fig. 2, suppose participants study such a mixed list where i_1, i_2, and i_3 are studied once and i_4, i_5, and i_6 are studied many times. The odds for j_7, which is similar to a weak item, is 0.2477 $((1.042 + 0.1359 + 0.3039 + 0.0017 + 0.0005 + 0.0039)/6)$. If i_1, i_2, and i_3 are studied many times and i_4, i_5, and i_6 are studied once, the odds for foil j_7, which is similar to a strong item, is 0.5152 $((1.7353 + 0.0001 + 0.0057 + 0.453 + 0.690 + 0.207)/6)$. For a mixed list, the odds and the resulting FAR will reflect the increasing match between the foil and the similar memory trace.

The above logic for pure and mixed lists applies to the model REM generally; it is not just a cleverly chosen numerical example. Fig. 4 shows simulated distributions of the odds for a foil that is moderately similar to a single studied item. In REM, similar items are constructed by generating two random vectors (i.e., drawing each feature from the geometric distribution with parameter g). One of those vectors is deemed the target and the other the foil. Each feature of the target is independently copied to the foil with some probability determined by the similarity parameter. If the parameter is equal to 1 then the two vectors are identical and if the parameter is equal to 0 then the two vectors are randomly similar (i.e., unrelated). A similarity parameter of.40 was used for these simulations.[2] The top half of the figure contains distributions for pure strong and pure weak lists and the bottom half contains distributions for a mixed list. The pure lists contained 60 items studied once or five times and the mixed list contains 30 items studied once and 30 items presented five times. All other parameter values are those previously declared. Fig. 4 shows that the simulated distributions correspond to the numerical example. For pure lists, the odds value for a foil similar to a strongly encoded item is lower than the odds for a foil similar to a weakly encoded item. The pattern is reversed when repetitions are mixed within a list.

Fig. 5 plots the predicted probability of calling an item “studied” (P(old)) as a function of the number of study presentations and the similarity between the foil and a single studied item. The leftmost panel shows predictions for pure lists. The targets (top line) and unrelated foils (bottom line) demonstrate the strength based mirror effect. The lines in the middle refer to foils that share some similarity with a single studied item. All of the previous discussion applies to moderately similar foils (i.e., those with a similarity parameter between .25 and .45) where the FAR decreases with increasing target repetition. However, foils that are very similar (i.e., a similarity parameter greater than .50) to a studied item mimic the pattern of HRs because the match between these foils and the item to which they are similar is so large that it dominates the odds. I will return to a discussion of highly similar foils in the General Discussion but focus on testing the predictions for moderately similar foils.

[2] In REM unrelated items share features by chance. The expected overlap between randomly generated vectors is determined by the g parameter. For example, with $g = .35$, the average overlap between unrelated vectors is 21.21%. The actual overlap between two similar vectors is a combination of the similarity parameter and the overlap due to the geometric distribution. With the similarity parameter of 0.40, the average overlap between two similar vectors is $(.40)(1.0) + (1 - .40)(.2121) = .527$.

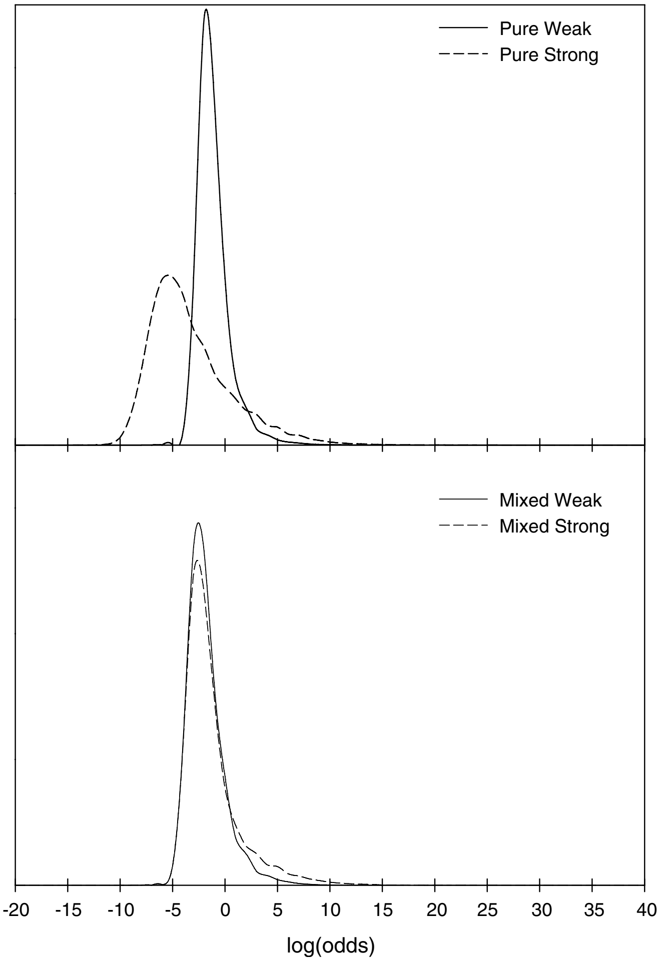

Fig. 4. Simulated distributions of the log odds for a similar foil. The top panel shows the familiarity of a similar foil following study of a pure weak or pure strong list. The bottom panel shows the case where a foil is similar to a strong item from a mixed list or a weak item from a mixed list.

The middle panel of Fig. 5 shows the predicted P(old) for targets, unrelated foils, and foils similar to a single studied item following study of a mixed list where half of the list items receive one study presentation and half receive five study presentations. There is a single FAR for unrelated items because this simulation involves a single study list. For all foils that share features with a studied item, the FAR increases as the corresponding study item receives additional study.

