Cognitive Psychology 59 (2009) 297–319

Contents lists available at ScienceDirect

Cognitive Psychology

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

The distribution of subjective memory strength: List strength and response bias

Amy H. Criss *

Department of Psychology, Syracuse University, Syracuse, NY 13244, United States

ARTICLE INFO

Article history:
Accepted 19 July 2009
Available online 17 September 2009

Keywords:
Episodic memory
Mathematical models
Mirror effects
Encoding
Recognition memory
Signal detection theory
Strength

ABSTRACT

Models of recognition memory assume that memory decisions are based partially on the subjective strength of the test item. Models agree that the subjective strength of targets increases with additional time for encoding however the origin of the subjective strength of foils remains disputed. Under the fixed strength assumption the distribution of memory strength for foils is invariant across experimental manipulations of encoding. For example, the subjective strength of foils may depend solely on the pre-experimental history of the item, thus encoding manipulations have no impact. In contrast, under the differentiation assumption the subjective strength of foils depends on the nature of the traces stored in episodic memory. If those traces are well encoded, the subjective strength of foils will be lower than the case where noisy traces are stored (e.g., when targets received minimal encoding). The fixed strength and differentiation accounts are tested by measuring direct ratings of memory strength. In Experiments 1 and 2, item strength is varied via repetition and in Experiment 3 response bias is varied via the relative proportion of targets on the test list. For all experiments empirical distributions of memory strength were obtained and compared to the distributions predicted by the two accounts. The differentiation assumption provides the most parsimonious account of the data.

1. Introduction

Recognition memory experiments require participants to endorse target items that were studied on an earlier list and reject foil items that were not studied. Manipulations that improve recognition

* Fax: +1 315 443 4085.
E-mail addresses: amy.criss@gmail.com, acriss@syr.edu.

0010-0285/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2009.07.003

memory accuracy often do so via a mirror effect: the simultaneous increase in the probability of correctly endorsing a target item (hit rate, HR) and decrease in the probability of erroneously endorsing a foil item (false alarm rate, FAR, e.g., Glanzer & Adams, 1990). One such example is the strength based mirror effect (SBME). The SBME is the finding that HRs increase and FARs decrease when accuracy is improved by increasing the strength of the studied items, typically via increased encoding time or repeated presentations of the study items (e.g., Stretch & Wixted, 1998). For example a strong list might contain targets, each studied five times and a weak list might contain targets, each studied once. HRs are higher and FARs are lower following the strong list than following the weak list. That participants are better able to recognize a strong target which received multiple opportunities for more encoding than a weak target is not a surprise. Of greater theoretical interest is why the FAR differs between the strong and weak lists. Two explanations will be considered here – the criterion shift assumption and the differentiation assumption.

1.1. The criterion shift assumption

One class of models adopts the *criterion shift assumption* wherein participants adopt a more stringent criterion following a strongly encoded list than a weakly encoded list. This assumption is adopted by models from two classes: global matching models where variability increases with strength resulting in the prediction of a higher FAR following a strong than weak list (see Shiffrin, Ratcliff, & Clark, 1990) and those that assume the subjective strength of unrelated foils is not affected by the encoding conditions, called fixed strength models and illustrated in the top panel of Fig. 1 (e.g., Cary & Reder, 2003; Stretch & Wixted, 1998; Verde & Rotello, 2007; a subset of the dual process models discussed in Yonelinas (2002)).[1]

The *fixed strength assumption* was first adopted as a convenient means of constraining signal detection theory (e.g., Lockhart & Murdock, 1970; Parks, 1966; Wickelgren & Norman, 1966). Signal detection theory (SDT) as applied to recognition memory assumes that the subjective strength of targets and foils are two overlapping normal distributions as illustrated in the top panel of Fig. 1.[2] Participants select some criterion (the vertical lines in Fig. 1) and any item evoking a subjective response greater than the criterion is endorsed, other items are rejected. Thus in the basic yes–no recognition memory paradigm, SDT requires five parameters (mean and standard deviation for each distribution and location of the criterion) for two data points (HR and FAR). Fixing the foil distribution as the standard normal (mean = 0 and standard deviation = 1) eliminated two free parameters. When multiple conditions were compared, the foil distribution across all conditions was constrained in the same way thus the fixed strength assumption was born. Though this assumption might be described as a simple convenience, it has been justified by attributing the subjective memory strength evoked by foils during a recognition memory test to the pre-experimental familiarity of the foil item. Thus manipulations of the history of the foil (e.g., normative word frequency) but not encoding conditions (e.g., repetition during the study list) were suggested to affect the subjective memory strength of foils (e.g., Lockhart & Murdock, 1970). The fixed strength assumption has persisted over decades and has moved beyond SDT models. Indeed, a number of current episodic memory models including both single and dual process models have embraced the fixed strength assumption.

Stretch and Wixted (1998) adopted the fixed strength assumption within a SDT model of recognition memory and applied it directly to the SBME (see also Hirshman, 1995). They proposed that additional encoding of items on the strong list increases the mean of the target distribution, hence the increase in the HR while the foil distribution does not change as a function of encoding conditions (see top panel of Fig. 1). However, participants adopt a stricter criterion for the strong list resulting in a lower FAR for that list despite the fixed foil distribution. The global matching models referred

[1] The term subjective strength refers to the scalar output generated from comparing the test item (and context) to the contents of episodic memory. In the literature it is also called subjective response, memory strength, global match, familiarity, activation, etc.

[2] An unequal variance model is illustrated because that is the model advocated by Mickes, Wixted, and Wais (2007) who introduced the direct ratings paradigm and by Stretch and Wixted (1998) who first applied SDT to the SBME. Further, in the top panel of Fig. 1 the variance of the strong and weak target distributions are equal. These choices were made for ease of illustration and do not represent a prediction or an assumption by a particular model or by the author.

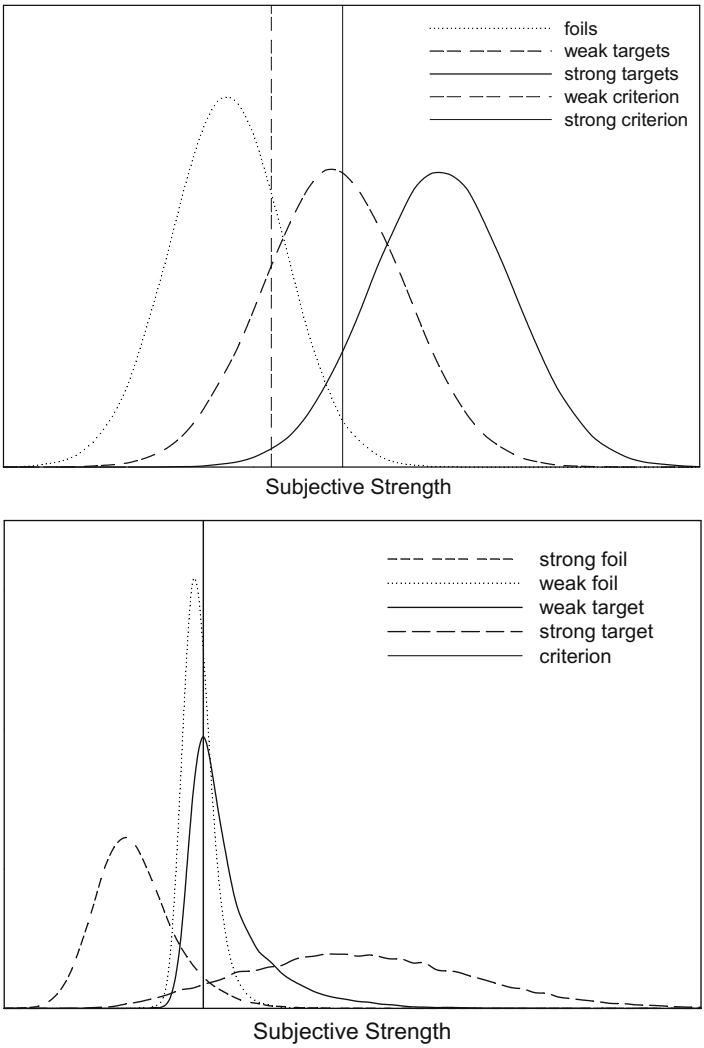

Fig. 1. The top panel shows an example of a signal detection theory account of the strength based mirror effect. The mean of the target distribution is greater for a strong than a weak list. The mean of the foil distribution is constant for strong and weak lists (the fixed strength assumption) but the criterion (the vertical line) changes between the two lists producing the strength based mirror effect. The bottom panel shows simulated distributions of REM (Shiffrin & Steyvers, 1997) for targets and foils following study of a weak and a strong list. Differentiation increases the mean of the target distribution and decreases the mean of the foil distribution for strong lists relative to weak lists. For clarity of illustration, the plotted distributions are the log of the decision variable. In the actual model, the decision is based on the untransformed value. Note that the two models are not on the same scale, so the absolute position on the *x*-axis cannot be compared in across the top and bottom panels.

to above predict an increase in the FAR following a strong compared to a weak list and must also adopt a criterion change to account for the strength based mirror effect. Finally some dual process models especially those attributing FARs to pre-experimental familiarity (or semantic memory) must adopt the criterion shift to account for an empirical SBME. Thus the criterion change explanation for the SBME was widely adopted with little question or controversy and was consistent with nearly all models at the time. Numerous studies have investigated the SBME by manipulating the type of stimuli, the type of memory test, or the retention interval (e.g., Cary & Reder, 2003; Hockley & Niewiadomski,

2007; Kim & Glanzer, 1993; Singer, Gagnon, & Richards, 2002; Singer & Wixted, 2006). Studies using the SBME paradigm or other manipulations have addressed how the criterion is set (Brown & Steyvers, 2005), whether it depends on the perceived difficulty of the study list or the experienced difficulty of the test items (e.g., Benjamin & Bawa, 2004; Brown, Lewis, & Monk, 1977; Hirshman, 1995; Verde & Rotello, 2007), and whether or not the criterion changes within a single list. However, the notion that the SBME resulted from a criterion shift was not challenged until recently.

