Journal of Experimental Psychology:
Learning, Memory, and Cognition
2010, Vol. 36, No. 2, 484–499


0278-7393/10/$12.00 DOI: 10.1037/a0018435

# Differentiation and Response Bias in Episodic Memory: Evidence From Reaction Time Distributions

Amy H. Criss
Syracuse University

In differentiation models, the processes of encoding and retrieval produce an increase in the distribution of memory strength for targets and a decrease in the distribution of memory strength for foils as the amount of encoding increases. This produces an increase in the hit rate and decrease in the false-alarm rate for a strongly encoded compared with a weakly encoded list, consistent with empirical data. Other models assume that the foil distribution is unaffected by encoding manipulations or the foil distribution increases as a function of target strength. They account for the empirical data by adopting a stricter criterion for strongly encoded lists relative to weakly encoded lists. The differentiation and criterion shift explanations have been difficult to discriminate with accuracy measures alone. In this article, reaction time distributions and accuracy measures are collected in a list-strength paradigm and in a response bias paradigm in which the proportion of test items that are targets is manipulated. Diffusion model analyses showed that encoding strength is primarily accounted for by changes in the rate of accumulation of evidence (i.e., drift rate) for both targets and foils and manipulating the proportion of targets is primarily accounted for by changes in response bias (i.e., starting point). The diffusion model analyses is interpreted in terms of predictions of the differentiation models in which subjective memory strength is mapped directly onto drift rate and criterion placement is mapped onto starting point. Criterion shift models require at least 2 types of shifts to account for these findings.

*Keywords:* episodic memory, memory models, recognition memory, reaction time distributions, diffusion model

Recognition memory experiments ask participants to endorse target items that were studied on an earlier list and reject foil items that were not studied. Manipulations that improve recognition memory accuracy often do so via a mirror effect: the simultaneous increase in the probability of correctly endorsing a target item (hit rate, or HR) and decrease in the probability of erroneously endorsing a foil item (false-alarm rate, or FAR; e.g., Glanzer & Adams, 1990). Stretch and Wixted (1998) delineated two classes of mirror effects: stimulus based and strength based.

## Stimulus-Based Mirror Effects

One example of a stimulus-based mirror effect is the word frequency mirror effect (WFME) wherein words of low environmental frequency (LF) have higher HRs and lower FARs than words of high environmental frequency (HF). The WFME continues to be the center of much debate, as it has been attributed to several underlying mechanisms (e.g., Dennis & Humphreys, 2001; Glanzer & Adams, 1990; Malmberg & Nelson, 2003; McClelland & Chappell, 1998; Reder et al., 2000; Shiffrin & Steyvers, 1997). Though accounts of the WFME vary at the conceptual level and in the details, most share the assumption that the WFME is a function of properties of the stimuli themselves. Studies have also demonstrated mirror effects based on other stimulus properties such as letter frequency, orthographic neighborhood size, neighborhood density, and pictures versus words (Criss & Malmberg, 2008; Glanc & Greene, 2007; Glanzer & Adams, 1985; Heathcote, Ditton, & Mitchell, 2006).

## Strength-Based Mirror Effects

In contrast, the strength-based mirror effect (SBME) arises not from properties of the stimuli but from encoding conditions. For example, HRs are higher and FARs are lower for study lists in which accuracy is improved by increasing study time (e.g., Ratcliff, Clark, & Shiffrin, 1990; Stretch & Wixted, 1998). Higher HRs for targets from a strongly encoded list compared with a weakly encoded list are predicted by all models. Of greater theoretical interest is why the FAR differs between the strong and weak lists. This is puzzling because there are no differences between weak and strong foils other than encoding conditions, and foils, by definition, are not presented for encoding. Two possible explanations have been explored in detail in the literature and are considered here (the criterion shift assumption and differentiation); discussion of other possibilities is reserved for the General Discussion.

I thank Scott Brown, Simon Dennis, Andrew Heathcote, and Jeff Starns for excellent commentary and reviews. I am indebted to Joachim Vandekerckhove for quick and thorough responses to all my Diffusion Model Analysis Toolbox queries and for making this excellent program freely available. Finally, I thank Eric-Jan Wagenmakers for being a patient and generous sounding board as I wrapped my brain around the diffusion model.

Correspondence concerning this article should be address to Amy H. Criss, Department of Psychology, Syracuse University, 430 Huntington Hall, Syracuse, NY 13244-2340. E-mail: acriss@syr.edu

## The Criterion Shift Assumption

The criterion shift assumption is a metacognitive strategy wherein participants become aware that memory accuracy for a strong list is very high either during encoding or during the initial test trials (e.g., Hirshman, 1995; Stretch & Wixted, 1998; Verde & Rotello, 2007) and adopt a stringent criterion based on this knowledge, much like participants adopt a strict criterion when they are informed that the majority of test items will be foils (e.g., Rotello, Macmillan, Hicks, & Hautus, 2006). Use of a stricter criterion reduces the FAR. Thus, the reduction in FAR for a strong relative to a weak list is accounted for by a more lenient criterion for the weak list. This assumption is adopted by models from two classes: global matching models in which a stronger study list results in higher variability and thus a higher FAR following a strong than a weak list (see Shiffrin, Ratcliff, & Clark, 1990) and those models that assume the subjective strength of unrelated foils is not affected by strength of the encoded list (e.g., Cary & Reder, 2003; Stretch & Wixted, 1998; Verde & Rotello, 2007). The latter class of models is referred to as fixed-strength models (illustrated in Figure 1). Fixed-strength models follow from an early constraint on signal detection theory (e.g., Lockhart & Murdock, 1970; Parks, 1966) in which the foil distribution was assumed to be constant across all encoding conditions. This assumption was initially adopted as a simple convenience and was justified by attributing subjective memory strength to the preexperimental familiarity of the foil item. Thus, manipulations of the history of the foil (e.g., word frequency) but not encoding conditions (e.g., study time) were assumed to affect the subjective memory strength of foils. The fixed-strength assumption has persisted over decades and is no longer considered a mere convenience. In fact, this assumption has been incorporated in a number of single- and dual-process theories of memory.

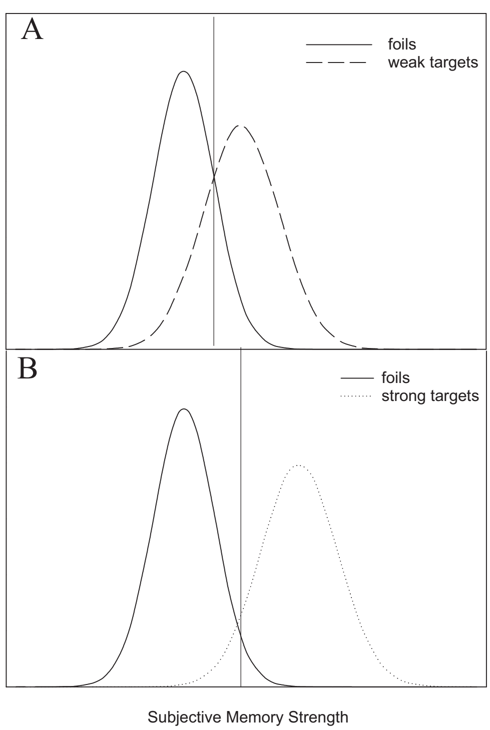


*Figure 1.* An illustration of the models used by Stretch and Wixted (1998) to account for the strength-based mirror effect in (A) a weak list and (B) a strong list. The mean of the target distribution is greater for a strong than a weak list. The mean of the foil distribution is constant for strong and weak lists (the fixed-strength assumption). The criterion (the vertical line) changes between the two lists, producing the strength-based mirror effect.