As shown with a numerical example, simulated odds distributions, and predicted response probabilities, REM predicts that the match between a foil and a memory trace stored following study of a similar item increases as the studied item is better encoded. However, the final decision about whether to call a test item "studied" or not is based on the match between the test item and all episodic memory traces (i.e., the odds). The remaining list items, via differentiation, can overturn this effect. If the repetitions are varied between-list, as in the left panel of Fig. 5, the FAR for moderately similar foils should be lower for a strong than a weak list. This prediction is tested in Experiment 1. Equating (approximately) the unrelated study items by mixing repetitions within a list allows the similarity effect to govern the overall match to memory as shown in the middle panel of Fig. 5. Here, the FAR for moderately similar foils are higher for those similar to a strong study item than those similar to a weak study item. Experiment 2 tests this prediction. In Experiment 3, I consider REM's predictions when differentiation is prevented and test the resulting predictions.

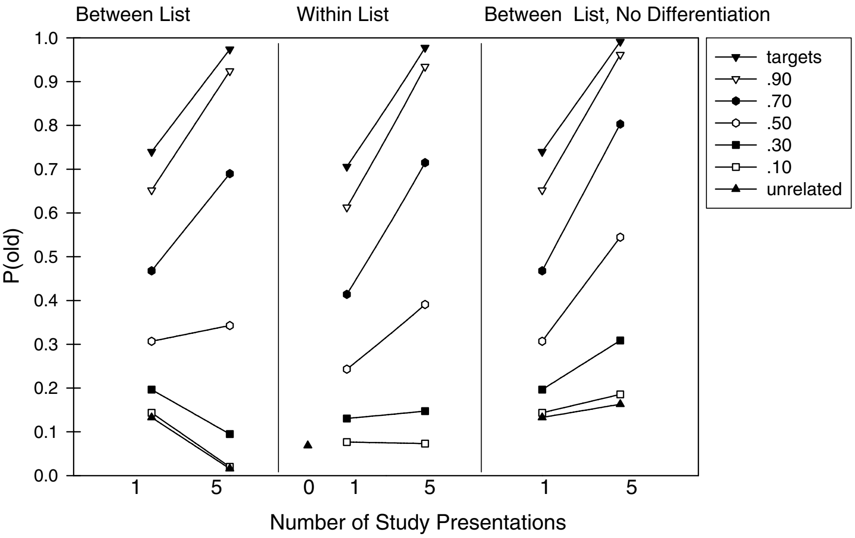

Fig. 5. Predicted probability of responding "old" ((*P*(old)) as a function of target strength and similarity between the test item and a single studied item. The left panel shows the case where strength is manipulated between-list, with each list containing either weak or strong items. The middle panel shows the case where strength is manipulated within-list, such that there is a single study list with half weak and half strong items. The right panel is identical to the left panel with one exception. For the right panel, repeated presentations are stored as a new trace in episodic memory rather than updating a single stored memory trace.

Experiment 1

The first experiment is a test of the counter-intuitive prediction that the FAR for moderately similar foils is lower following study of a strong than a weak list. Moderately similar foils were defined as rhyming words that differed from a target word by a single letter (e.g., boat and coat were one stimulus pair). This choice was based on prior studies using categories of orthographically and phonemically similar words where REM was fit with a similarity parameter of around .30 (Criss & Shiffrin, 2004b). Features in REM are abstract and reflect orthography, phonology, and semantics, among other information. Rhyming foils that differ by a single letter obviously share surface characteristics with their corresponding targets, but share no semantic features (other than arbitrary associations due to the participants' life experiences). Thus, classifying rhyming foils as moderately similar seems a reasonable choice.

Participants

A total of 46 people from the Carnegie Mellon University community received partial course credit or $7 per hour for participating in the experiment.

Stimulus materials

The word pool consisted of 56 pairs of rhyming words that differed by a single letter. The two words were of equal length and every attempt was made to equate the two words on normative word frequency. The average log frequency for the first item of each pair was $M = 8.57$ ($SD = 2.22$) and the average for the second item was $M = 8.44$ ($SD = 2.10$), where the designation of first or second is arbitrary (from the Hyperspace Analog to Language corpus, see Balota et al., 2002; Lund & Burgess, 1996). The difference between the log frequency for each pair of items was computed and the average difference across all pairs was $M = 0.85$ ($SD = 0.56$). Whether the first item served as the target and the second served as the rhyming foil or vice versa was randomly selected for each participant. The unrelated foils were the remaining items in the target pool that were not assigned to the study list for that individual participant.

Design

The study list consisted of 36 words, each studied for 2.5 s followed by a blank screen for 500 ms. One group of participants studied each item once (weak list) and the other group studied each item five times (strong list). In the strong list each, of the 36 items was studied before any one item repeated for five consecutive rounds. Within each round the order of words was randomly assigned. Participants engaged in 45 s of arithmetic between the study and test lists. The total time between the final presentation of each study item and the beginning of the test list was equal on average for both groups. The test list consisted of 54 items, 18 of each of the following: targets, unrelated foils, and rhyming foils. The test was self-paced yes–no recognition with a 500 ms blank screen separating each test trial. Instructions warned that some test items might rhyme with a studied item and others might not but the participant should only respond "studied" to those exact items presented on the list and reject all others.

Results and discussion

An alpha level of .05 was adopted for all statistical tests in this manuscript. An independent samples t-test confirmed that the hit rate was greater for items studied five times than items studied once, $t(41)=2.56$, $p = .014$. A 2×2 mixed design analysis of variance (ANOVA) was conducted with type of foil (unrelated or rhyming) as the within-subject factor and list strength (weak or strong) as the between-subject factor. The FAR for rhyming foils exceeded that for unrelated foils $F(1, 41) = 18.45$, $p < .001$, $MSE = .01$ and the FAR for the weak list was greater than the FAR for the strong list $F(1, 41) = 13.66$, $p = .001$, $SEM = .03$. The interaction approached significance $F(1, 41) = 3.08$, $p = .087$, $SEM = .01$ due to the greater impact of repetition on rhyming foils than unrelated foils. These findings, plotted in Fig. 6, confirm the model predictions, plotted in the left panel of Fig. 5, for moderately similar foils and a between-list manipulation of item strength.