1.2. The differentiation assumption

Criss (2006) showed that models adopting differentiation (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) provide a very different account for the strength based mirror effect. These models differ in the exact details of implementation and produce competing predictions for some paradigms (see Criss & McClelland, 2006). However, they share two critical features that underlie the SBME. The first is that repetitions of an item within the same context results in the updating of a single memory trace for that item. Updating a single memory trace results in a more accurate representation of the target item with each encoding opportunity. The more accurate the memory trace, the better it matches when the target is presented during a recognition memory test. The second assumption is that the more accurate a given memory trace, the less similar it is to unrelated items. For a foil probe, all stored memory traces are unrelated (that is they only match features by chance) and mismatch features of the foil probe to some extent. The better encoded the memory traces, the greater the mismatch to a foil probe and the lower the FAR. This logic still applies when the test probe is a target because all but one memory trace is unrelated. For the single memory trace that *is* related to the target probe (because it was stored when the target was studied) the match tends to be large and that match grows as the target items receives additional encoding. The matching rule is multiplicative and therefore the match between a target probe and the single memory trace that it matches outweighs the mismatch between the target probe and the remaining unrelated memory traces, thus the HR increases as target items receive additional encoding. This competition is more obvious when the test probe is similar but not identical to a target such as when a foil is similar to a studied item. Criss (2006) showed that the differentiation models a priori predict a reversal in the pattern of FARs for such foils when the composition of the study list changed from pure strength to mixed strength and confirmed these predictions in a series of experiments.

These two differentiation models do not assume that the distributions of subjective memory strength are Gaussian (or any other form). Instead, the distributions of subjective memory strength are simulated based on the encoding and retrieval processes of the model (see the section titled "The REM Model" for an example). The bottom panel of Fig. 1 illustrates the predicted distributions of subjective memory strength for targets and foils following a strong and a weak list. The distribution of subjective memory strength increases for targets and simultaneously decreases for foils following a strong list compared to a weak list. This prediction of the differentiation models follows directly from the encoding assumptions (i.e., memory traces are updated with repetition) and retrieval assumptions (i.e., the decision rule where mismatching features decrease the subjective memory strength). In fact, Criss (2006) demonstrated that disrupting the encoding assumptions by storing a new memory trace with each repetition rather than updating a single memory trace for each item disrupts the SBME predictions. Under these conditions, the models predict that the FAR should be greater for a strong than a weak list (i.e., the opposite of the pattern predicted by the model when memory traces are updated with repetition during encoding). Thus the mirror pattern for the SBME is a direct consequence of the encoding mechanism assumed in differentiation models and is not simply a convenient assumption. The same mechanisms that are responsible for the SBME also underlie predictions for the null list strength effect, among other findings, and are thus critical components of the model. Note that differentiation models, like all models, require the participant to set a criterion for endorsing a test item as one from the list. However, the placement of the criterion carries no explanatory power for the strength based mirror effect.

1.3. Likelihood models

The astute reader will note that SDT is a measurement model that is able to adopt either a fixed strength/criterion shift assumption or a differentiation/fixed criterion assumption and these

assumptions cannot be discriminated for the yes-no recognition memory paradigm. (though note that this flexibility does not apply to all models that adopt the criterion shift assumption, e.g., Cary and Reder (2003) and a subset of the global matching models). Differentiation in SDT simply refers to the ordering of the distributions and not the encoding and retrieval processes that underlie the placement of the distributions (which is an integral part of the differentiation models as described earlier). A differentiation/fixed criterion version of SDT could arise from at least two factors: by assumption or by using an informed likelihood ratio decision rule. An informed likelihood ratio decision rule takes into account how well the test item *actually* matches the contents of memory and how well the item is *expected* to match memory based on properties of the stimulus and the experimental design. In other words, the system is informed about the experiment and the stimulus and uses this information in the decision process on a trial-to-trial basis. For example, such a model requires higher memory strength to endorse an item expected to produce a strong match to memory than an item expected to produce a lower match to memory (e.g., due to properties of the stimulus itself or the encoding conditions). These models necessarily produce a mirror effect when one class of items is better remembered than another. The SDT likelihood model has been rejected as a memory model (e.g., see Morrell, Gaitan, & Wixted, 2002; Stretch & Wixted, 1998) and fully informed likelihood models have been criticized on other grounds (e.g., see Balakrishnan & Ratcliff, 1996; Hintzman, 1994; McClelland & Chappell, 1998). Thus predictions for this class of models are not considered in this manuscript, though I return to this issue in the discussion.

1.4. Direct ratings paradigm

The differentiation models attribute the SBME to the encoding and decision rule inherent in the episodic memory system, while the models adopting the criterion shift assumption attribute the SBME to meta-cognitive assessment of the task. In principle, these two explanations are fundamentally different. In practice, they have been difficult to discriminate in part because the data provide little constraint on the models. The direct ratings paradigm introduced by Mickes et al. (2007) provides a method of obtaining additional data to further constrain the modeling assumptions under consideration.

Mickes et al. (2007) asked participants to report the strength of their memory in relatively fine detail (on a scale of 1–20 or 1–99). Using this direct rating method, they demonstrated that recognition memory behaves in accord with basic principles of signal detection theory. Namely, they showed an overlap between the empirical distribution of memory strength for foils and targets and the standard deviation of the foil distribution was lower than the standard deviation of the target distribution with a ratio of approximately 0.80.

I use the direct ratings paradigm to gather additional data in an attempt to address whether models that adopt the criterion shift or models that adopt differentiation are a more accurate description of subjective memory strength in episodic memory. Experiments 1 and 2 are standard SBME experiments; response bias is manipulated in Experiment 3. In all experiments distributions of memory strength are reported and compared to predictions based on the differentiation and criterion shift assumptions.

2. Experiment 1

Participants studied a list of items and then provided a direct rating of their memory strength for targets and foils in the test that followed. Each participant received a weak study list followed by a direct ratings test and a strong study list followed by a direct ratings test. The weak list contained items studied a single time and the strong list contained items studied five times. Assuming the ratings provided by participants in this paradigm reflect the subjective match of the test item to memory (following Mickes et al., 2007) then all models predict higher ratings for strong than weak targets. Critically, the predictions for foils differ for the models under consideration. Differentiation models predict lower memory strength for foils following a strong list than a weak list. A subset of the global matching models (described in Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990) predicts the opposite. The fixed strength assumption predicts that memory strength should not differ for foils following a strong or weak list.

2.1. Methods

2.1.1. Participants

A total of 30 Syracuse University students received partial course credit for participation.

2.1.2. Stimulus materials

The word pool consisted of 476 words ranging between 3 and 15 letters in length ($M = 7.64$) and between 3 and 12.7 log normative word frequency ($M = 7.84$) in the Hyperspace Analog to Language corpus (Balota et al., 2002; Lund & Burgess, 1996).[3]

2.1.3. Design

Participants received one weak study list (each item studied once) and one strong study list (each item studied five times) with a test following each list. For the strong list each target was studied before any item repeated for five consecutive blocks. Within each block, the order of words was randomly chosen. Participants were unaware of the blocks as each began immediately following the prior block with no interruption. The order of the weak and strong lists was randomly assigned for each participant. There was a short break between the two conditions indicating that a new study-test cycle was beginning.

The study lists each consisted of 50 words studied for 2.5 s followed by a blank screen for 500 ms. A 45 s arithmetic task separated the study and test lists. The test lists consisted of all 50 target items and an equal number of foil items randomly chosen from the word pool and randomly intermixed. The test was self-paced with a 500 ms blank screen separating each trial. During the memory test participants were asked to rate the strength of their memory on a 1–20 scale (20 = very strong memory). Instructions were read to the participants prior to beginning the experiment and appeared on the computer monitor prior to each test list. Following Mickes et al. (2007), participants were carefully instructed to reserve the rating of one for cases where they were absolutely sure that the item was not on the list (and they had absolutely no memory for any details of the item occurring on the list) and reserve the rating of 20 for cases where they were absolutely sure that the item was on the list and they remembered details about its occurrence. The instructions stated that participants should give a rating from the upper half of the scale to items they remembered from the study list and a rating from the lower half of the scale to items they did not remember being on the study list. Participants were further instructed that the specific rating given to any individual item should be carefully chosen based on the details of their memory for that item.

2.2. Results and discussion

The average strength rating for weak targets, weak foils, strong targets, and strong foils was computed separately for each participant (see Table 1). Targets ($M = 14.17$, $SD = 2.00$) had significantly higher direct ratings than foils ($M = 6.47$, $SD = 2.34$) demonstrating that participants were able to discriminate between targets and foils $t(29) = 13.72$, $P < .001$, $d = 3.54$. The strength rating for strong targets was greater than weak targets, $t(29) = 7.58$, $P \leqslant .001$, $d = 1.53$ as predicted by both the differentiation and criterion shift models. Critically, the pattern of results for the foils supports the differentiation models: The strength rating for foils following a strong list was lower on average than the direct ratings for foils following a weak list, $t(29) = 2.77$, $P = .005$, $d = 0.44$.

3. Experiment 2

The previous experiment demonstrates that direct strength ratings follow the pattern predicted by differentiation models. This experiment is identical except participants are first asked to explicitly state whether the item was studied or not and then give a direct strength rating. There are two advantages to this design. First, direct strength ratings and the strength based mirror effect in the HRs and

[3] For comparison the words ranged from 1 to 245 per million ($M = 18.66$) in Kucera and Francis (1967).