## Differentiation Models

Differentiation models were developed in part to account for the null list-strength effect and the WFME (Criss & McClelland, 2006; McClelland & Chappell, 1998; Shiffrin et al., 1990; Shiffrin & Steyvers, 1997). Differentiation models predict an SBME in the distributions of subjective memory strength, as shown in Figure 2, for the same reason they produce the null list-strength effect, though this went largely unnoticed in the SBME literature until recently. Two critical properties of the differentiation models underlie the SBME (Criss, 2006). The first is that repetition of an item results in updating the single memory trace for that item, resulting in a more accurate representation of the target with each encoding opportunity. The more accurate the memory trace, the better it matches its corresponding target presented during a memory test and the less well it matches any unrelated item presented during test. The second critical feature is that the models consider positive evidence when a feature matches and negative evidence when a feature mismatches. When traces stored in episodic memory are relatively complete (e.g., following a strong list), there is more negative evidence for an unrelated foil than when the traces are relatively incomplete (e.g., following a weak list). For these reasons, the distribution of subjective memory strength increases for targets and simultaneously decreases for foils following a strong list compared with a weak list. This prediction of the differentiation models follows directly from the encoding (i.e., memory traces are updated with repetition) and retrieval (i.e., the decision rule in which mismatching features decrease the subjective memory strength) assumptions. In fact, 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 (i.e., in this case the FARs for a strong list are higher than FARs for a weak list; see Criss, 2006, for relevant data and retrieving effectively from memory [REM] model simulations and Murnane & Shiffrin, 1991, for relevant data and a global matching model perspective). Differentiation models (like all models) require a criterion for endorsing a test item as "studied" or "not studied." There are many circumstances when the criterion changes (e.g., Criss, 2009; Xu & Malmberg, 2007). However, to account for the basic SBME, the placement of the criterion need not vary.

The differentiation models attribute the SBME to the encoding and decision processes inherent in the episodic memory system, whereas the other models under consideration attribute the SBME to a criterion shift based on metacognitive assessment. These two explanations are fundamentally different, at least in principle. In practice they have been difficult to discriminate with HRs and FARs. The goal of the current article was to take a more comprehensive look at data from the SBME paradigm by collecting both accuracy and reaction time (RT) measures. This full set of data is used to evaluate how additional encoding time influences the

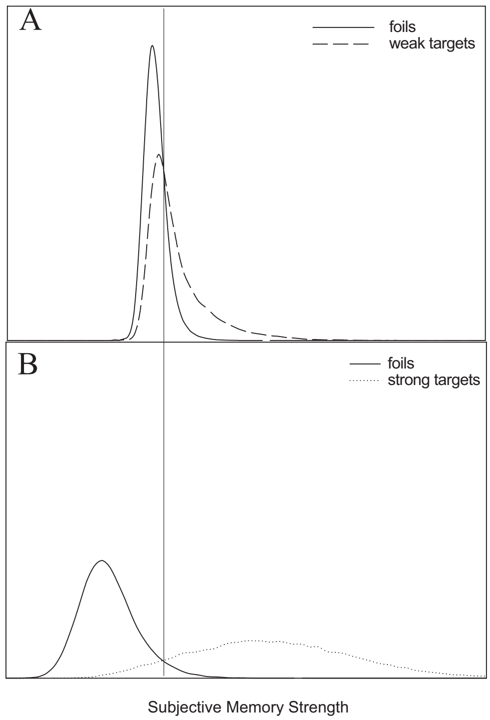


*Figure 2.* Simulated distributions of a differentiation model (retrieving effectively from memory; Shiffrin & Steyvers, 1997) for targets and foils following study of (A) a weak list and (B) a strong list. Differentiation increases the mean of the target distribution and decreases the mean of the foil distribution for strong lists relative to weak lists. The criterion need not change across lists to account for the strength-based mirror effect. 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.

distribution of subjective strength for foils and changes in criterion placement.

## The Ratcliff Diffusion Model (RDM)

To do this, the RDM shown in Figure 3 (Ratcliff, 1978; Ratcliff & McKoon, 2008) is employed. The RDM has been applied to a wide range of tasks in which one of two possible decisions is made in a relatively short period. In the model, presentation of a stimulus triggers the sequential sampling and accumulation of evidence until sufficient evidence is gathered to support one of two responses. The rate at which evidence accumulates is governed by the drift rate parameter, which is normally distributed across trials with a mean $v$ and variance $\eta$. Mean drift rate is determined by the quality of evidence provided by the stimulus. When the quality of evidence is high, the absolute value of the drift rate is high, and responses are both fast and accurate. When the quality of evidence is poor, the absolute value of the drift rate is small, and responses are both slower and less accurate. For example, in a recognition memory task the quality of evidence is determined by subjective memory strength (e.g., the match between the test word and episodic memory; Ratcliff, 1978; Ratcliff, Thapar, & McKoon, 2004). The separation between the response boundaries is governed by the $a$ parameter. For convenience the lower boundary ("not studied" in a recognition memory task) is set to zero, and the upper boundary is denoted $a$. Speed–accuracy trade-offs are modeled by changes in the boundary separation. Wider boundaries result in longer but more accurate decisions than narrow boundaries. For a fixed-boundary separation, the point at which the decision process begins to accumulate evidence is determined by the starting point parameter, which varies across trials according to a uniform distribution with mean $z$ and range $s_z$ such that $0 \leq z - s_z < z + s_z \leq a$. Prior to stimulus presentation, participants may be biased toward one of the response options based on incentives, instructions, or personal preference. This response bias is modeled by the mean starting point (Edwards, 1965; Voss, Rothermund, & Voss, 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008; but see Diederich & Busemeyer, 2006). Note that starting point does not affect the accumulation of evidence, and vice versa, much like response criterion and subjective memory strength are often assumed to be independent in signal detection theory. The time to encode the stimulus and execute a motor response, called nondecision time, is modeled by a uniform distribution with mean $T_{er}$ and range $s_t$.

One additional parameter, the drift criterion (also called the zero point of the drift rate), provides the criterion value of drift rate above which evidence accumulates toward the upper boundary and below which evidence accumulates toward the lower boundary (Ratcliff, 1978, 1981, 1985, 1987; Ratcliff, VanZandt, & McKoon, 1999). The effective rate of accumulation is the drift rate of the stimulus (i.e., memory strength) minus the drift criterion. This parameter is rarely discussed and in fact is not mentioned in the three published RDM toolboxes (Vandekerckhove & Tuerlinckx, 2007, 2008; Voss & Voss, 2007; Wagenmakers, van der Maas, & Grasman, 2007). There are at least three likely reasons that this parameter has been underinvestigated. First, it is almost always set equal to zero, and so positive drift rates accumulate toward one boundary, and negative drift rates accumulate toward the opposite boundary. Second, the drift criterion and drift rate mimic each other in predictions of both accuracy and RT distribution, and it is difficult to identify the two parameters separately (Ratcliff & McKoon, 2008). Third, it is not entirely clear what psychological or experimental factors should influence the drift criterion; it is unclear what factors produce a response bias that affects how information is accumulated (but for some evidence in nonmemory tasks, see Diederich & Busemeyer, 2006; Ratcliff, 1985; Ratcliff et al., 1999).

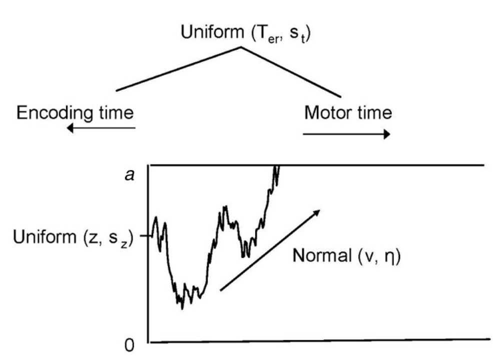


*Figure 3.* An illustration of Ratcliff's (1978) diffusion model.