Experiment 2

This experiment is designed to test REM's prediction that following study of a mixed list the FAR to moderately similar foils will increase, reflecting the increasing match between the test item and the memory trace corresponding to the similar studied item.

Participants

A total of 38 people from the Carnegie Mellon University community received partial course credit or $7 per hour for participation in the experiment.

Stimulus materials

The stimuli were identical to those in Experiment 1.

Design

The study list consisted of 36 words, each studied for 2.5 s followed by a blank screen for 500 ms. Half of the items were presented a single time (weak items) and half were presented five times (strong items). Each of the 18 strong items was studied before any one item repeated for each of the first four presentations. Within each round, the order of words was randomly chosen. The fifth presentation of the strong items was randomly intermixed with the presentation of the weak items. Between the study and test lists participants engaged in 45 s of arithmetic. Thus, the total time between the final presentation of the study item and the beginning of the test list was equal on average for both conditions.

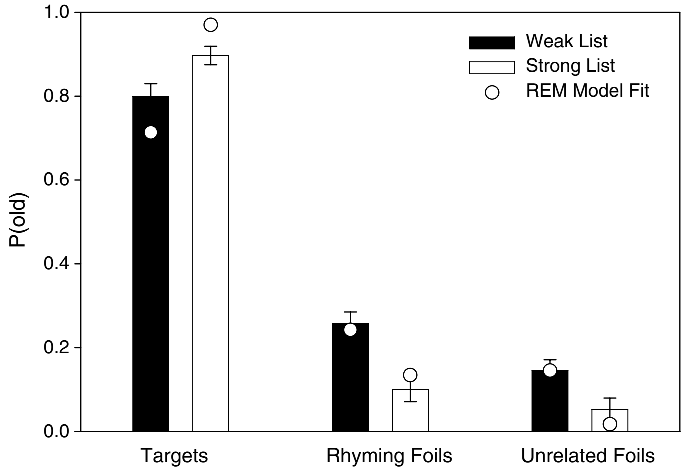

Fig. 6. Data from Experiment 1 where strength is varied between-list. Error bars represent 1 standard error above and 1 standard error below the mean.

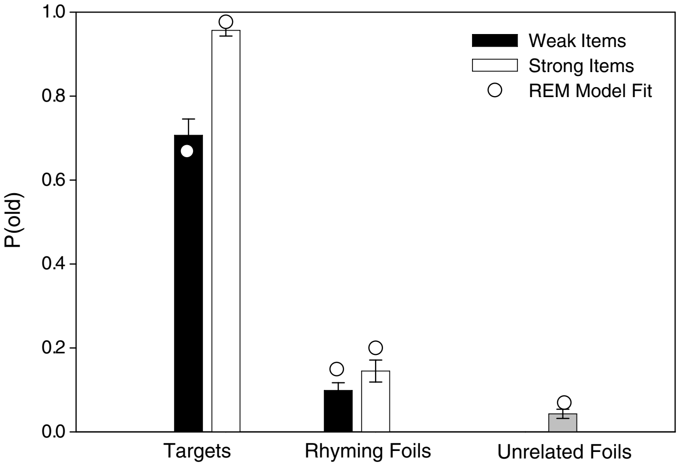

Fig. 7. Data from Experiment 2 where strength is varied within-list. Error bars represent 1 standard error above and 1 standard error below the mean.

The test list consisted of 54 items, 18 of each of the following: targets, unrelated foils, and rhyming foils, half of the targets and half of the rhyming foils came from the weak items and half came from the strong items. The test was self-paced yes–no recognition with a 500 ms blank screen separating each test trial. The test instructions were identical to Experiment 1.

Results and discussion

A paired samples *t*-test confirmed that the hit rate was higher for items studied multiple times $t(35) = 7.88$, $p < .001$. The FAR for foils that rhymed with a strong item was greater than the FAR for foils that rhymed with a weak item $t(35) = 2.26$, $p = .030$. Finally, an ANOVA and post hoc tests confirmed that the FAR for unrelated foils was lower than the FAR to rhyming foils. These findings, shown in Fig. 7, confirm the model predictions, plotted in the middle panel of Fig. 5, for moderately similar foils and a within-list manipulation of item strength.

Experiment 3

All of the reasoning so far refers to the case where perfect differentiation occurs, that is, every repetition of a given item is stored in the same memory trace. What if this is disrupted so that multiple traces of the same word are stored in memory? The right panel of Fig. 5 shows the model predictions when differentiation is prevented. In this simulation each repetition is stored as a new memory trace. For example, if the word coat is presented five times, the result is the storage of five noisy copies of the lexical/semantic representation of coat. The simulations shown in the right and left panels of Fig. 5 are identical except that the left panel incorporates 100% differentiation (each repetition updates a single memory trace) and the right panel incorporates 0% differentiation (each presentation results in storage of an additional memory trace). Both simulations refer to pure lists. There is a striking difference between the model predictions for moderately similar foils: the FAR decreases when differentiation is allowed but increases when differentiation is completely prevented.