Table 1
Results for Experiments 1 and 2 where item strength was manipulated between lists.

	Weak	Strong
	Strength ratings	
Experiment 1		
Foil	7.03 (.42)	5.90 (.52)
Target	12.34 (.40)	15.99 (.47)
Experiment 2		
Foil	6.32 (.38)	5.65 (.43)
Target	12.39 (.38)	15.96 (.36)
	P("studied")	
Foil	0.18 (.02)	0.13 (.03)
Target	0.64 (.03)	0.90 (.02)

Note: Mean values are reported. Strength ratings vary between 1 and 20. Standard errors of the mean are reported in parentheses.

FARs are measured simultaneously. Second, this design allows a test of a prediction made by the criterion shift account. According to one possible criterion shift account, participants adopt a stricter criterion following a strong list than following a weak list. This predicts that the subjective strength for rejected items will be higher for the strong list than the weak list. For example, let us assume that the criterion for rejecting an item is nine for the weak list and raised to the more conservative value of 11 for the strong list (on a 20 point scale). The strength rating for rejected items should range from approximately 1–9 for the weak list and approximately 1–11 for the strong list, thus the average should be higher in the latter case even when the underlying strength distributions are identical.

3.1. Methods

3.1.1. Participants

A total of 36 Syracuse University students received partial course credit for participation. Five were excluded because they failed to follow directions (e.g., Debriefing revealed that these participants misused the direct ratings scale, treating it as a confidence scale where 20 meant that they were highly confident of their prior yes/no decision. Their average strength rating following a "no" response was more than 1 standard deviation above the mean and the difference between mean ratings following "no" and "yes" responses were small, supporting their confession). The analyses reported below are based on the remaining 31 participants.

3.1.2. Stimulus materials

The stimuli were identical to Experiment 1.

3.1.3. Design

The design was identical to Experiment 1 with one exception. Before providing a strength rating on a scale of 1–20, participants answered yes or no to the question "Was this item on the study list?"

3.2. Results and discussion

The results of this study (see Table 1) replicated the first experiment. The direct strength ratings for foils ($M = 5.98$, $SD = 2.07$) were lower than targets ($M = 14.17$, $SD = 1.77$), indicating above chance discrimination $t(30) = 15.61$, $P < .001$, $d = 4.25$. As predicted by differentiation models the strength rating for foils following a strong list were lower than foils following a weak list ($t(30) = 2.14$, $P = .021$, $d = 0.30$) and the strength rating for strong targets was greater than weak targets, $t(30) = 9.53$, $P < .001$, $d = 1.74$.

Replicating many prior studies, a strength based mirror effect was obtained. Strong lists had higher HRs ($t(30) = 10.04$, $P < .001$, $d = 2.01$) and lower FARs ($t(30) = 2.34$, $P = .013$, $d = 0.37$) than weak lists (see Table 1).

One interpretation of the criterion shift assumption predicts that the average strength rating for all items called "not studied" should be greater for the strong than the weak list. Contrary to this prediction the direct ratings for rejected items following the strong list ($M = 4.80$, $SD = 2.09$) were numerically lower than ratings following the weak list ($M = 5.24$, $SD = 1.81$) though this difference was not significant by traditional criteria, $t(30) = 1.53$, $P = .069$, $d = 0.23$.

The direct ratings were further analyzed with separate 2×2 ANOVAs for "studied" and "not studied" responses (item type: target vs. foil and condition: strong vs. weak). Note that this data is based on the subset of participants who had observations in each cell (the total number of subjects in each condition can be derived from the reported degrees of freedom), thus the following results should be interpreted with some caution. For "not studied" responses, there no main effect of strength ($F(1, 29) = 0.15$, $P = .903$), a main effect of item type $F(1, 29) = 30.77$, $P < .001$ but these were qualified by a significant interaction between the two variables, $F(1, 29)=5.13$, $P = .031$. Direct strength ratings tend to be higher for strong than weak targets but lower for strong than weak foils. For "studied" responses, ratings were higher for targets than foils $F(1, 23) = 205.66$, $P < .001$ and for strong than weak lists ($F(1, 23) = 13.88$, $P = .001$), with no interaction between the two variables $F(1, 23) = 1.67$, $P < .209$. Distributions of direct ratings as a function of response, condition, and item type were computed for all 31 participants and the mean of the distributions is plotted in Fig. 2.[4]

4. Comparison of Experiments 1 and 2

The direct strength ratings in Experiments 1 and 2 were directly compared using a $2 \times 2 \times 2$ mixed analysis of variance (ANOVA) with strength of list (strong or weak) and test item type (target or foil) as within-subject factors and experiment as a between-subject factor. Confirming the previous analyses, there was a main effect of list strength ($F(1, 59) = 51.31$, $P < .001$), a main effect of item type ($F(1, 59) = 428.83$, $P < .001$) and an interaction between the two ($F(1, 59) = 119.04$, $P < .001$). There was no main effect of experiment and no interactions between experiment and the other variables. Given the comparable results from the two experiments, Fig. 3 plots the combined response distributions averaged across all 61 participants. Like Mickes et al. (2007), the target and foil distributions overlap and tend to be truncated. The variance of the ratings distribution for each condition was computed for each participant. These values where then subjected to paired t-tests. Consistent with Mickes et al. (2007), the variance of the target distributions was greater than the variance of the foil distributions for the weak list, $t(65) = 9.45$, $P < .001$ (26.92 and 12.52, respectively) and strong list, $t(65) = 3.02$, $P = .004$ (16.18 and 11.79, respectively).

4.1. Distributional analysis

Observation of Fig. 3 suggests that the distributions for weak and strong foils differ and that the distributions for weak and strong targets differ. This observation was directly tested with a two-sample Kolmogorov–Smirnov (K-S) test comparing the weak and strong foil distributions for each participant and the weak and strong target distributions for each participant. K-S tests on the foil distributions demonstrated that 62% (38 of 61) of the participants produced different distributions (the test was marginally significant with a P-value of .056 for another three participants). For the target distributions, 75% (46 of 61) of the participants produced different distributions (the test was marginally significant with a P-value of .056 for another four participants). Empirical cumulative distribution functions (cdf), averaged over individual participants are plotted in Fig. 4. As established

[4] In this and other figures in this manuscript error bars are excluded for visual clarity but may be obtained by contacting the author.

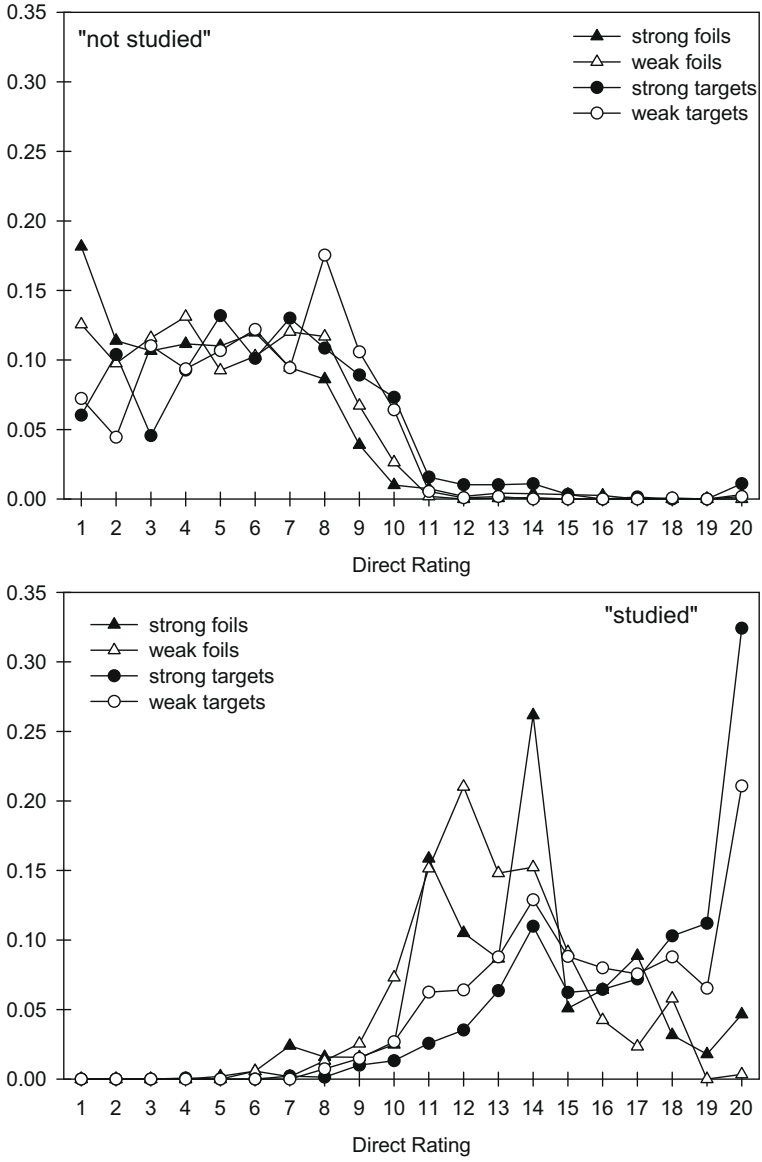

Fig. 2. Empirical distributions of direct ratings for strong and weak lists of Experiment 2 as a function of whether the participant first responded "studied" (bottom panel) or "not studied" (top panel).

by the K-S tests, the distributions for the weak and strong conditions differ for both targets and foil items, in accord with differentiation models.