## Response Bias

A common empirical method for eliciting response bias is to manipulate payoffs or the prior probability of a target. As just noted, there are three ways to characterize bias in the diffusion model—changes in the starting point ($z$), boundary separation ($a$), or drift criterion—and all have been informally described as analogous to the criterion in signal detection theory (e.g., Ratcliff & McKoon, 2008; Wagenmakers et al., 2008, 2007). Both starting point and drift criterion can produce similar changes in the probability of endorsing an item as studied—$P$("studied")—but they differ in how RT distributions are affected (see Ratcliff & McKoon, 2008, for a thorough discussion of this). An unbiased starting point is located at the midpoint of the boundary separation (e.g., $z = a/2$). If the starting point is biased such that it is closer to the "studied" boundary, then "studied" responses are more likely to occur and are faster, and "not studied" responses are less likely to occur and are slower, and this is true for both correct and incorrect decisions. Ratcliff and McKoon (2008) reported that changes in the starting point affect the tail and leading edge of the distribution by a 2:1 ratio with typical parameters; thus, the biased RT distribution appears shifted relative to the unbiased distribution. In contrast, typical changes in the drift rate (or the drift criterion) produce a 4:1 ratio for correct responses, spreading the tail of the biased distribution relative to the unbiased distribution.

Experiment 1 makes use of a classic manipulation of response bias, changing the probability of a target appearing on the test list, and asks whether this is better characterized by changes in the starting point, which is independent of the accumulation of evidence, or by changes in the drift criterion, which directly influences the manner in which the system processes evidence provided by the stimulus. Some have suggested that drift criterion may be adjusted based on payoffs or prior probabilities of the stimuli (Ratcliff, 1987, p. 279; Ratcliff et al., 1999, Experiment 1), whereas others have demonstrated that payoffs or priors alter starting point (Voss et al., 2004; Wagenmakers et al., 2008). Finally, Diederich (2008; Diederich & Busemeyer, 2006) suggests that both starting point and drift criterion are influenced by payoffs and stimulus probabilities depending on fluctuations in attention under time pressure. None of these experiments employ a recognition memory task.

Ratcliff and Smith (2004) manipulated the prior probabilities of targets in a recognition memory experiment with the goal of evaluating four sequential sampling models. Though they did not explicitly seek to evaluate a starting point versus drift criterion diffusion model (e.g., they let both freely vary for all conditions), they reported that a single drift criterion for all conditions provided a sufficient fit to the data. Thus, there is tentative support that a response bias manipulation in recognition memory is better characterized by changes in starting point. In Experiment 1, I sought to test whether changes in the starting point or drift criterion better account for change in behavior as a function of the relative proportion of targets on the test list.

## Experiment 1

Participants studied a list of items and then provided a binary decision for targets and foils in a recognition memory test that followed. The proportion of test items that were targets was manipulated. The literature suggests that HRs and FARs will increase in step with the proportion of targets on the test list (e.g., Criss, 2009; Rotello et al., 2006). RT distributions and response probabilities were used conjointly to discriminate between models in which response bias does and does not interact with evidence accumulation.

### Method

**Participants.** Nineteen members of the Syracuse University subject pool participated to fulfill course requirements.

**Stimulus materials.** The word pool consisted of 2,150 words between 4 and 11 letters in length and between 0.69 and 13.25 log frequency ($M = 8.46$) in the hyperspace analog to language corpus (Balota et al., 2002).[1]

**Design.** The experiment was divided into two sessions, each consisting of 10 study–test blocks. A brief break between each study–test block indicated that a new block was beginning. Each study list consisted of 50 unique words presented a single time for 1.5 s with a 750-ms interstimulus interval. Each test list consisted of 50 items with a 750-ms interstimulus interval between trials. The construction of the test list varied between the two sessions, with condition randomly assigned to session for each subject. During one session, 35 targets and 15 foils were tested on each list, labeled the "biased-to-say-studied" condition. During the other session, 35 foils and 15 targets were tested on each list, labeled the "biased-to-say-not-studied" condition. Participants were accurately informed about the nature of the test lists. During the memory test, participants were instructed to place the left index finger on the *C* key (labeled "no") and the right index finger on the *M* key (labeled "yes") and maintain that placement during the entire set of test trials. Participants were informed that their RT was being measured and were asked to respond as quickly as possible to the question "Was this word on the list you just studied?" without sacrificing accuracy. For each participant, words were randomly assigned to condition, and no item was presented on more than one list throughout the four sessions. The experiment was conducted with the Psychophysics Toolbox in MATLAB (Brainard, 1997).

### Results and Discussion

All participants completed all sessions for a total of 1,000 responses as follows: 350 target and 150 foil observations in the biased-to-say-studied condition and 150 target and 350 foil observations in the biased-to-say-not-studied condition. Two participants were eliminated for performing at chance (HR–FAR was less than .02), and one was eliminated for producing almost no FARs (less than .04 in each condition), which prevented model analysis. All analyses of variance (ANOVAs) are repeated measures unless otherwise noted.

**Accuracy.** As obvious in Figure 4, the bias manipulation was effective. A 2 (type of test item) × 2 (bias condition) ANOVA demonstrated main effects of item type and bias condition. Targets were called "studied" more often than foils, $F(1, 15) = 89.19$, $p <$

[1] For comparison, the words were between 11 and 49 ($M = 21.60$) per million in Kučera and Francis (1967).

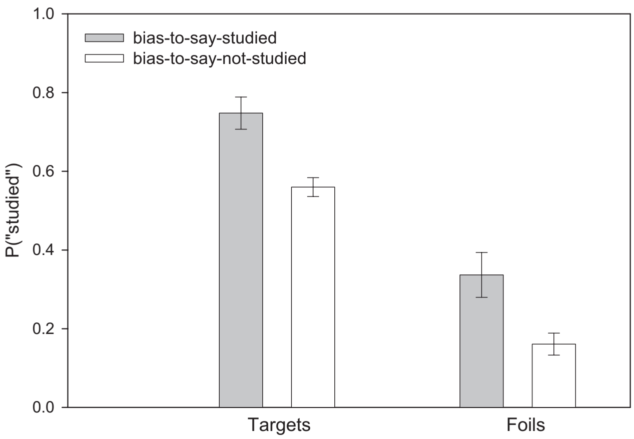


*Figure 4.* Accuracy data for Experiment 1. Error bars represent one standard error above and one below the mean.

.001, $MSE = .029$. Test items were called "studied" more often in the bias-to-say-studied condition than the bias-to-say-not-studied condition, $F(1, 15) = 23.41$, $p < .001$, $MSE = .023$. There was no interaction between the two variables, $F(1, 15) = 0.063$, $p = .805$, $MSE = .011$.

**RT.** In a quantile probability plot, RT quantiles are plotted as a function of the probability of a response, allowing the simultaneous observation of both speed and accuracy. Figure 5A shows such a plot for this experiment. The left panel displays foil trials, and the right panel displays target trials. Gray symbols refer to the bias-to-say-studied condition, and the white symbols refer to the bias-to-say-not-studied condition. Circles represent incorrect responses (FARs in the left panel and misses in the right panel), and triangles represent correct responses (correct rejections in the left panel and HRs in the right panel). The .1, .3, .5, .7, and .9 RT quantiles are plotted vertically for each condition (e.g., the middle point represents the median RT, the lowest point represents the RT below which only 10% of responses lie, and the uppermost point represents the RT above which 10% of responses fall) The location of each RT distribution along the *x*-axis is determined by the average probability correct for that condition.

The effect of bias condition on response probability is obvious in Figure 5A. In the left panel, gray circles are plotted to the right of white circles, indicating higher FARs for the bias-to-say-studied condition. Similarly, in the right panel, gray triangles are to the right of white triangles, indicating a higher HR in the bias-to-say-studied condition. The difference between the RT distributions for the two bias conditions appears to be a shift in the distributions, affecting both the fastest and slowest responses to targets and foils to a similar degree. Specifically, the more common response is faster for each condition, and the less common response is slower. The diffusion model takes into account changes in response probability and RT and provides explanations for these observed changes in terms of cognitive processes.