What contributes to the increase in FARs when differentiation is prevented? First, the match between a foil and any unrelated memory trace is the same for strong and weak lists because information is not accruing in the same memory trace. Second, preventing differentiation turns an item strength manipulation into a list length manipulation because more traces are stored in memory with each repetition. This is evident in the predictions for unrelated foils. The right panel of Fig. 5 shows that the FAR to unrelated foils increases slightly as the number of repetitions increase. Third, there are now multiple traces that share features with the similar foil (rather than just one) and each provides evidence that the foil item was studied. Returning to the example in Fig. 2, the odds for a similar foil following study of a weak list is 0.472, as illustrated earlier. The odds for a foil similar to an item from a strong list where differentiation was prevented is 0.70 $((5*1.042 + 0.1359 + 0.3039 + 0.453 + 0.69 + 0.207)/10)$, assuming the strong item was studied five times.[3] When repetitions are stored in a new trace rather than accumulated in a single trace,

[3] This is not exactly accurate. The five stored traces of item i_1 would not be identical but would have different likelihood ratios depending on the exact feature values stored in each of the five traces.

the FAR for foils similar to an item from a strong list is higher than the FAR for foils similar to an item from a weak list.

The model predicts yet another interaction: for a between-list manipulation of repetition the FAR should decrease if differentiation is allowed but increase if differentiation is prevented. Experiment 3 tests these predictions. Murnane and Shiffrin (1991) suggested that it was possible to prevent differentiation by embedding words in different sentences. The intuition is that the new sentence context might encourage the participant to think of the repeated word from a different perspective with each presentation and induce storage of a new memory trace. Following Murnane and Shiffrin (1991), I attempt to prevent differentiation by repeating items in different sentences. This results in three conditions: the weak list (each sentence studied once), the strong list (each sentence studied three times), and the different sentences list (three different sentences studied for each target word). Comparison of the weak and strong lists should mimic the pattern of data found in Experiment 1 and in the left panel of Fig. 5. Comparison of the weak and different sentences lists should follow the predictions in the right panel of Fig. 5. Of course, these predictions reflect the most extreme cases where either no differentiation or perfect differentiation occurs. A mixed case where repeated presentations sometimes lead to the updating of a single memory trace and sometimes result in the storage of additional memory traces would yield results somewhere between the two extremes.

Participants

A total of 102 people from the Carnegie Mellon University community received partial course credit or $7 per hour for participation in the experiment.

Stimulus materials

The test stimuli were taken from the same pool as Experiment 1. A fixed set of 36 of the 56 rhyming pairs from Experiment 1 were used as targets and rhyming foils. One item from each of the remaining 20 pairs was randomly chosen and those items constituted the unrelated foil pool. The study stimuli were sentences, ranging in length from three to nine words including one target word per sentence. Three sets of 36 sentences were constructed loosely based on the LAFF sentence database (MIT Speech Communication Group, 2005). Each set contained one sentence including one of the target words. With the exception of pronouns, prepositions, conjunctions, and generic verbs (e.g., can, have, is, etc.) no content words other than the target words were repeated in any of the 108 sentences. For example, the following were sentences for the stimulus pair boat (target) and coat (foil): The boat is easy to support; A small boat is coming tonight; Redeem the ticket on the boat.

Design

The study list consisted of sentences, each studied for 3 s followed by a blank screen for 500 ms. Participants were randomly assigned to one of three groups. One group of participants studied one set of sentences once each (e.g., 36 different sentences and 36 total trials). Another group of participants studied one set of sentences three times each (e.g., 36 different sentences and 108 total trials). Each sentence in the set was presented before any sentence repeated and the order was randomly assigned for each of the three rounds. For both of these groups, the set of sentences was randomly chosen for each participant. The remaining group studied each of the three sets of sentences once each (e.g., 108 different sentences and 108 total trials) and constitute the different sentences list. One complete set was presented before the next set began and the order of sets and order of sentences within a set was randomly assigned for each participant. The first and second groups correspond to the weak and strong lists just as in Experiment 1. The third group is my attempt to prevent differentiation and is the critical comparison group. Between the study and test lists participants engaged in 45 sec of arithmetic. The total time between the final presentation of each study item and the beginning of the test list was equal on average for all groups. The test list consisted of 36 words, 18 targets, 9 rhyming foils, and 9 unrelated foils. The test was self-paced yes–no recognition with a 500 ms blank screen separating each test trial. As before, instructions warned that some foils might rhyme with a studied word.

Results and discussion

Fig. 8 plots the P(old) for the three types of test items and the three study conditions. Separate ANOVAs were conducted on the targets, unrelated foils, and rhyming foils. Each test confirmed an overall difference in P(old) between the weak, strong, and different sentences conditions for targets ($F(2, 99) = 8.99$, $p < .0001$, $MSE = .014$), unrelated foils ($F(2, 99) = 3.48$, $p = .035$, $MSE = .015$), and rhyming foils ($F(2, 99) = 3.59$, $p = .031$, $MSE = .027$). All remaining statements are based on significant Bonferroni-adjusted post hoc tests.

Embedding words in sentences during study produced a typical strength based mirror effect: the HR was higher and the FAR to unrelated foils was lower for the strong than weak condition. In addition, the FAR for similar foils was lower for the strong than weak

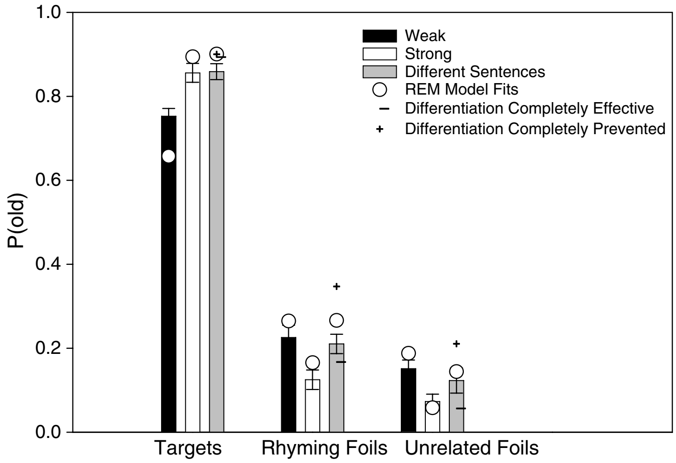

Fig. 8. Data from Experiment 3 where strength is varied between-list and the study list contains sentences. Error bars represent 1 standard error above and 1 standard error below the mean.

list. This set of data nicely replicated Experiment 1 (the single item version) and again confirmed predictions of REM.