At face value, these results are consistent with the differentiation models: the distribution of subjective memory strength decreases for foils and increases for targets following a strong list compared to a weak list. However, any ratings method, including the one used here, requires participants to map their internal memory strength onto an artificial scale provided by the experimenter (1–20 in this case). It is possible that participants adopt different mapping functions between lists. Though this cannot be strictly disconfirmed, the possibility is addressed in the next experiment.

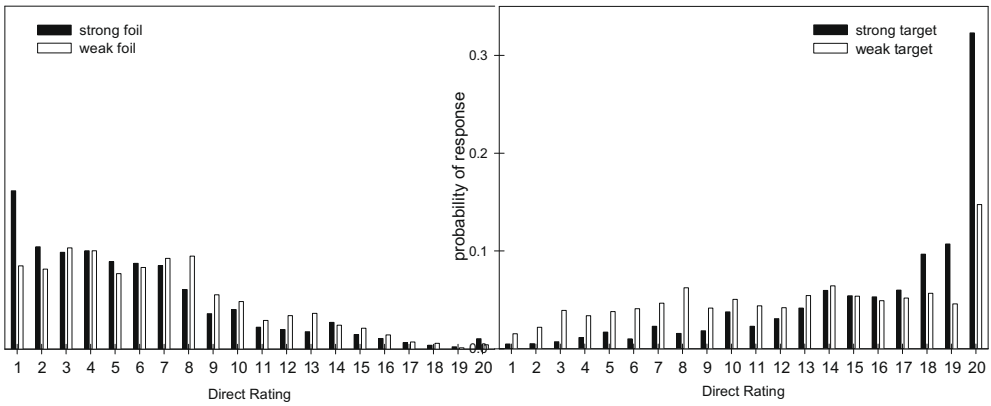

Fig. 3. Empirical distributions of the direct ratings for foils and targets following strong and weak lists collapsed over both Experiments 1 and 2. Foils are in the left panel and targets are in the right panel.

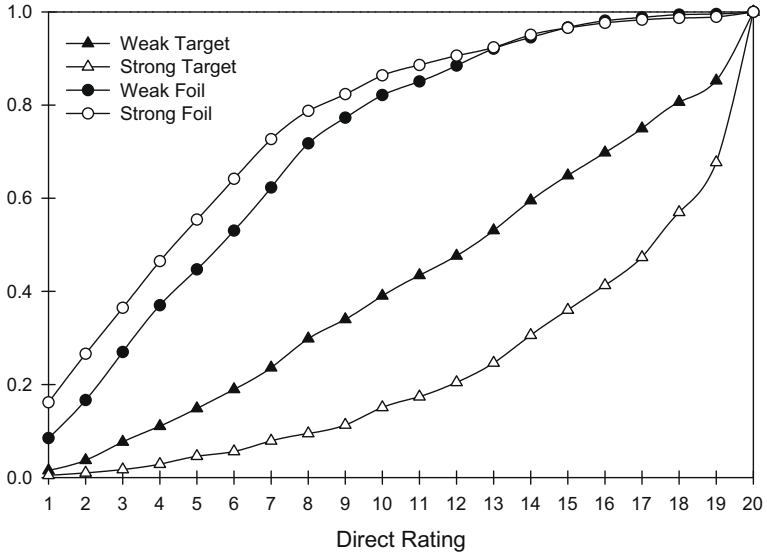

Fig. 4. The mean empirical cumulative distribution function for Experiments 1 and 2, showing the difference between the weak and the strong conditions for both foils and targets.

5. Experiment 3

According to the fixed strength assumption, the foil distribution remains constant despite encoding manipulations and any change in FAR must result from a change in criterion location between conditions. To account for the results of Experiments 1 and 2, theorists adopting the criterion shift assumption could assume that the mapping between internal memory strength and the ratings scale changes between a strong and weak list producing the illusion of a change in the foil distribution. In essence this suggests that the direct ratings method does not measure memory strength per se but instead measures some combination of memory strength and response bias. If this is the case, then manipulations that affect response bias (but not memory strength) should be reflected in the empirical distributions in a similar fashion. In the following experiment response bias is manip-

ulated by altering the number of target items on the test list while holding study conditions constant. If the direct ratings conflate response bias and memory strength then the manipulation should result in separate distributions for the two different bias conditions. If instead, direct ratings reflect memory strength then the manipulation should not change the distributions. To ensure that the response bias manipulation was successful, participants were asked for a binary "studied" or "not studied" decision and then asked for a direct rating on the 1–20 scale used throughout the manuscript.

5.1. Methods

5.1.1. Participants

A total of 40 Syracuse University students received partial course credit for participation. Three participants were excluded based on the same criteria as Experiment 2; reported analyses are based on the remaining 37 participants.

5.1.2. Stimulus materials

The stimuli were identical to Experiment 1.

5.1.3. Design

Participants received two study-test cycles with the composition of the test list varying between cycles. Each study list contained 100 words each presented once for 2.5 s followed by 500 ms interstimulus interval. The composition of the test list varied between lists. In the *70% targets condition*, 70 test items were targets and 30 test items were foils. In the *30% targets condition*, 30 of the test items were targets and 70 of the test items were foils. The order of the two conditions was randomly assigned for each participant with a break between the two conditions indicating that a new study-test cycle was beginning. Participants were correctly informed about the composition of the test list after each study list.

Participants engaged in 45 s of arithmetic after each study list. Each of the 100 test trials was self-paced and separated by a 500 ms blank screen. The memory test consisted of a yes–no decision followed by a direct rating of memory strength, identical to Experiment 2. Instructions for the direct strength ratings were identical to those used in the prior experiments.

5.2. Results and discussion

Increasing the number of targets on the test list caused participants to more readily endorse test items as "studied" but did not change ratings of memory strength as shown in Table 2.

A 2×2 ANOVA (test item: target vs. foil and condition: *70% targets* vs. *30% targets*) was conducted with the probability that an item is endorsed as studied (P("studied")) as the dependent measure. The ANOVA revealed a higher rate of endorsement for targets than foils, $F(1, 36) = 127.49$, $P < .001$, and for items in the *70% targets* condition than in the *30% targets* condition, $F(1, 36) = 6.01$, $P = .019$. The interaction was not significant, $F(1, 36) = 2.16$, $P = .151$. The experimental manipulation successfully elicited a change in response bias for responding "studied" such that participants were more willing to endorse any test item during *the 70% targets* condition than the *30% targets* condition. The lack of a significant interaction suggests that there was no difference in overall accuracy between the two conditions, rather the increased willingness to endorse an item was similar in magnitude for both targets and foils.

The distributions of direct ratings shown in Fig. 5 suggest that there is no difference in rated strength between the two conditions. This observation was confirmed by a 2×2 ANOVA (test item: target vs. foil and condition: *70% targets* vs. *30% targets*) on mean direct ratings which revealed higher ratings for targets than foils, $F(1, 36) = 67.60$, $P < .001$, no difference between the *70% targets* and *30% targets* conditions, $F(1, 36) = 1.51$, $P = .226$, and no interaction between the two variables, $F(1, 36) = 0.92$, $P = .344$.

The difference in P("studied") but lack of difference in mean direct strength ratings for the two bias conditions is an interesting finding that many readers might not have anticipated. Thus the nature of

Table 2
Results from Experiment 3. Each study word was studied once. The proportion of target items on the test list varied.

	Condition	
	70% targets	30% targets
Strength ratings		
Foil	8.14 (.40)	7.65 (.40)
Target	11.53 (.33)	11.36 (.40)
P("*studied*")		
Foil	0.33 (.03)	0.25 (.03)
Target	0.64 (.03)	0.60 (.03)

Note: Mean values are reported. Strength ratings vary between 1 and 20. Standard errors of the mean are reported in parentheses.

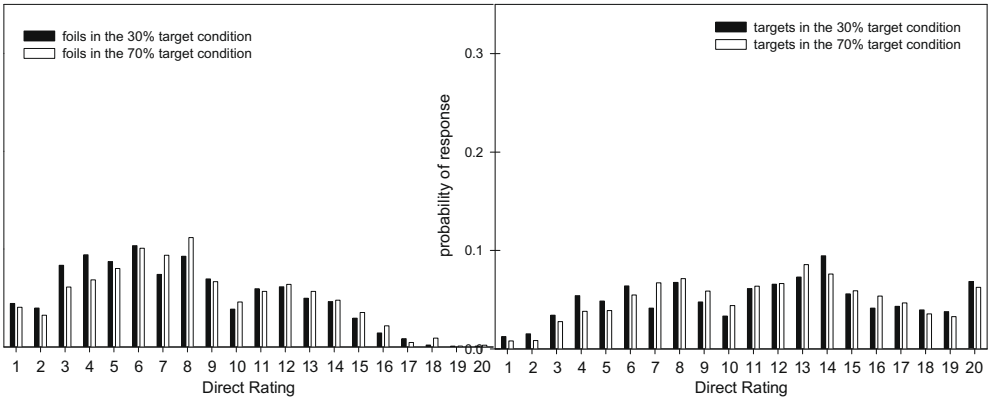

Fig. 5. The empirical distributions of direct ratings for Experiment 3 as a function of bias condition. Foils are in the left panel and targets are in the right panel.

these ratings was further analyzed through a $2 \times 2 \times 2$ ANOVA (response: "studied" vs. "not studied", test item: target vs. foil and condition: *70% targets* vs. *30% targets*). Note that this data is based on the subset of participants who had observations in each cell (the total number of subjects in each condition can be derived from the reported degrees of freedom), thus the following results should be interpreted with some caution. Main effects of response ($F(1, 33) = 178.80$, $P < .001$) and of test item $F(1, 33)=57.87$, $P < .001$ were qualified by a significant interaction between the two variables, $F(1, 33) = 57.49$, $P < .001$. All other effects failed to reach significance. Direct ratings reflect objective and subjective memory in that they are higher for targets ($M = 14.22$, SE = .35) than foils ($M = 12.19$, SE = .30) called "studied" but not different for targets ($M = 6.65$, SE = .31) and foils ($M = 6.41$, SE = .35) called "not studied." If the participant has no memory for the item, they say "not studied" followed by a low strength rating regardless of the objective status of the item. If the participant does remember the item, they say "studied" followed by a rating that is consistent with the objective state of the item. Perhaps details of the actual study event contribute to the higher rating for targets than foils. Distributions of direct strength ratings for items as a function of response, condition, and item type were computed for each individual participant and the mean of those distributions is plotted in Fig. 6.