**Diffusion model analysis.** Participants occasionally respond very fast or very slow presumably because of lapses in attention or fast guesses. These outliers and contaminants distort calculations of the mean and standard deviation of RT (and distort the recovered parameters of the RDM) and should be eliminated from the data or explicitly modeled when possible (e.g., Ratcliff, 1993; Ratcliff & Tuerlinckx, 2002). Outliers were eliminated prior to analysis. Ninety-four responses (0.59%) exceeding 3 s and 996 responses (6.23%) faster than 300 ms were eliminated.[2]

The best fitting model was selected with Akaike information criterion with a finite sample correction (AICc) and Bayesian information criterion (BIC), which take into account not only the accuracy of the fit but also the number of free parameters used to obtain that fit (e.g., Akaike, 1978, 1979; Burnham & Anderson, 2002). AICc and BIC values for a set of models can be transformed into Akaike weights and BIC model weights, respectively, which can be interpreted as the probability that the model is the best model among the set under consideration (Wagenmakers & Farrell, 2004). These model selection techniques are useful but are limited in that they consider only number of parameters as a measure of complexity and neglect functional form (see Myung, 2000, for a review of this issue). Further, the penalty for number of parameters is harsher for BIC than AICc, which sometimes creates conflicting model selection results.

For each participant for each of the eight conditions (the factorial combination of bias condition, target or foil, and response), the .10, .30, .50, .70, and .90 RT quantiles were computed. Quantiles were averaged across participant, and the average data were fit by minimizing chi-square with the Diffusion Model Analysis Toolbox (Vandekerckhove & Tuerlinckx, 2007, 2008).[3]

Two models were fit to the data. In the *drift criterion model*, the drift rate was fit as follows: Target and foil drift rates were allowed to vary freely with the constraint that the difference between the target and foil drift rates was constant across condition (i.e., $v_{\text{target, bias-to-say-studied}} - v_{\text{foil, bias-to-say-studied}} = v_{\text{target, bias-to-say-not-studied}} - v_{\text{foil, bias-to-say-not-studied}}$). Drift criterion is the difference between the corresponding drift rates in each condition (e.g., $v_{\text{target, bias-to-say-studied}} - v_{\text{target, bias-to-say-not-studied}}$). All other parameters were held constant across condition, and the starting point was fixed at the midpoint of the response boundaries.[4] In the *starting point model* the drift rate was allowed to differ for targets and foils (and these values were identical for both bias conditions), and the starting point was allowed to vary across bias conditions. Table 1 reports AICc and the Akaike weights and BIC and BIC model weights (e.g., Akaike, 1978, 1979; Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004) for the two models, and both agree that the starting point model fits the group data better. The best fitting parameter values for the starting point model are reported in Table 2, and

[2] For both Experiments 1 and 2, several cut points for fast RTs were adopted including no cut point, 200 ms, and 300 ms. The cut point made no difference in the model selection results for 84% of the participants. When cut point did change model selection results, it was frequently due to inconclusive model selection with no cut point. To maintain consistency with the recognition memory RT literature (e.g., Ratcliff, Thapar, & McKoon, 2004), I report fits using 300 ms as the cut point.

[3] In addition, I conducted fits to quantile distributions using the multinomial likelihood loss function (S. Brown & Heathcote, 2003; Heathcote & Brown, 2004; Heathcote, Brown, & Mewhort, 2002). The parameter values were similar and the qualitative pattern of results for critical parameters were identical.

[4] A drift criterion model with starting point fixed across condition but not constrained to the midpoint did not fare any better.

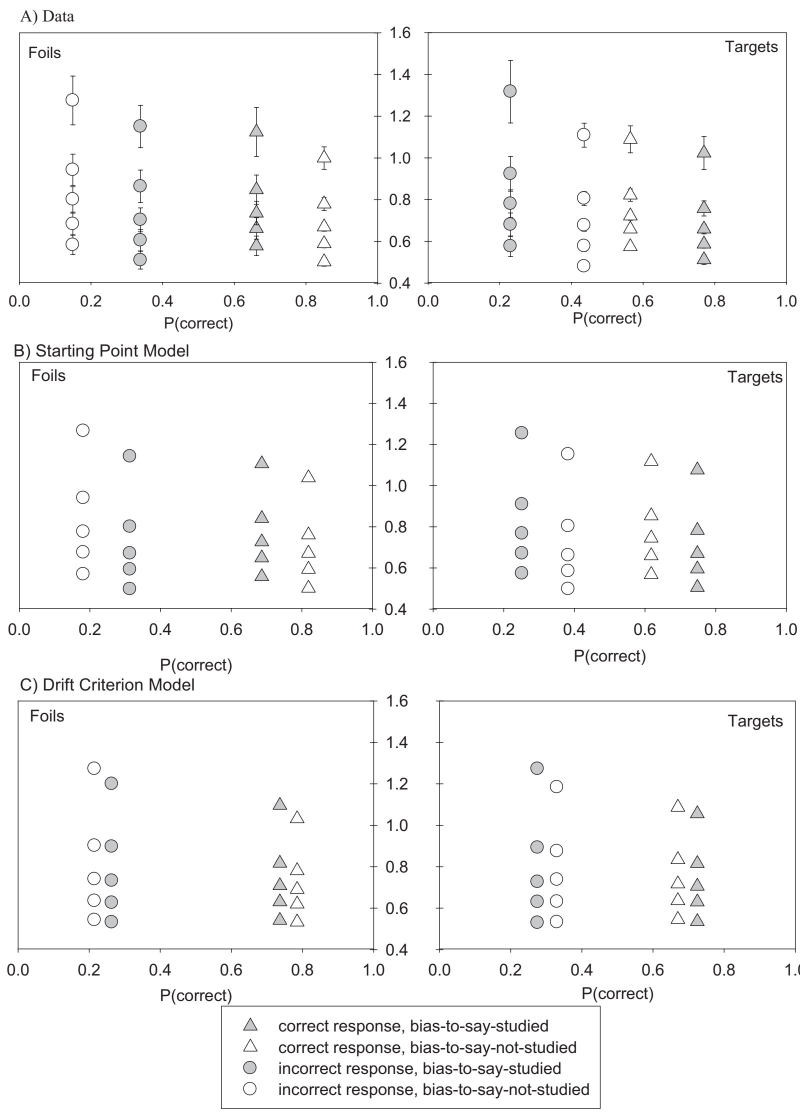


*Figure 5.* (A) The empirical reaction time distributions for Experiment 1, (B) predictions from a starting point model (with parameters from the left column of Table 3 excluding drift criterion, which was set to zero), and (C) predictions from a drift criterion model (with parameters from the group values in Table 3 excluding starting point, which was set to .06815).

they are as expected: The starting point for the bias-to-say-studied condition is higher (closer to the "studied" boundary) than the starting point for the bias-to-say-not-studied condition. The two models were also fit to individual participants. Overall, 13 participants were best fit by the starting point model, two were best fit by the drift criterion model, and one was inconclusive (AICc favored the starting point model, but BIC model weights were nearly equal, .53 for the drift criterion and .47 for the starting point model). The mean of the parameters for the starting point model across individual fits are similar to those obtained by fitting the group data as can be seen in Table 2.

Table 1
*Model Selection Values for Group Data in Experiment 1*

| Model | AICc | $w_i$(AICc) | BIC | $w_i$(BIC) |
|---|---|---|---|---|
| Starting point | 68232 | 1.00 | 68301 | 1.00 |
| Drift criterion | 68571 | 0.00 | 68632 | 0.00 |

*Note.* AICc = Akaike information criterion with a finite sample correction; BIC = Bayesian information criterion.