Next compare the weak list and the different sentences list. Repeating items in the context of different sentences improved memory for those items as evidenced by the higher HR for the different sentences than the weak condition. However, the FAR for unrelated foils did not differ between the two conditions and neither did the FAR for rhyming foils. The latter two findings are not consistent with the model predictions when differentiation is completely prevented. REM predicts that FARs to unrelated and rhyming foils should increase with target repetition, if each item presentation is stored as a new trace. If each presentation of a repeated item is stored in the same memory trace, then the model predicts that the FARs should decrease with target repetition. The data fall in the middle, the FARs are not rising or falling, suggesting imperfect differentiation.

Why was the experimental manipulation not successful at completely preventing differentiation? For one, the sentences were not explicitly constructed to highlight different meanings of the target words. In addition, the set of 108 sentences presented during the different sentences condition contained more unique words than the other conditions, possibly adding noise. Other strategies to prevent differentiation might include presenting individual items with different encoding tasks where each task is chosen to illustrate different properties, presenting words with different definitions, or presenting words in different temporal contexts. Further research is needed to flesh out the conditions leading to different amounts of differentiation and the effects of various levels of differentiation on subsequent memory performance.

Evaluation of the REM model fits for Experiments 1–3

The open circles in Figs. 6–8 are REM's fit to the data. Best fits were not obtained due to the complexity of the parameter space. Rather, the data were fit by hand. The values of the similarity parameter and u were allowed to vary. All other parameters were held constant at the values listed earlier (i.e., $c = .70$, $g = .35$, $M = 20$, $criterion = 1$); these values are taken from previous implementations of REM and should not be considered free parameters for the present purposes. The same similarity parameter (.33) was used for all experiments. A value of $u = 0.28$ was obtained for Experiments 1 and 2. In Experiment 3, sentences were studied and it seemed reasonable to allow the probability of storing a feature to differ from Experiments 1 and 2, in lieu of implementing limited capacity, thus $u = .22$. An additional parameter governed the probability that a new trace was stored on each successive repetition of an item. To strengthen the case that differentiation was partially prevented, Fig. 8 displays the predictions of the model for the different sentences condition when differentiation is completely prevented (parameter value 1) and when differentiation is completely successful (parameter value 0). This parameter was allowed to vary and reasonable fits resulted from a value of 0.30, as also shown in Fig. 8. If a new trace was not stored, then the memory trace stored on the first presentation was updated (see Malmberg, Holden, & Shiffrin, 2004 & Raaijmakers, 2003 for similar modeling strategies). For simplicity, simulations of Experiment 3 only stored target items in memory, ignoring the remaining words in the sentences (and ignoring the potential for emergent associative features e.g., Criss & Shiffrin, 2005, 2004a; Murnane & Shiffrin, 1991). Each simulated participant had the same parameter values and the model was fit to group data rather than individual participants. More suitable parameter values might be obtained if the model were fit to individual participants (i.e., Estes & Maddox, 2005). Despite the limited parameter search, fitting group data, and the use of few free parameters, the model fits with relatively high accuracy.

Observation of Fig. 5 may lead the reader to mistakenly believe that the model can predict any pattern of FARs if the appropriate similarity parameter is chosen. Indeed, the similarity parameter determines whether the FAR increases, decreases, or remains constant with target repetitions. However, the similarity parameter also determines the overall level of FAR for similar foils, with higher levels of similarity leading to more false alarms. Both the overall FAR rate and the pattern of FAR as a function of target repetition are jointly determined by the similarity parameter. For example, the model cannot predict, a very high FAR (like those for the similarity parameter greater than .50) that decreases with target repetition. As for the specific value of the similarity parameter, it is important to note that similarity is not an attribute that can be objectively measured and imported into models. Rather, similarity depends on the context including such factors as the set of items being considered and the bias of the rater (e.g., Goldstone, 1994; Tversky, 1977). Until more sophisticated measures of similarity are constructed and/or REM incorporates concrete measurable features (which requires an accurate way to identify and measure the features belonging to a concept, see Steyvers, 2000 for one attempt), one cannot verify that any stimuli correspond to an absolute level of similarity. Nevertheless, the fact that the similarity parameter jointly determines the overall level of FAR for similar foils and the pattern across target repetition provides a reasonable degree of confidence that the similarity parameter is not arbitrary. Indeed the value of the similarity parameter was held constant for fits to the current set of experimental data. These fits produced reversals in the FAR pattern with list composition, exactly as predicted by the model a priori.

General discussion

The strength based mirror effect naturally falls out of the REM due to the accrual of information in a single memory trace during encoding. The higher HR for strong than weak items is the result of a greater match between a test item and its corresponding memory trace. The lower FAR following a strong than a weak list is due to the greater mismatch between a test item and memory traces stored during the study of other items. REM makes additional predictions about the FAR for foils that are moderately similar to a studied item. If item strength is manipulated between-list, the FAR to foils that are similar to a strong item will be less than the FAR for foils that are similar to a weak item. However, if item strength is manipulated within-list the reverse is predicted. These predictions were confirmed in Experiments 1 and 2, respectively. In the third experiment, *S*s studied target words embedded in sentences. When sentences were presented once or three times each, in a between-list design, the results mimicked Experiment 1. When the same word was presented in three different sentences, the HR increased but the FARs to unrelated foils and to similar foils did not change relative to the once presented condition. According to the model, this is best explained by assuming imperfect differentiation meaning that a repetition is sometimes stored as a new episodic trace and other times an existing episodic trace is updated.