5.2.1. Distributional analysis

Observation of Fig. 5 and the ANOVA based on mean ratings suggests that the distribution of memory strength for foils from the two bias conditions do not differ and likewise the target distributions from the two bias conditions do not differ. These observations were directly tested with a

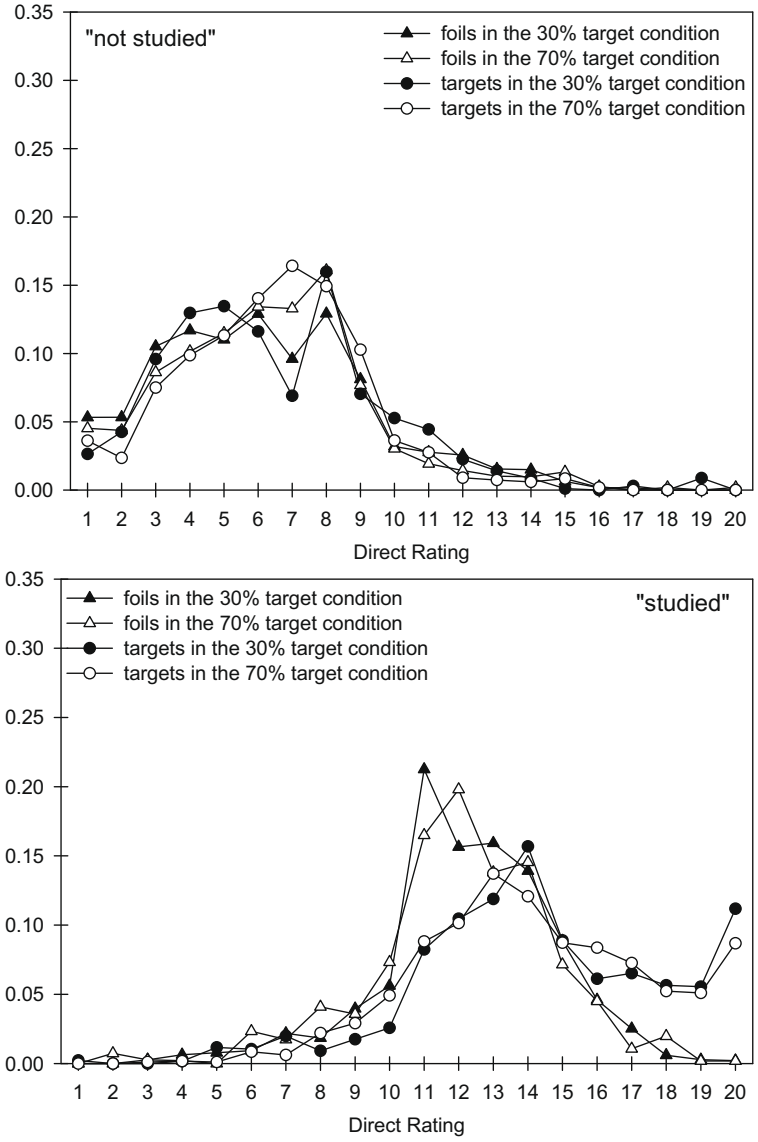

Fig. 6. The empirical distributions of direct ratings for Experiment 3 as a function of bias condition and binary response given by participants prior to the strength rating. Items called "not studied" are in the top panel, items called "studied" are in the bottom panel.

two-sample K-S test comparing the *30% target* condition and the *70% target* condition separately for targets and foils for each participant. K-S tests on the foil distributions revealed that the majority of participants (62%, 23 of 37) produced indistinguishable distributions for the two bias conditions (though one of those was marginally significant with a *P*-value of .0588). The target distributions showed a similar pattern with 59% (22 of 37) of participants producing distributions for the two bias conditions that were not different. The empirical cumulative distribution function, averaged over individual cdfs is plotted in Fig. 7. As established by the K-S tests, manipulating the proportion of targets on the test list did not alter the ratings of memory strength.

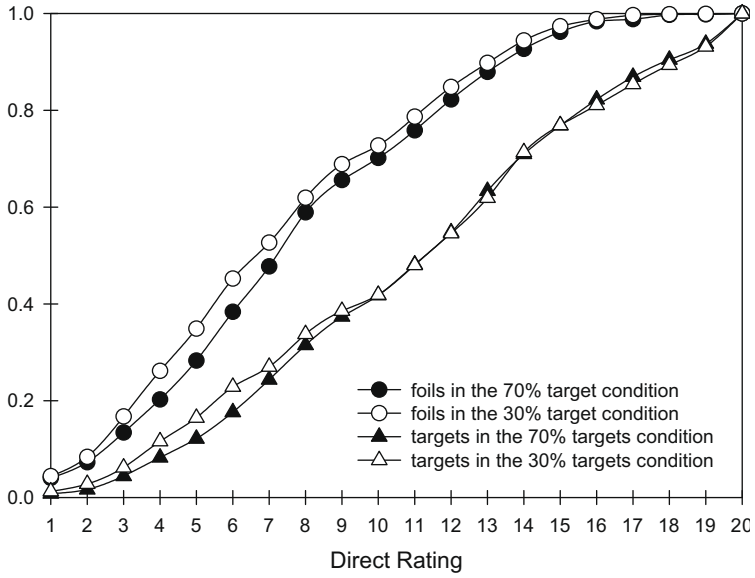

Fig. 7. The mean empirical cumulative distribution functions for Experiment 3, showing no difference between the 30% target and 70% target conditions.

I further analyzed the distributions of direct ratings for the participants who did produce different distributions for the two conditions. I conducted a directional K-S test to assess the nature of the differences in distributions. For target items, seven participants gave higher ratings to targets in the 70% target condition than in the 30% target condition (the remaining eight participants did the opposite). For foil items, five participants gave higher ratings to foils in the 70% target condition than in the 30% target condition (the remaining nine did the opposite). Within the minority of participants who did produce different distributions of direct strength ratings for the two bias conditions, there is not strong support for the fixed strength assumption.

The manipulation of the proportion of targets on the test list altered willingness to endorse an item as "studied" but did not change the direct ratings of memory strength. Note that the average difference in FAR between conditions is .08 here and .05 in Experiment 2. Also recall that the same word pool is used in all experiments, thus a model adopting the fixed strength assumption would assume that the distribution of subjective strength for foils is the same for all conditions in all experiments presented in this paper. The only way to allow changes the FAR is to change the location of the criterion for responding "studied." Because the change in FAR is similar in Experiments 2 and 3, it follows that the criterion location should be similar according to the fixed strength assumption. However, the empirical distributions in Experiments 1 and 2 are different (Figs. 3 and 4) for the majority of participants but the empirical distributions are not different (Figs. 5 and 7) for the majority of the participants in Experiment 3.

Under the fixed strength assumption, the different distributions for weak and strong foils in Experiments 1 and 2 are attributed to a different mapping between subjective memory strength and the ratings scale in the two conditions and the greater FAR for the weak condition is due to a more liberal criterion setting. Likewise the higher FAR for the *70% targets* condition in Experiment 3 is due to a more liberal criterion setting however, this is not accompanied by a mapping that produces distributions similar to those found in Experiments 1 and 2. In order to account for this full pattern of data, a fixed strength model could propose that participants adopt a mapping between their actual memory strength and the 1–20 scale that is different in all four conditions (weak, strong, 30% targets, 70% targets) despite the fact that the overall change in "studied/not studied" criterion location is approximately the same and thus the change in FAR is similar across experiments.

In contrast, the differentiation account requires no ad hoc assumptions. Instead, the pattern of data is consistent with the idea that participants faithfully report their subjective memory strength and do so consistently between lists in a given experiment and that subjective memory strength follows the predictions of the differentiation models such that the foil distribution decreases in strength as contents of the studied list become more accurate and complete representations of the studied items. A criterion for responding "studied" is placed along the scale as appropriate given the experimental conditions or instructions.

6. General discussion

In Experiments 1 and 2, memory strength was manipulated by repeating the studied items. The result was an increase in HR, decrease in FAR, and a corresponding change in the distributions of reported memory strength such that the target distribution shifted up in strength and the foil distribution shifted down in strength following a strong list. This is the pattern of data expected by the differentiation assumption if the direct ratings paradigm elicits ratings of subjective memory strength (i.e., see bottom panel of Fig. 1). In Experiment 3, response bias was manipulated by changing the proportion of targets on the test list and informing participants of that manipulation. The resulting data showed an increase in the HR and FAR but no change in the distributions of reported memory strength. Again, this pattern of data supports differentiation models under the assumption that the direct ratings paradigm elicits subjective memory strength and manipulations of response bias alter the location of the criteria for calling an item "studied."