AICc and BIC indicate that the starting point model better captures the quantitative details of the data. Inspection of simulated data suggests that the drift criterion model fails to capture the differences between bias conditions in the leading edge of the RT distribution. Figure 5B shows simulated data generated by a starting point model, and Figure 5C shows simulated data generated by a drift criterion model. Five thousand simulated trials were run for

Table 2
*Best Fitting Parameters of the Starting Point Model for Data From Experiment 1*

| Parameter | Interpretation | Value (group data) | Mean value (individual fits) |
|---|---|---|---|
| $T_{er}$ | mean nondecision time | 0.5109 | 0.4988 |
| $s_t$ | range of nondecision time | 0.2166 | 0.2317 |
| $a$ | boundary separation | 0.1291 | 0.1368 |
| $s_z$ | range of starting point | 0.0811 | 0.0704 |
| $\eta$ | variance of the drift rate | 0.1593 | 0.1819 |
| $v_{\text{target}}$ | drift rate for targets | 0.1171 | 0.1372 |
| $v_{\text{foil}}$ | drift rate for foils | −0.1635 | −0.1847 |
| $z_{\text{BSS}}$ | starting point for bias-to-say-studied lists | 0.0787 | 0.0834 |
| $z_{\text{BSNS}}$ | starting point for bias-to-say-not-studied lists | 0.0492 | 0.0526 |

*Note.* BSS = bias-to-say-studied; BSNS = bias-to-say-not-studied.

each model. All parameters, excluding the drift criterion and starting point, were held constant for the two models at values reported in the left column of Table 3 (which correspond to the best fitting mixed model, allowing both drift criterion and starting point to vary across bias condition; see Mixed Model section). For the starting point model predictions, drift criterion was set to zero. For the drift criterion model predictions, starting point was set to the average (.06815). A biased starting point increases the probability of giving the response nearest the starting point and increases the speed of that response. At the same time, the competing response is less likely and slower. Changes in the starting point shift the RT distribution, and this is obvious at both the fastest and slowest quantiles (e.g., compare the adjacent white and gray symbols in Figure 5B). In contrast, changes in the functional drift rate via a nonzero drift criterion increase the speed and accuracy of correct responses by spreading the distribution rather than shifting the distribution. This is especially obvious in the tail of the distribution (i.e., slowest responses) where the favored response is faster than the nonfavored response. Differences in the leading edge (i.e., fastest responses) are negligible. The empirical data, shown in Figure 5A, show pronounced changes in both the tail of the distribution and the leading edge, consistent with a starting point model.

**Mixed model.** It is plausible that both drift criterion and starting point vary with the proportion of targets on the test list. As just illustrated, the starting point alone captures the qualitative pattern and quantitative details of the empirical data very well. Nevertheless, allowing the drift criterion to vary provides more flexibility to fit the specific quantitative details, and therefore a mixed model will always fit at least as well as or better than the simpler, nested model and is therefore of little use in model selection. For the group data, the mixed model is preferred over the starting point or drift criterion model (see Table 4). The starting point parameters in the mixed model follow the same pattern (closer to the "studied" boundary for the bias-to-say-studied condition) as the starting point model. The drift criterion parameter shifts the target drift rate higher and the foil drift rate lower in absolute value for the bias-to-say-studied condition relative to the other condition (see best fitting parameters in Table 3). In contrast to the group data, for the majority of participants, the starting point model is preferred ($N$ = 9). Only three individuals are best fit by a mixed model, another three are inconclusive, and one is best fit by a drift criterion model. The parameter values for the group fit and the average value over individual fits are very close with the exception of the drift criterion. The drift criterion for the group data is more than double the value obtained by averaging over individual fits, likely due to the needless use of the parameter (at least for nine of the participants).

Figure 6 plots the fits of both the starting point model (denoted with X) and the mixed model (denoted with +) for comparison. The starting point model provides a very good fit to the data, and most predicted values fall within two standard errors of the data. The improvement in fit provided by the mixed model is small in magnitude and can be characterized as a very slight exaggeration of $P$ ("studied") differences between bias conditions. For RT, the mixed model slightly counteracts the effect of starting point on error responses; that is, the favored response is slightly slower for errors in each condition under the mixed model, and the competing response is slightly faster (compared with the starting point model). In summary, support for the mixed model is inconsistent. Group data favor the mixed model, but individual participant data do not. The starting point must vary between bias conditions to account for the qualitative

Table 3
*Best Fitting Parameters of the Mixed Model to Data From Experiment 1*

| Parameter | Interpretation | Value (group data) | Mean value (individual fits) |
|---|---|---|---|
| $T_{er}$ | mean nondecision time | 0.5172 | 0.4990 |
| $s_t$ | range of nondecision time | 0.2192 | 0.2270 |
| $a$ | boundary separation | 0.1376 | 0.1406 |
| $s_z$ | range of starting point | 0.1000 | 0.0811 |
| $\eta$ | variance of the drift rate | 0.2125 | 0.1956 |
| $v_{\text{target}}$ | drift rate for targets in bias-to-say-not-studied lists | 0.1097 | 0.1402 |
| $v_{\text{foil}}$ | drift rate for foils bias-to-say-not-studied lists | −0.2161 | −0.1875 |
| Drift criterion | constant added to the target and foil drift rates for the bias-to-say-studied lists | 0.0509 | 0.0209 |
| $z_{\text{BSS}}$ | starting point for bias-to-say-studied lists | 0.0816 | 0.0832 |
| $z_{\text{BSNS}}$ | starting point for bias-to-say-not-studied lists | 0.0547 | 0.0557 |

*Note.* BSS = bias-to-say-studied; BSNS = bias-to-say-not-studied.

Table 4
*Model Selection Values Including a Mixed Model for Group Data in Experiment 1*

| Model | AICc | $w_i$(AICc) | BIC | $w_i$(BIC) |
|---|---|---|---|---|
| Mixed | 68199 | 1.00 | 68275 | 1.00 |
| Starting point | 68232 | 0.00 | 68301 | 0.00 |
| Drift criterion | 68571 | 0.00 | 68632 | 0.00 |

*Note.* AICc = Akaike information criterion with a finite sample correction; BIC = Bayesian information criterion.

pattern of data, especially changes in the leading edge of the RT distributions. Allowing the drift criterion to vary in addition to the starting point improves small quantitative details of the fit, at least to the group data.

The goal of this experiment was to determine whether manipulating the proportion of targets tested, a classic bias manipulation in recognition memory, influences the accumulation of evidence or the starting point of the decision process. The data offer strong support for the latter in the form of changes in the starting point, consistent with the findings of Ratcliff and Smith (2004). Including changes in drift criterion improves small quantitative details in the group data. In the next experiment, list strength is manipulated, and the question of interest is whether the SBME is due to changes in the starting point or evidence accumulation.

## Experiment 2

In Experiment 2, participants studied a list of items and then provided a binary decision for targets and foils in a recognition memory test that followed. Word frequency and list strength were orthogonally manipulated and should result in a simultaneous SBME and WFME for *P* ("studied"). Word frequency serves as a point of comparison with previous studies, which have demonstrated that the effect of word frequency is best described by differences in drift rate (e.g., Ratcliff et al., 2004). The critical data are the RT distributions for list strength and whether they are best described by changes in the drift rate or starting point in the RDM.

The memory models under consideration differ in criterion and/or the mean of the subjective memory strength, as seen in Figures 1 and 2. The differentiation models predict an increase in subjective memory strength for targets and a decrease in subjective memory strength for foils as a function of list strength with a fixed criterion across condition. Within the RDM, the mean of the memory strength distributions are interpreted as mean drift rates, and the fixed criterion is interpreted as a fixed starting point.

Interpreting the criterion shift models in the RDM is slightly more complicated, as there are at least two reasonable possibilities. One interpretation is that the mean of the memory strength distributions are analogous to the mean drift rates and the location of the criterion is analogous to the starting point. This interpretation is consistent with Experiment 1 and Ratcliff & Smith (2004), in which behavioral response bias maps onto starting point in the diffusion model. A second possibility is that criterion changes following a strength manipulation are qualitatively different from criterion changes following a manipulation of stimulus probabilities. Whereas changes in stimulus probability (i.e., such as in Experiment 1) are best described by changes in starting point, criterion changes following a strength manipulation are best interpreted as changes in the drift criterion. A change in the drift criterion between strong and weak lists means that the encoding conditions for the targets determine how evidence is accumulated during test.