Implications for models

Many other accounts of the strength based mirror effect assume that participants adopt a more strict criterion for saying "old" following a strong than a weak study list. The match to memory for the two sets of foils is not different, rather *S*s are simply more or less biased to say "old." One might ask whether the new data refute this criterion placement account. Unfortunately, there is no simple answer. All models, including REM, require a criterion parameter for responding old or new and the criterion shift assumption can be adopted in any modeling framework. As mentioned earlier, it has been adopted by both single and dual process models, attesting to its universal applicability. Thus, it is only fair evaluate individual models.

Signal detection theory

Fig. 1 depicts SDT applied to memory. This is the type of model often adopted when discussing the strength based mirror effect (e.g., Stretch & Wixted, 1998; Verde & Rotello, in press). This framework assumes a single distribution for unrelated foils, one that does not change as a function of the strength of the studied items. There is no doubt that the current set of data could be fit by SDT, with the following results.[4] First, similar foils form a third distribution, with a mean greater than the unrelated foils. Second, the mean of the similar foil distribution increases as the study items increase in strength but the mean of the unrelated foil distribution remains unchanged (as in Fig. 1). A lenient criterion for the weak relative to the strong list would then account for the data from both unrelated and similar foils as in Experiments 1 and 2. Turning to Experiment

[4] Thanks to John Dunn for confirming this.

3, the target distribution for the different sentences condition has a mean located between the weak and strong conditions. A slightly more lenient criterion is used for the weak than the different sentences condition. The criterion for the strong condition is considerably more strict than the other two conditions.

Despite the ability of this model to fit the data, conceptual problems exist. Some readers might find it peculiar to assume the similar foil distribution is affected by the strength of the studied items but the unrelated foil distribution is not. In fact, one application of SDT to foils belonging to the same category as a strong or a weak studied item explicitly assumed a single foil distribution regardless of target strength (Morrell et al., 2002). Further, the above discussion implies that the criterion is set based on the perceived memorability of the study list (e.g., Benjamin & Bawa, 2004; Brown, Lewis, & Monk, 1977; Hirshman, 1995), a proposal that has been rejected by some (e.g., Verde & Rotello, in press). The fundamental issue here is that signal detection theory is a theory about how a signal is detected in the presence of noise. It is a descriptive model that can fit existing data and has proven successful for a wide variety of tasks for many years (Green & Swets, 1966; Macmillan & Creelman, 1991; Wickens, 2001). However, SDT is not a theory about the encoding and retrieval processes that underlie human memory and thus has no mechanistic account of memory and no theoretical basis for the resulting fits. The purpose of this manuscript is not to rule out the differential placement of the criterion for different experimental conditions (see Brown & Steyvers, 2005; Brown, Steyvers, & Hemmer, in press; Heit, Brockdorff, & Lamberts, 2002 for relevant discussion on the process of criterion selection). Rather, the goal is simply to present an opposing account and test predictions generated by that account. Additional studies are required to determine the extent to which differentiation or criterion shifts (or neither or both) underlie the strength based mirror effect and other related phenomenon.

Fully informed likelihood models

Often the criterion shift account is contrasted with fully informed likelihood models. Fully informed likelihood models assume that the memory system takes into account the statistics of the distribution of familiarity values associated with both the old and the new stimuli for each test condition of the experiment. The likelihood of obtaining the familiarity value associated with a given test item under each of the two relevant distributions is computed and used as the decision variable. For the strength based mirror effect, fully informed likelihood models produce four separate distributions for weak and strong targets and foils (similar in appearance to Fig. 3). These distributions arise because the decision process knows that a weak item is weak and a strong item is strong and takes that into account when calculating the likelihood ratio. However, it is important to be aware that the REM and SLiM are not members of the class of fully informed likelihood models, contrary to typical citation patterns in the literature. As extensively discussed, the simulated distributions of REM pictured in Fig. 3 arise are due to the accrual of repetitions in a single memory trace. REM and SLiM do make use of likelihood ratios, owing to the confusion. However, these likelihood ratios are *subjective* likelihoods based on the degree to which a test item matches the contents of memory (see Eqs. (2) and (3)). Unlike fully informed likelihood models, no knowledge of the parameters specific to the experimental conditions are assumed when computing likelihood ratios in either REM or SLiM (see Criss & McClelland, in press for further discussion).

Attention Likelihood Theory (ALT) is a fully informed likelihood model and has been extensively applied to mirror effects arising from manipulating study time, normative word frequency, part of speech, among other variables (e.g., Glanzer & Adams, 1985, 1990). ALT is a local access model in which the corresponding item in memory, but no other stored memory traces, participates in the decision. ALT cannot account for changes in behavior due to the composition of the study list, as found for moderately similar foils. The underlying distributions for similar foils and targets do not differ as a function of whether item strength is varied between or within-list, thus ALT should predict the same pattern of data in both cases, in contrast to the data.

Other models

SAC (Cary & Reder, 2003; Reder et al., 2000) has adopted the criterion shift assumption to account for strength based mirror effect. SAC assumes that words are represented by individual nodes rather than features and has not yet been extended to account for similarity between targets and foils. Further, SAC assumes that false alarms are due to the pre-experimental familiarity of the test item and pre-experimental familiarity, by definition, is not affected by encoding conditions. It is not clear how SAC could account for the data from the moderately similar foils without additional assumptions.