6.1. The REM Model

The experimental results are consistent with the qualitative pattern of data inherent in the Retrieving Effectively from Memory model (REM; Shiffrin & Steyvers, 1997). Nonetheless, I tested the hypothesis that the mapping between internal subjective memory strength and the direct ratings scale was constant between strong and weak lists within the REM framework. The approach was to estimate the mapping between the distributions of memory strength in REM and the empirical ratings (on the 1–20 scale) for the weak list (i.e., what is the criterion for memory strength in REM that maps onto an empirical rating of 1, 2, and so on). This set of mappings was then used to predict distributions for the strong list and for items from both lists called "studied" and "not studied."

I briefly describe the REM model and refer readers interested in additional details to the original paper (Shiffrin & Steyvers, 1997) or to a detailed application of REM to the SBME (Criss, 2006). In REM, an incomplete and error-prone copy of the studied item is stored in episodic memory along with the current context. Episodic memory features ($n = 20$) are initialized as zeros, indicating a lack of information. Feature values range from 1 to infinity (in theory, in practice the largest value is often less than 30 and depends on the g parameter) and are sampled from a geometric distribution ($g = .35$). During study each zero is replaced with a feature value with some probability ($u = 0.04$) per unit of time (t) otherwise a value for that feature is not stored (thus storage is incomplete). Given that a feature value is stored, the correct value is stored with some probability ($c = 0.70$) otherwise, a random feature value is selected from the geometric distribution (thus storage is error-prone). Additional study results in the storage of more features (i.e., replacing the remaining zeros) but not the correction of previously stored feature values.

For simplicity assume that context features perfectly isolate the encoded list for comparison during the memory test.[5] During retrieval, item features of test item j are compared to each trace stored in memory, indexed by i, and a likelihood ratio is computed as follows,

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

[5] See Criss and Shiffrin (2004b) or Shiffrin and Steyvers (1997) for a discussion of context. Simulations using the context features will result in the same pattern of data for the current manipulations.

This is the REM equation for the subjective likelihood that the test stimulus j matches memory trace i for simulated subject k. The number of non-zero features that mismatch is nq and the number of non-zero features that match and have the value v is nm. Features with a value of 0 do not contribute to the decision because zero indicates a lack of information. The decision about whether test stimulus j was studied or not is based on the subjective memory strength, defined as the average of the likelihood ratios (the odds). If the average is greater than some criterion (*criterion* = 1), test stimulus j is called "studied" otherwise it is called "not studied"

That REM predicts a SBME and a shift in the distributions for strong and weak foils (and targets) has been demonstrated elsewhere (e.g., Criss, 2006). Of interest here is a demonstration that the mapping between the experimenter-provided scale (1–20) and the internal memory scale (i.e., the odds value described above) need not change between the weak and strong list to account for the pattern of empirical distributions obtained here. Ordinal regression was used to obtain the criteria along the theoretical subjective memory strength dimension that best fit the empirical distributions of direct ratings for the weak lists.[6] These criteria were then used to generate predicted distributions for strong and weak lists for items called "studied" and "not studied." Fig. 8 shows the predicted distributions and can be compared to Fig. 2 showing the empirical distributions from the same analysis.

To obtain stable estimates for each of the criteria it is critical to have a large number of observations for each rating and to have each rating present in the analysis, thus the model was fit to group data, not to individual participants. Fitting individual participants has obvious advantages, especially if the goal is to recover the true parameters that generated the participant data (c.f., Cohen, Sanborn, & Shiffrin, 2008). The goal of this analysis is less ambitious and that is to simply demonstrate that the REM model can predict distributions that are qualitatively similar to observed distributions using the same criteria for strong and weak lists. The empirical distribution of direct ratings for targets was generated by simply compiling distributions from each individual participant into one large distribution. Thus the empirical distribution for targets consisted of a total of 3050 observations (50 per participant, 61 participants). Empirical distributions for foils were similarly constructed.

The theoretical distribution of subjective memory strength for targets and foils were obtained as follows. All parameter values identified above were held constant at the standard values (e.g., see Shiffrin & Steyvers, 1997 and Criss, 2006) except the t parameter ($t = 10$ for weak targets and $t = 17$ for strong targets) which was fit to the empirical values of P ("studied") from Experiment 2.[7] Then 61 simulated participants were ran and the odds value for each test item was collected. Because the odds values are highly skewed (especially the target distributions) they were log transformed.

Values from the weak theoretical distributions served as predictor (covariate) variables and values from the empirical distributions for the weak list served as response (dependent) variables. To generate pairs of predictors and responses that served as input to the ordinal regression, each observation from the empirical target distribution was paired with one randomly selected observation from the theoretical target distribution. Empirical–theoretical pairs for foils were similarly chosen. The ordinal regression was given a total of 6100 pairs (one empirical and one theoretical value both drawn for the respective target or the respective foil distributions). The regression was not informed which pairs corresponded to targets and which corresponded to foils (just as participants do not know which are which but must make a prediction based on the subjective strength of the test item).

The 19 parameters estimates (thresholds) obtained from the ordinal regression are criteria along the subjective memory dimension in REM (i.e., log odds) that best fit empirical ratings for the weak lists. Next, histograms of strong and weak, target and foil distributions were created using these same criteria. To compare the predicted distributions to empirical distributions, histograms were computed separately for items that would be called "studied" and those that would be called "not studied" by the

[6] DeCarlo (2003) demonstrates how to use ordinal regression to fit SDT models in SPSS. Note that the method here is slightly different. SDT assumes subjective strength is normally distributed. Here the actual theoretical distributions (not an assumed Gaussian) are used as predictors.

[7] The current parameter for the strong list ($t = 17$) underestimates the HR. A higher value of t combined with a more lenient criterion for the strong list accurately captures the HR (and FAR) for the strong condition. This is not a persistent flaw of the model, many prior applications to the SBME and study repetitions in general have not faced this problem, thus it is not pursued further. Note that this criterion shift just described in the REM framework is opposite that required by a fixed strength assumption.

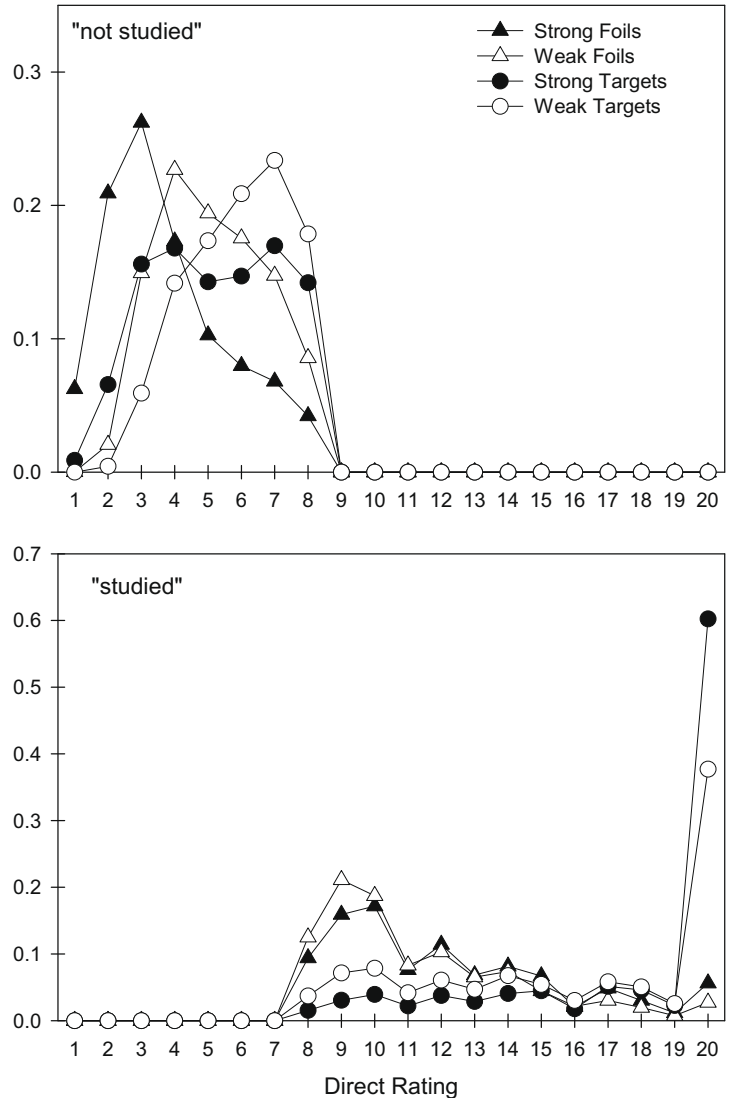

Fig. 8. The predicted distribution of ratings for strong and weak targets and foils given a "studied" or "not studied" response. The criteria mapping the REM distributions to the direct ratings scale is based on fitting the weak condition only and using those same criteria for the strong condition.

model. The default criterion of 1 (0 on a log scale) was used to determine whether an item was classified as "studied" or not and the same criterion was used for all conditions. Note that there are both "studied" and "not studied" responses given a direct rating value of 8 because a log odds of zero is falls within the criterion value for a direct rating response of eight.

The results shown in Fig. 8 are similar to the empirical distributions plotted in Fig. 2. The bottom panels show that HRs receive higher ratings than FARs. Strong hits receive higher ratings than weak hits, especially in the highest rating, although the model gives a higher proportion of responses to the highest value of 20 than do people (perhaps because participants are instructed to reserve 20 for extreme cases). In the model this results from the extended heavy tail of the target distributions

(e.g., see bottom panel of Fig. 1). The top panel of Fig. 8 shows lower ratings to correct rejections (CRs) than to misses and lower ratings for strong CRs than weak CRs. For "not studied" responses, the model assigns fewer of the items to the lowest value of 1 than do people. In the model this is because the foil distributions truncate somewhat abruptly with almost no tail, in contrast to the extreme tail of target distributions. The discrepancies between the data and the model could potentially be corrected with an exhaustive search for the best fitting parameters, accounting for individual variability in the empirical distributions of ratings, and/or allowing parameters to vary between conditions. The important point illustrated here is that the mapping between memory strength and the experimenter-provided rating scale need not change between the weak and strong lists within the REM framework to account for the general patterns of observed data.