Applying the RDM to data from the SBME paradigm allows for an unambiguous test between differentiation models and starting point version of the criterion shift model. However, the differentiation model and the drift criterion version of the criterion shift model cannot be discriminated because the drift rate and drift

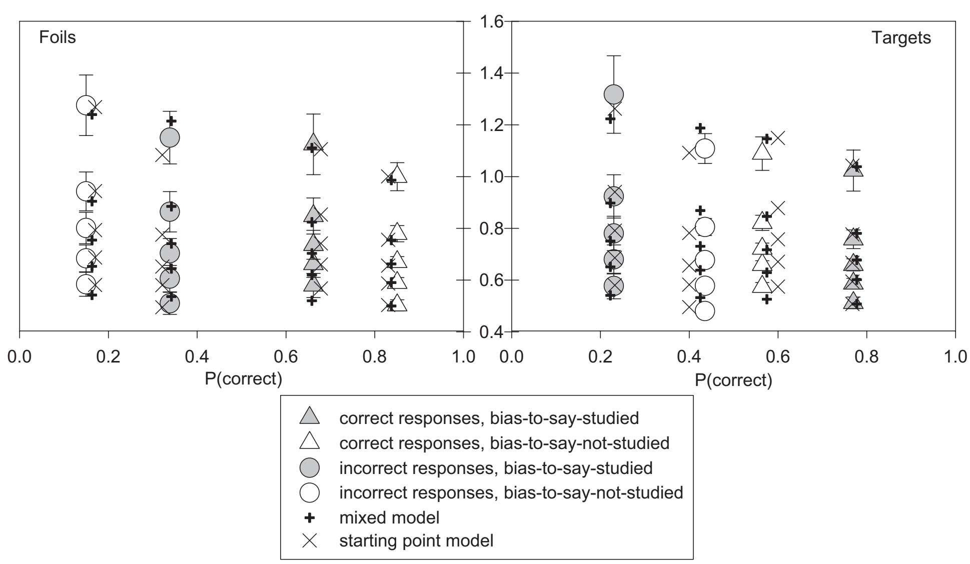


*Figure 6.* The empirical reaction time distributions for Experiment 1 and fits from the best fitting starting point (corresponding parameters in the left column of Table 2) and mixed model (corresponding parameters in the left column of Table 3). Error bars are standard errors of the mean.

criterion mimic each other in terms of both accuracy and RT predictions.

## Method

**Participants.** Sixteen members of the Syracuse University community participated for $10/hr.

**Stimulus materials.** The word pool consisted of 800 LF and 800 HF words between 4 and 11 letters in length. HF words ranged between 9 and 13 log frequency ($M = 10.46$) in the hyperspace analog to language corpus (Balota et al., 2002), and LF words ranged between 3.5 and 6 log frequency ($M = 5.22$).[5]

**Design.** The experiment was divided into four sessions. Each session consisted of four study–test blocks including one of each of the following lists: strong HF, strong LF, weak HF, and weak LF. The order of the study–test blocks was randomly assigned anew for each participant and each session. A brief break between each study–test block indicated that a new block was beginning. Each study list consisted of 50 unique words. For the weak lists, study words were presented a single time for 1.5 s with a 750-ms interstimulus interval. For the strong lists, study words were shown for five such presentations, and the entire set of 50 words was presented before any word repeated. Each test list consisted of all 50 targets and 50 foil items randomly intermixed. The test was self-paced, with a 750-ms blank screen separating each trial. During the memory test, participants were instructed to place the left index finger on the *C* key (labeled "no") and the right index finder on the *M* key (labeled "yes") and maintain that placement during the entire set of test trials. Participants were informed that their RT was being measured and were asked to respond as quickly as possible to the question "Was this word on the list you just studied?" without sacrificing accuracy. For each participant, words were randomly assigned to condition, and no item was presented on more than one list throughout the four sessions. The experiment was conducted with the Psychophysics Toolbox in MATLAB (Brainard, 1997).

## Results and Discussion

Thirteen participants completed all sessions for a total of 200 target and 200 foil observations in each condition. The other three participants participated in three sessions each (due to failure to show for the final session, $N = 1$, or loss of data from one session, $N = 2$). These three participants are included in the analysis presented below based on 150 observations per condition.

**Accuracy.** Separate 4 (session) × 2 (word frequency) × 2 (strength) ANOVAs showed no main effect of session and no interaction between session and any other variable for hits (all $F$s $< 1.8$ and $p$s $> .15$) or false alarms (all $F$s $< 1.9$ and $p$s $> .13$); thus data were collapsed over session. As expected, both a WFME and an SBME were observed (see Figure 7). Separate 2 (strength) × 2 (word frequency) ANOVAs were conducted for HRs and FARs. LF targets had higher HRs than HF targets, $F(1, 15) = 40.92$, $p < .001$, $MSE = .004$, and strong targets had higher HRs than weak targets, $F(1, 15) = 83.16$, $p < .001$, $MSE = .004$. LF foils had lower FARs than HF foils, $F(1, 15) = 18.40$, $p = .001$, $MSE = .003$, and foils following a strong list had lower FARs than foils following a weak list, $F(1, 15) = 19.94$, $p < .001$, $MSE = .003$. Word frequency and strength did not interact for foils, $F(1, 15) = 0.47$, $p = .501$, $MSE = .001$. The benefit from repetition was slightly larger for HF than LF targets; however, this effect was not significant, $F(1, 15) = 3.28$, $p = .09$, $MSE = .002$.

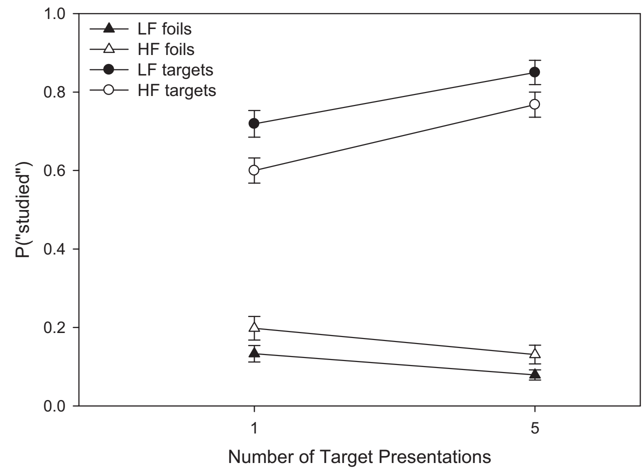


*Figure 7.* Accuracy data for Experiment 2 showing word frequency and strength-based mirror effects. Error bars represent one standard error above and one below the mean. LF = low frequency; HF = high frequency.

**RT.** Figure 8 shows RT distributions for this experiment. Figure 8A shows data for HF words, and Figure 8B shows data for LF words. The left panels display foil trials, and the right panels display target trials. Gray symbols refer to test items following a weak list, and white symbols refer to test items following a strong list. Triangles are correct response (correct rejections in the left panels and HRs in the right panels) and circles are incorrect responses (FARs in the left panels and misses in the right panels). The error bars represent one standard error above and one below the mean. The .1, .3, .5, .7, and .9 RT quantiles are plotted vertically for each condition.

The SBME is evident in the order of the four conditions. Strong foils are to the left of weak foils, and strong targets are to the right of weak targets, with the exact location determined by the empirical $P$("studied"). The WFME is evident by comparing Figures 8A and 8B. The LF observations in Figure 8B are more extreme along the $x$-axis—targets are higher and foils are lower—than the HF observations in Figure 8A.