There are many other models of episodic memory that deserve mention. Only the SLiM model incorporates differentiation and makes predictions that are qualitatively similar to those reported here.[5] Dennis and Humphreys (2001) proposed a context-noise model where the recognition decision is based on the match

[5] Simulated distributions and response probabilities for SLiM are available upon request.

between the reinstated study context and the prior contexts stored with the test item. In this model, no items from the study list participate in the decision and the model predicts no effect of list length, list strength, or list composition and cannot account for the current data without additional assumptions. Murdock (2003) accounts for both the word frequency mirror effect and the spacing effect by assuming that the less familiar the item, the greater the benefit of study. In the experiments used here, foils from the strong and weak conditions come from the same pool and are equally familiar. It is not clear how the model could account for the strength based mirror effect or the consequences for similar foils. Other global matching models (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988; Murdock, 1997; Murdock et al., 2001; Nosofsky, 1988) do not incorporate the assumption of differentiation – they do not assume that a strengthened item becomes less similar to other unrelated items. Instead they assume that a new trace is stored with each repetition of a studied item, making a strength manipulation similar to a length manipulation. As a result, they make predictions qualitatively similar to those shown in the right column of Fig. 5, regardless of whether study repetitions are manipulated between or within-list. This is not to say that any of the models discussed above are ruled out by the current set of data. Rather they too could adopt an ad hoc criterion shift at the expense of a less parsimonious account. Only REM and SLiM predict the current set of data as a natural consequence of the encoding process.

Limitations of REM

REM has been successfully applied to a range of different memory tasks (e.g., judgments of frequency, associative recognition, list discrimination, cued recall, perceptual identification, and lexical decision, see Malmberg et al., 2004; Criss & Shiffrin, 2004a, 2005; Diller, Nobel, & Shiffrin, 2001; Schooler, Shiffrin, & Raaijmakers, 2001; Wagenmakers et al., 2004, respectively). However, REM is not without problems. REM has not been extended to important findings such as serial position functions, forgetting curves, or the spacing effect. More problematic is that REM suffers a bit of an identity crisis when applied to associative recognition (AR). The original paper proposed that a pair was represented by two concatenated vectors and AR was accomplished by comparing the double-long vector to memory in the same way that single item recognition was accomplished. Later papers adopted a cued recall process (Diller et al., 2001) or a recall-to-accept mechanism (Xu & Malmberg, in press). My own work establishes the necessity of emergent associative features that are unrelated to item features and implements this assumption in REM (Criss, 2005; Criss & Shiffrin, 2004b, 2005). The representational constraints and the retrieval process are not necessarily in conflict. However, we have yet to agree on the exact conditions requiring emergent associative features and the exact conditions requiring recall in addition to familiarity.

Related empirical findings

Two studies manipulating item strength and similarity deserve further mention. First consider Morrell et al. (2002). They conducted experiments somewhat similar to Experiment 2 but found no difference in FAR for similar foils as a function of item strength (but see Benjamin & Bawa, 2004; Brown et al., in press). For example, in Experiment 2 of Morrell et al, participants studied 20 items from category A 5 times each and 20 items from category B 1 time each (or vice versa so that category B items were repeated and category A items were not). The similar foils were unstudied items from each category. The relevant comparison is the FAR to items in a given category when that category was strong versus weak and Morrell et al. found no difference between these FARs.[6]

Shiffrin et al. (1995) also presented categories of related items at study with each item from a given category studied once or many times. They also manipulated the number of exemplars per category, varying from 2 to 9 on a list with a total of 145 unique words. This ratio is somewhat closer to the ratio used here, relative to the Morrell et al. (2002) study. In addition, Shiffrin et al. used less obvious categories in an effort to prevent participants from adopting encoding strategies based on the structure of the study list (e.g., generating unstudied category members during study). In seven of eight cases (e.g., comparing strong vs. weak for a fixed category length) the FAR to foils from the strong category were numerically higher than the FAR to foils from the weak category. The average increase in FAR over all eight conditions was 0.0385. The many conditions in their study probably contributed to a lack of statistical power to detect this small difference. Shiffrin et al. and Morrell et al. concluded that the strength of similar studied items has little effect on the overall familiarity of a test item. I offer the alternative conclusion that in fact the strength of similar studied items does influence the overall familiarity of a foil item. However, that increase in familiarity can be dominated by the match of the other unrelated list items resulting in various interactions and reversals as seen in the current set of data.

[6] Morrell et al. (2002) claimed that likelihood models predict a decrease in FAR as a function of strength. They erroneously included citations to REM and SLiM whereas this paper clearly shows that the differentiation models predict an increase in the FAR.

Fig. 5 shows that the predicted pattern of data for highly similar foils (i.e., similarity parameter greater than .50) is different from the predicted pattern for the moderately similar foils used in these experiments. Plurality reversed foils seem to fall in the category of highly similar foils. FARs to such foils typically rise for the initial target presentation then remain relatively constant or decrease slightly as the number of target presentations approaches double digits (Hintzman & Curran, 1995; Hintzman, Curran, & Oppy, 1992). The differentiation models provide different accounts of this pattern: REM adopted a recall-to-reject mechanism but SLiM maintained the same familiarity-based process used for unrelated foils (Malmberg et al., 2004 & McClelland & Chappell, 1998, respectively). All of these studies manipulated target repetition within-list. An untested prediction of the models is what happens when target repetition is manipulated between-list. REM and SLiM predict that the FAR for very similar foils should mimic the HR and rise with target presentation, if responses are based on familiarity (at least for the relatively low number of repetitions used in the current simulations).