6.2. Criterion Shifts

The data are consistent with the criterion shift assumption implemented in many models if several auxiliary assumptions are adopted. Of course, the specific auxiliary assumptions depend on the specific model under consideration. As an illustration, consider one possibility – the unequal variance SDT model shown in the top panel of Fig. 1. The first assumption is that participants are less willing to endorse an item as studied following a strong than a weak list which accounts for the reduced FAR in the former condition (the standard criterion shift assumption). The second assumption is that within a single experiment an individual participant adopts a different mapping between their internal subjective memory strength and the 1–20 scale for the tests following a weak and a strong list (this explains the different distributions of direct ratings for foils in Experiments 1 and 2). The final assumption is that the mapping between the internal memory strength and the 1–20 scale was different depending on the number of targets on the test list (e.g., for Experiment 3). However, the mapping adopted by participants for the conditions in Experiment 3 does not resemble that used in Experiments 1 and 2 despite the fact that the change in FAR between conditions is similar in magnitude for both experiments (and presumably the change in criterion is also similar in magnitude). Of course both the bias manipulation and the strength manipulation are between-list manipulations and one cannot be sure that participants adopt the same strategy for both lists. Further, the bias and strength manipulations were conducted in separate experiments with different participants. A more powerful demonstration would be to replicate the results reported here in a within-subject manipulation.

This set of data cannot unequivocally confirm or disconfirm the fixed strength assumption. In part, this is due to the flexibility of the criterion settings allowed in models adopting the fixed strength assumption. Some theorists might find the assumption that the mapping between internal memory strength and the experimental scale is free to take any form for any experimental condition unsatisfactory; others might judge that flexibility to be an asset. There exists no explicit memory model of when, how, and to what degree criteria change between conditions. Thus a model that relies on criterion placement as the primary factor determining memory performance cannot be directly tested and confirmed or disconfirmed. Consequently, debates about whether other experimental manipulations such as the similarity between targets and foils or normative word frequency alter the criterion or memory strength appear with some regularity in the literature (e.g., Benjamin & Bawa, 2004; Glanzer & Adams, 1990; Miller & Wolford, 1999; Wixted & Stretch, 2000). An alternative for SDT models of recognition memory is to adopt a differentiation account. This remedy is not obviously applicable to models that assume FARs are due to pre-experimental familiarity or semantic memory (e.g., Cary & Reder, 2003 and some versions of dual process models outlined in Yonelinas, 2002).

Despite being consistent with both differentiation models and some criterion shift accounts, the current set of experiments represents significant progress. Prior to Criss (2006), few if any papers discussed differentiation as a theoretical mechanism underlying the strength based mirror effect. That differentiation and the fixed strength assumption remain in competition with no unambiguous winner attests to both the difficulty of addressing the question and the importance of bringing resolution to the debate. To aid progress in understanding the theoretical mechanisms underlying recognition memory, modelers should begin to incorporate specific, testable claims about how and when criteria change.

6.3. Classes of mirror effects

In their seminal work on mirror effects, Glanzer and Adams (1985, 1990) demonstrated the mirror effect with a number of different experimental manipulations (e.g., word frequency, study time, concreteness, pictures vs. words, etc.) that resulted in one class of items that were better remembered than another class of items via a higher hit rate and lower false alarm rate. They further demonstrated that the mirror effect was typically accompanied by concentering of the distributions (Glanzer, Adams, & Iverson, 1991). That is, the distribution of targets and foils of the class of items with poorer performance are both shifted toward the midpoint relative to the distributions for the condition with higher performance. These findings of concentering and of a higher HR and lower FAR for the class of items with superior memory were so ubiquitous that they were deemed a "regularity of memory" and many refer to "the" mirror effect (Glanzer, Adams, Iverson, & Kim, 1993, c.f., Criss & Shiffrin, 2004a). Glanzer and colleagues championed the attention likelihood model, one example of a fully informed likelihood ratio model, where the decision about whether or not to call an item "studied" takes into account properties of the test stimulus and experimental conditions. Fully informed likelihood models naturally and necessarily predict a mirror pattern.

Stretch and Wixted (1998) suggested a fundamental difference between the underlying cause of strength based (e.g., repetition during a study list) and frequency based (e.g., normative word frequency) mirror effects. Namely, they attributed the frequency based mirror effect to differences in the underlying distributions and the strength based mirror effect to a criterion shift.

The differentiation models were developed around the same time that Stretch and Wixted (1998) distinguished between the two types of mirror effects. They too attribute the word frequency mirror effect and strength based mirror effects to different theoretical mechanisms (though not the same mechanisms as Stretch and Wixted). According to these models, uncommon words tend to be composed of uncommon features (e.g., come from a distribution with a lower value of the g parameter in REM) and common words tend to be composed of common features (e.g., higher value of the g parameter). Critically, uncommon features are more diagnostic resulting in superior performance for low frequency words (see Criss and Malmberg (2008), Criss and Shiffrin (2004a), Malmberg, Steyvers, Stephens, and Shiffrin (2002), Malmberg, Zeelenberg, and Shiffrin (2004) for extensive discussion of the REM account of word frequency). The differentiation models are "partially-informed" likelihood models with respect to word frequency because the average frequency of features in the environment (c.f., the inclusion of the long run average of the g parameter in Eq. (1)) is taken into account in the decision rule which contributes to the word frequency mirror effect. However, the frequency of any individual test item is not considered in the decision rule. Differentiation models are uninformed as to the experimental conditions and such information is not taken into account in the decision rule (e.g., see the lack of parameters specific to the experimental list such as u in Eq. (1)). Both word frequency and repetition act to create different distributions of memory strength in the differentiation models and dual mirror effects are predicted when word frequency and repetition are manipulated within an experiment. The word frequency mirror effect is due to feature diagnosticity and the list strength mirror effect is due to differentiation.

Among other findings, Stretch and Wixted (1998) demonstrated that participants do not produce a within-list mirror effect even when the test items that belong to strong and weak conditions were made transparently clear (e.g., by using different font colors and/or explicit instructions). If the decision rule took into account the information about the test item, then a within-list mirror effect should have been observed. On this basis, they ruled out fully informed likelihood ratio models as an explanation for this class of mirror effects. This same argument does not apply to the differentiation models because they have no parameter to incorporate information about the encoding and/or test conditions into the decision rule. Thus the finding of a null within-list mirror effect for unrelated foils is consistent with differentiation models. Differentiation models could assume that participants adopt different criteria for items classified as strong and weak during test, but empirical evidence does not yet support this strategy.

A second approach to define strong vs. weak conditions following a mixed study list is to test foils that are similar to strong or weak targets. In such a paradigm, participants could adopt one of three methods for making a decision: (1) adopt different criteria for the strong vs. weak conditions; (2)

include a category cue in the probe to preferentially compare the test item to members of the same category; and (3) ignore the relationship between category and strength and compare the test item to all memory traces. Strategy 1 predicts a within-list SBME for both classes of models. Predictions for Strategies 2 and 3 are difficult to intuit within differentiation models because both category similarity and repetition will influence the decision and the two variables interact based on the degree of similarity between the test and study items (e.g., see Criss, 2006; Criss & Shiffrin, 2004b; Malmberg, Holden, & Shiffrin, 2004).

The simulations in Criss (2006) showed that differentiation models predict that a similar foil matches a strong memory trace *better* than a weak memory trace (opposite the pattern for an unrelated foil). These simulations considered the case where a foil is similar to one strong or one weak target and the predictions were confirmed empirically. Morrell et al. (2002) conducted experiments where foil items were similar to multiple items from the same category. They presented two categories of items to participants with all members of one category in the weak condition and all members of the other category in the strong condition. No within-list mirror effect was observed, evidence against fully informed likelihood ratio models. Predictions of differentiation models do not mimic predictions of fully informed ratio models for reasons already explained. Simulations and additional empirical results using this type of paradigm seem necessary to clarify the implications for differentiation models. In sum, fully informed likelihood models face serious challenges in accounting for the data presented here. In contrast, there is little compelling evidence against differentiation models as an explanation for the strength based mirror effect.

The model of Dennis and Humphreys (2001) is not a fully informed likelihood ratio model in that it does not take into account the parameter values for each individual test item. However, it may take into account some general properties of experiment, thus predictions about the SBME depend on how this is implemented. Starns (2009) recently applied that model to the strength based mirror effect by treating the model as a partially-informed likelihood model and adjusting the amount of evidence required for a "studied" response based on the expected difficulty of the test list (such information was provided in the experimental instructions). Implications of the current data for this model await further development of the theory and its application to the SBME.

6.4. Response bias

Rotello, Macmillan, Hicks, and Hautus (2006) manipulated the expected proportion of targets on the test list and collected remember–know responses, yes–no judgments, and confidence ratings. They report substantial response bias in both the HR and FAR and in ROCs. In particular, the ROCs for the different bias conditions are composed of different data points that lie on the same curve, as expected by SDT if bias manipulations act to simply shift the criteria. In contrast, Experiment 3 reported here demonstrated response bias in *P*("studied") but not in the distributions of direct ratings (although there is a slight hint of a similar finding in the mean direct ratings and mean cdf). A number of differences exist between the methods used in the two studies. For example, Rotello et al. asked for a studied/not studied response followed by a confidence rating (very sure, somewhat sure, guessing) and a remember/know judgment. In contrast, the direct ratings method used here emphasized memory strength and the remembered details (or lack thereof) of the memory for the test item rather than confidence in the accuracy of the participant's response. Another difference in procedure is that Rotello et al. conducted a between-subject manipulation whereas Experiment 3 manipulated response bias within-subject (necessary in order to conduct K-S tests on an individual participant) perhaps resulting in an anchoring effect in the direction of the first memory test (the differences in mean *P*("studied") between condition are twice as large in Rotello et al.'s data).