The changes in RT as a function of word frequency and strength are very similar. Word frequency primarily serves to spread out the tail of the RT distributions; correct responses to LF words are faster than correct responses to HF words, especially in the slower responses for both targets and foils. Likewise, strength spreads the tail of the RT distributions; correct responses to words following a strong list are faster than words following a weak list for both targets and foils. The diffusion model takes into account both changes in response probability and RT as a function of condition as described next.

**Diffusion model analysis.** Empirical RT distributions were computed, and the diffusion model was fit with the Diffusion

[5] For comparison, HF words were less than 50 per million ($M = 130.66$) and LF words were between 2 and 10 per million ($M = 3.31$) in Kučera and Francis (1967).

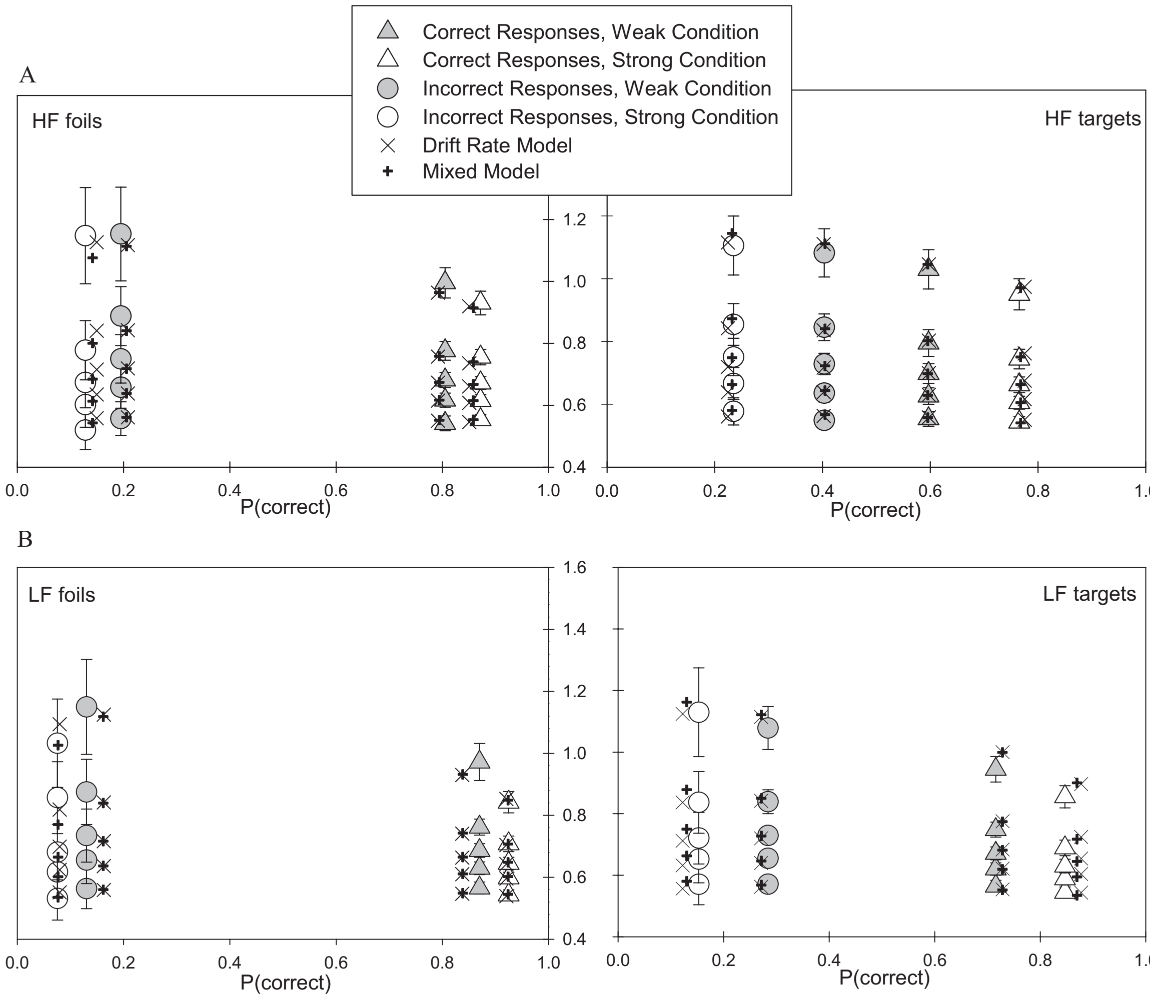


*Figure 8.* The empirical reaction time distributions for Experiment 2 for (A) high-frequency (HF) words and (B) low-frequency (LF) words. Parameters for the drift rate fits are in Table 6 (group values). Parameters for the mixed model fit are in Table 8 (group values).

Model Analysis Toolbox in MATLAB (Vandekerckhove & Tuerlinckx, 2007, 2008). Outliers were removed from each participant's data prior to further analysis. Twenty-six responses (0.11%) were slower than 3 s and 50 responses (0.20%) were faster than 300 ms, all of which were eliminated (following Ratcliff et al., 2004; see Footnote 2).

For each participant for each of the 16 conditions (the factorial combination of word frequency, list strength, target or foil, and response), the .10, .30, .50, .70, and .90 quantiles of the RT distribution were computed. The quantiles were averaged across participant, and the two models were fit to the group data.

In both models all parameters except the drift rate and starting point were held constant across condition. Word frequency is treated as a stimulus-based effect and attributed to changes in the drift rate, which is analogous to changes in the distribution of subjective memory strength. As discussed in the introduction, this is a common assumption shared by the majority of models and shared by the models considered in this article (cf. Benjamin, 2003; Hoshino, 1991). The difference between the models is the nature of the SBME. In the *drift rate model*, there were eight drift rate parameters, one for HF weak targets, HF strong targets, HF weak foils, HF strong foils, and their LF counterparts. In this model the starting point was not free to vary but rather fixed to be unbiased (e.g., the midpoint of the response boundaries; $z = a/2$). Note that as described in the introduction to this experiment, this drift rate model can be interpreted as a differentiation model or as a drift criterion version of a criterion shift model. The *starting point model* had six drift rate parameters, one for HF weak targets, HF strong targets, HF foils, and their LF counterparts. In addition, there were two starting point parameters, one for the weak lists and one for the strong lists. AICc and the Akaike weights and BIC and BIC model weights for each model are shown in Table 5 (e.g., Akaike, 1978, 1979; Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004), and both measures agree that the drift rate model provides the best fit. The models were also fit to individual participants. The drift rate model was favored by 11 participants (two provided weak or moderate evidence), and the starting point model was favored by five participants.

Table 5
*Model Selection Results for Fits to Group Data in Experiment 2*

| Model | AICc | $w_i$(AICc) | BIC | $w_i$(BIC) |
|---|---|---|---|---|
| Drift rate | 108604 | 1.00 | 108709 | 1.00 |
| Starting point | 108745 | 0.00 | 108851 | 0.00 |

*Note.* AICc = Akaike information criterion with a finite sample correction; BIC = Bayesian information criterion.