Summary

Differentiation models offer an alternative account of the strength based mirror effect, one based on the contents of episodic memory rather than on criterion placement. Differentiation is simply the accumulation of information in a single memory trace with additional repetitions, resulting in a reduction in the match between a strong target and an unrelated foil compared to a weak target and an unrelated foil. The similarity between a target and similar foil rises as a function of the strength of the studied item. However, the overall decision is based on the total match to memory which includes the match between the foil and the other unrelated studied items. Thus the FAR to similar foils might rise or fall depending on the balance between the strength of the many unrelated studied items and the strength of the single related studied item.

When repetitions are varied between-list, the FAR to unrelated foils and to similar foils decreases (Experiment 1) but when repetitions are mixed within a list, the FAR to similar foils increases with the strength of the similar item (Experiment 2), just as predicted. If differentiation is completely eliminated, the FARs to similar foils should rise even when repetitions are varied between-list. Embedding repetitions of target words in different sentences was partially successful in preventing differentiation (Experiment 3). REM a priori predicted the complex pattern of data reported here and produced good quantitative fits. A criterion shift account cannot be ruled out but requires additional and perhaps unsatisfactory assumptions.

References

Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., & Simpson, G. B. (2002). The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords. Available from http://elexicon.wustl.edu/, Washington University.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51*, 159–172.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology, 29*, 461–473.

Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(4), 587–599.

Brown, S., Steyvers, M., & Hemmer, P. (in press). Modeling experimentally induced strategy shifts. *Psychological Science*.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Memory and Cognition, 49*, 231–248.

Criss, A. H. (2005). The representation of single items and associations in episodic memory. Doctoral dissertation, Indiana University, 2004. *Dissertation Abstracts International-B, 65*(12), 6882.

Criss, A. H., & McClelland, J. L. (in press – this issue). Differentiating the differentiation models: A Comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language*, doi:10.1016/j.jml.2006.06.003.

Criss, A. H., & Shiffrin, R. M. (2004a). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory and Cognition, 32*(8), 1284–1297.

Criss, A. H., & Shiffrin, R. M. (2004b). Context noise and item noise jointly determine recognition memory: A comment on Dennis & Humphreys (2001). *Psychological Review, 111*(3), 800–807.

Criss, A. H., & Shiffrin, R. M. (2005). List discrimination and representation in associative recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1199–1212.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review, 109*, 710–721.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452–477.

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 414–443.

Dobbins, I., & Kroll, N. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1186–1198.

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

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin and Review, 12*, 403–408.

Gibson, E. (1940). A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review, 47*, 196–229.

Gibson, E. (1969). *Principles of Perceptual Learning and Development*. New York: Appleton-Century-Crofts.

Gibson, J., & Gibson, E. (1955). Perceptual learning – differentiation or enrichment? *Psychological Review, 62*, 32–41.

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

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition, 13*(1), 8–20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 5–16.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*, 125–157.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, UK: Wiley.

Heit, E., Brockdorff, N., & Lamberts, K. (2002). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin and Review, 10*(3), 718–723.

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

Hintzman, D. L., & Curran, T. (1995). When encoding fails: Instructions, feedback, and registration without learning. *Memory and Cognition, 23*(2), 213–226.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(4), 667–680.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list strength paradigm. *Journal of Experimental Psychology: Learning Memory and Cognition, 21*, 302–331.

Hockley, W. E. (1994). Reflections of the mirror effect for item and associative recognition. *Memory and Cognition, 22*, 713–722.

Hockley, W. E., & Niewiadomski, M. W. (in press). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory and Cognition*.

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology, 33*, 36–67.

Kim, K., & Glanzer, M. (1993). Speed versus accuracy instructions, study time, and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 638–665.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers, 28*, 203–208.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

Malmberg, K. J., Holden, J., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old–new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 319–333.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 734–760.

Metcalfe-Eich, J. M. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review, 92*, 1–38.

MIT Speech Communication Group. (2005). Lexical access from features sentences. Available from http://hdl.handle.net/1721.1/27283.

Morrell, H., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal detection-based models of recognition memory. *Journal of Experimental Psychology-Learning, Memory, and Cognition, 28*, 1095–1110.

Murnane, K., & Shiffrin, R. M. (1991). Word repetitions in sentence recognition. *Memory and Cognition, 19*(2), 119–130.

Murdock, B. B. (1997). Context and mediators: A theory of distributed associative memory (TODAM2). *Psychological Review, 104*(4), 839–862.

Murdock, B. B. (2003). The mirror effect and the spacing effect. *Psychonomic Bulletin and Review, 10*, 570–588.

Murdock, B., Smith, D., & Bai, J. (2001). Judgments of frequency and recency in a distributed memory model. *Journal of Mathematical Psychology, 45*, 564–602.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 54–56.

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

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 163–178.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 294–320.

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*, 803–881.

Schooler, L., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A model for implicit effects in perceptual identification. *Psychological Review, 108*, 257–272.

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

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 179–195.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – Retrieving effectively from memory. *Psychonomic Bulletin and Review, 4*, 145–166.

Steyvers, M. (2000) Modeling semantic and orthographic similarity effects on memory for individual words. Dissertation, Psychology Department, Indiana University.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1397–1410.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Verde, M., & Rotello, C. (in press). Memory strength and the decision process in recognition memory. *Memory and Cognition.*

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory and Cognition, 20*(3), 291–302.

Wagenmakers, E. J. M., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology, 48*, 332–367.

Wickens, T. D. (2001). *Elementary signal detection theory*. Oxford University Press.

Xu, J., & Malmberg, K. J. (in press). Modeling the effects of verbal- and non-verbal pair strength on associative recognition. *Memory and Cognition.*

Zaki, S. R., & Nosofsky, R. M. (2001). Exemplar accounts of blending and distinctiveness effects in perceptual old new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(4), 1022–1041.