A number of other experiments manipulating response bias (i.e., through either payoffs or varying base rates) fail to replicate this critical prediction of SDT (e.g., Balakrishnan, 1998; Mueller & Weidemann, 2008; Van Zandt, 2000) and may provide insight as to why the distributions of ratings did not differ for the bias conditions in Experiment 3. One possibility is the difference in shape of ROCs across bias conditions is due to a fixed decision criteria accompanied by changes in the variance of the signal and noise distributions, perhaps due to the use of different stopping rules during stimulus sampling in the different bias conditions (Balakrishnan, 1999). Another possibility is that the mapping

between the internal response and the experimenter-provided response scale is noisy (Mueller & Weidemann, 2008). If the level of noise increases with the number of responses categories, so that more noise accompanies confidence ratings or direct memory strength ratings than a binary decision, the shape of the ROC may differ. Finally, there is evidence that information continues to accumulate between two successive judgments and that the information contributing to the 2nd decision (e.g., confidence rating) is not necessarily the same as the information that lead to the1st decision (e.g., yes/no; Curran, Debuse, & Leynes, 2007; Malmberg & Xu, 2007; VanZandt & Maldonado-Molina, 2004). In the current experiments the direct strength rating came after the yes/no decision and may have been based on more accurate information. The inconsistencies in the data and the underlying cause of differences in the shape of the ROC under response bias manipulations clearly deserves further investigation. With respect to Experiment 3, it is possible that any of these three factors are playing in a role in the lack of difference in the ratings distribution despite differences in *P*("studied") across bias conditions. However interesting, this does not detract from the overall message of the current manuscript that the fixed strength assumption deserves serious scrutiny as does the criterion shift account of the SBME. If anything, this discussion serves to illuminate the progress that could be made by moving away from simple SDT models to more comprehensive process models that include specific assumptions about encoding, representation, retrieval, and decision rules.

6.5. Differentiation models

Differentiation models successfully account for the data presented here and a wide range of other data without the use of differential mapping between strong and weak conditions and without the arbitrary or post hoc use of criterion shifts. Further, the differentiation models specify the processes underlying encoding and retrieval and are capable of providing testable predictions about human performance over a wide range of experimental paradigms and manipulations. In contrast, SDT is a model of decision making in the presence of noise. It makes no assumptions about the processes that underlie episodic memory or the mechanisms that drive encoding or retrieval. A simplifying assumption originally adopted in SDT applications to recognition memory, the fixed strength assumption, has persisted and has even been adopted in mechanistic models (e.g., Reder et al., 2000) despite little empirical evidence that the distribution of memory strength for foil items remains constant and/or is derived from pre-experimental familiarity. The issue of fixed criteria vs. fixed distributions is no longer one of convenience, as it might have been in the early years of mathematical modeling of recognition memory, but has become an important theoretical issue. It seems that substantial progress could be gained by focusing on process models that describe the encoding and retrieval stages in detail along in combination with the decision process.

In one such attempt, Criss (submitted for publication) measured reaction time in a two experiments similar to those presented here but with a simple yes/no response: a SBME manipulation and a bias manipulation. The accuracy results were prototypical and thus do not discriminate between differentiation and the criterion shift models. To gain leverage, the predictions from such models were mapped onto parameters of Ratcliff's diffusion model (RDM, Ratcliff, 1978). The RDM is a sequential sampling model used to analyze reaction time in tasks with two response options. Upon presentation of a stimulus information accumulates from the starting point until one response boundary is reached. The rate of accumulation of evidence (i.e., drift rate) measures the quality of evidence provided by the stimulus (e.g., subjective memory strength, see Ratcliff, 1978; Ratcliff, Thapar, & McKoon, 2004) and the starting point between the two response boundaries measures preferential bias toward one response (e.g., Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). Consistent with the differentiation framework, the bias manipulation was best characterized as changes in the starting point and the strength manipulation was best characterized as changes in the rate of information accumulation for both targets and foils.

Only two models (e.g., McClelland & Chappell, 1998 and Shiffrin & Steyvers, 1997) include differentiation and are consistent with the reaction data just described and distribution of direct ratings presented here. The differentiation account implies an active response to stimulus novelty. Indeed, much evidence suggests that animals (e.g., human infants and adults, rats, etc.) orient to a novel stimulus and that this novelty orienting response is supported by the some of the same neural structures

that support episodic memory (e.g., Dias & Honey, 2002; Fantz, 1964; Knight, 1984, 1996). The success of the differentiation models in accounting for recognition memory along with evidence for the intrinsic detection of novelty suggests that these models provide fertile grounds for further theoretical development and guidance in understanding episodic memory.

Acknowledgments

Mark Steyvers provided inspiration and guidance for theoretical and empirical aspects of this article, his contribution is appreciated. The author would like to thank Marc Howard, and members of his MEMlab for useful and thought provoking discussion. I thank Caren Rotello, John Wixted, Neil Macmillan, Bill Hockley, Mike Kahana, and several anonymous reviewers for useful commentary on this and previous versions of this manuscript.

References

Balakrishnan, J. D. (1998). Measures and interpretations of vigilance performance. Evidence against the detection criterion. *Human Factors, 40*, 601–623.

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1189–1206.

Balakrishnan, J. D., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance, 22*(3), 615–633.

Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., Simpson, G. B., et al. (2002). *The English lexicon project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords*, Washington University. <http://elexicon.wustl.edu/>.

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.

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.

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin and Review, 15*, 692–712.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55*(4), 461–478.

Criss, A. H. (submitted for publication). *Differentiation and response bias in episodic memory: Evidence from reaction time distributions*.

Criss, A. H., & Malmberg, K. J. (2008). Evidence in favor of the early-phase elevated-attention hypothesis: The effects of letter frequency and object frequency. *Journal of Memory and Language, 59*, 331–345.

Criss, A. H., & McClelland, J. L. (2006). 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, 55*(4), 447–460.

Criss, A. H., & Shiffrin, R. M. (2004a). Interactions between study task, study time, and the low frequency hit rate advantage in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*(4), 778–786.

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.

Curran, T., Debuse, C., & Leynes, P. A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 2–17.

DeCarlo, L. T. (2003). Using the PLUM procedure of SPSS to fit unequal variance and generalized signal detection models. *Behavior Research Methods, Instruments, and Computers, 35*, 49–56.

Dennis, S., & Humphreys, M. S. (2001). The role of context in episodic recognition: The bind cue decide model of episodic memory. *Psychological Review, 108*, 452–478.

Dias, R., & Honey, R. C. (2002). Involvement of the rat medial prefrontal cortex in novelty detection. *Behavioral Neuroscience, 116*(3), 498–503.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention familiar patterns relative to novel ones. *Science, 146*, 668–670.

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.

Glanzer, M., Adams, J. K., & Iverson, G. J. (1991). Forgetting and the mirror effect in recognition memory: Concentering of underlying distributions. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*(1), 81–93.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*(3), 546–567.

Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 201–205.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list strength paradigm. *Journal of Experimental Psychology: Learning Memory and Cognition, 21*, 302–313.
Hockley, W. E., & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory and Cognition, 35*, 679–688.
Kim, K., & Glanzer, M. (1993). Speed versus accuracy instructions, study time, and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 638–652.
Knight, R. T. (1984). Decreased response to novel stimuli after prefrontal lesions in man. *Electroencephalography and Clinical Neurophysiology, 70*, 9–20.
Knight, R. T. (1996). Contribution of human hippocampal region to novelty detection. *Nature, 383*(6597), 256–259.
Kucera & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.
Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74*, 100–109.
Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers, 28*, 203–208.
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*, 319–331.
Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory and Cognition, 30*(4), 607–613.
Malmberg, K. J., & Xu, J. (2007). On the flexibility and on the fallibility of associative memory. *Memory and Cognition, 35*(3), 545–556.
Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by Midazolam. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 540–549.
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.
Mickes, L., Wixted, J. T., & Wais, P. (2007). A direct test of the unequal variance signal detection model of recognition memory. *Psychonomic Bulletin and Review, 14*(5), 858–865.
Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review, 106*, 398–405.
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.
Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin and Review, 15*, 465–494.
Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review, 73*, 44–58.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.
Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experiment Psychology: Learning, Memory, and Cognition, 16*(2), 163–178.
Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50*, 408–424.
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.
Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. (2006). Interpreting the effects of response bias on remember–know judgments using signal-detection and threshold models. *Memory and Cognition, 34*, 1598–1614.
Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theroretical 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.
Singer, M., Gagnon, N., & Richards, E. (2002). Question answering strategy: The effect of mixing test delays. *Canadian Journal of Experimental Psychology, 56*, 41–57.
Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory and Cognition, 34*, 125–137.
Starns, J. (2009). Enhancing lure rejection by strengthening studied items: Contrasting encoding- and retrieval-based mechanisms from REM and BCDMEM. In *Talk presented at the context in episodic memory symposium*, West Palm Beach, FL.
Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1379–1396.
Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 582–600.
VanZandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*(6), 1147–1166.
Verde, M., & Rotello, C. (2007). Memory strength and the decision process in recognition memory. *Memory and Cognition, 35*(2), 254–262.
Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58*, 140–159.
Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology, 3*, 316–347.
Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review, 107*, 368–376.
Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.