The best fitting parameters for the drift rate model are reported in Table 6. Note that parameters obtained by fitting the group data are very similar to those obtained by fitting each participant and averaging parameter value over participant. Positive drift rates tend to accumulate toward the "studied" boundary, and negative drift rates tend toward the "not studied" boundary. All target conditions have positive values, and all foil conditions have negative values, consistent with expectations. Stimuli that provide a higher quality of evidence or a more extreme subjective memory strength (e.g., LF targets and foils) have a higher absolute value of the drift rate than stimuli resulting in poorer quality of evidence or less extreme values of subjective memory strength (e.g., HF targets and foils, respectively), consistent with models that attribute the WFME to different distributions of memory strength (e.g., Criss & Malmberg, 2008; Dennis & Humphreys, 2001; Glanzer & Adams, 1990; McClelland & Chappell, 1998; Reder et al., 2000; Shiffrin & Steyvers, 1997; Stretch & Wixted, 1998). Drift rate parameters as a function of study repetition are consistent with the differentiation models (Criss, 2006; Criss & McClelland, 2006; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Encoding conditions that lead to more evidence also have a higher drift rate (strong relative to weak targets). Critically, the absolute value of the drift rate for foils following a strong list is greater in magnitude than the value for foils following a weak list. The RT data and modeling disconfirm the starting point version of the criterion shift model but not the drift criterion version of the criterion shift model. It is mathematically impossible to discriminate between changes in the drift criterion parameter and changes in the foil drift rate parameter in this paradigm. Thus, either the differentiation models provide a more accurate description of episodic memory or there are multiple types of criterion shifts that have been lumped into a single class and modeled with a single parameter within memory models (or potentially both).

It is informative to note that the starting point model is inaccurate at both a qualitative and a quantitative level of analysis. In this model, drift rates for strong and weak foils are identical, which predicts approximately equal RTs and approximately equal FARs for the strong and weak foils, which is inconsistent with the empirical pattern of data. A starting point biased toward the "not studied" boundary for strong relative to weak lists would partially correct this by decreasing the FAR for the strong list. However, this also predicts faster "not studied" responses for strong than weak foils and slower "studied" responses for strong than weak foils. Further changes in starting point shift more than spread the distributions (as demonstrated in Figure 5). All these qualitative descriptions are inconsistent with the pattern of data seen in Figure 8.

**Mixed model.** As in Experiment 1, it is plausible that a mixed model allowing both drift rate and starting point to change with list strength may provide additional flexibility to predict the subtle quantitative details of the data beyond that provided by a drift rate model alone. However, the mixed model will always fit at least as well as or better than the simpler, nested model and is therefore of little use in model selection. Indeed, a mixed model with eight drift rate parameters (HF weak targets, HF strong targets, HF weak foils, HF strong foils, and their LF counterparts) and two starting point parameters (weak and strong lists) fit group data better than the original starting point or drift rate model (see Table 7). As shown in Table 8, the drift rate parameters for the mixed model follow the same pattern as for the drift rate model: LF and strong conditions have more extreme drift rates than HF and weak conditions for both targets and foils. Surprisingly, the starting point values are in the opposite direction of what would be predicted by a criterion shift memory model: The starting point is closer to the "studied" boundary for the strong list compared with the weak list. This may be a result of serial correlations (Laming, 1968; Treis-

Table 6
*Best Fitting Parameters of the Drift Rate Model for Experiment 2*

| Parameter | Interpretation | Value (group data) | Mean value (individual fits) |
|---|---|---|---|
| $T_{er}$ | mean nondecision time | 0.5175 | 0.5103 |
| $s_t$ | range of nondecision time | 0.1530 | 0.1866 |
| $a$ | boundary separation | 0.1221 | 0.1294 |
| $s_z$ | range of starting point | 0.0641 | 0.0421 |
| $\eta$ | variance of the drift rate | 0.2188 | 0.2283 |
| $z$ | starting point (constrained to equal $a$/2) | 0.0611 | 0.0647 |
| $v_{\mathrm{HFst}}$ | drift rate for HF strong targets | 0.2070 | 0.2306 |
| $v_{\mathrm{HFwt}}$ | drift rate for HF weak targets | 0.0685 | 0.0678 |
| $v_{\mathrm{HFsf}}$ | drift rate for HF strong foils | −0.2872 | −0.3194 |
| $v_{\mathrm{HFwf}}$ | drift rate for HF weak foils | −0.2228 | −0.2310 |
| $v_{\mathrm{LFst}}$ | drift rate for LF strong targets | 0.3244 | 0.3820 |
| $v_{\mathrm{LFwt}}$ | drift rate for LF weak targets | 0.1663 | 0.1833 |
| $v_{\mathrm{LFsf}}$ | drift rate for LF strong foils | −0.3917 | −0.4566 |
| $v_{\mathrm{LFwf}}$ | drift rate for LF weak foils | −0.2713 | −0.3094 |

*Note.* HF = high frequency; st = strong target; wt = weak target; sf = strong foil; wf = weak foil; LF = low frequency.

Table 7
*Model Selection Results Including the Mixed Model for Group Data in Experiment 2*

| Model | AICc | $w_i$(AICc) | BIC | $w_i$(BIC) |
|---|---|---|---|---|
| Mixed | 108469 | 1.00 | 108590 | 1.00 |
| Drift rate | 108604 | 0.00 | 108709 | 0.00 |
| Starting point | 108745 | 0.00 | 108851 | 0.00 |

*Note.* AICc = Akaike information criterion with a finite sample correction; BIC = Bayesian information criterion.

man & Williams, 1984; Verplanck, Collier, & Cotton, 1952). On average, a strong list produces more "studied" responses than a weak list (46% vs. 41%).

Fits of the mixed model to individual participants proved uncertain. An equal number of participants ($N = 5$) are best fit by the mixed model and the drift rate model. The starting point model is favored by two participants, and model selection was inconclusive for another four participants (BIC favored one model, AICc favored a different model, or the model weights were equally split among two models).

Figure 8 plots the data and fits of both the drift rate model (denoted by X) and the mixed model (denoted by +) for comparison. The drift rate model captures the pattern of data quite well, with all predicted values within or very near two standard errors of the mean value. The improvement in fit provided by the mixed model is a very slight slowing of incorrect "new" and speeding of incorrect "old" responses for the strong conditions. As was true for Experiment 1, support for the mixed model is inconsistent. Group data favor the mixed model; individual participants were split between the models. Drift rates must vary between strong and weak targets and strong and weak foils to account for the qualitative pattern of data—large changes in the tail of the distributions and mostly absent changes in the leading edge of the RT distributions. Allowing the starting point to vary as a function of list strength, in addition to drift rate, provides a small improvement in the quantitative details of the fit.

The goal of this experiment was to assess whether list strength alters decision criterion (e.g., starting point) or evidence accumulation (e.g., changes in the drift rate) for foils. The data offer strong support for the latter. Foils following a strong list have larger values of drift rates than foils following a weak list. Allowing the starting point to change with list strength improves small details of the fit for the mixed model. Notably, the best fitting change in starting point is in the opposite direction of that predicted by criterion change memory models: Participants were more biased toward the "studied" response boundary following a strong than a weak list. These data disconfirm the starting point version of the criterion shift model but are consistent with either the differentiation model or the drift criterion version of the criterion shift model.

## General Discussion

This article makes use of manipulations in recognition memory, stimulus probability, and list strength, for which accuracy alone is insufficient to discriminate between models. The goal of this article was to use the diffusion model as a measurement tool to provide converging evidence in the debate in the memory modeling literature between differentiation and criterion change accounts of encoding strength. In one experiment, stimulus probability was manipulated. "Studied" responses are more likely when targets are more probable at test, and "not studied" responses are more likely when foils are more plentiful at test. RT distributions shift such that the more likely response is faster and the less likely response is slower in both conditions. This pattern of data requires a model with differences in starting point, specifically with starting points closer to the more likely response. In a second experiment, list strength and word frequency were manipulated, and both LF and strong lists have higher HRs and lower FARs than HF and weak lists, respectively. RT distributions shift such that the tail of the distribution is faster for the more accurate conditions but the

Table 8
*Best Fitting Parameters of the Mixed Model to Data From Experiment 2*

| Parameter | Interpretation | Value (group data) | Mean value (individual fits) |
|---|---|---|---|
| $T_{er}$ | mean nondecision time | 0.5131 | 0.5105 |
| $s_t$ | range of nondecision time | 0.1528 | 0.1812 |
| $a$ | boundary separation | 0.1208 | 0.1306 |
| $s_z$ | range of starting point | 0.0530 | 0.0416 |
| $\eta$ | variance of the drift rate | 0.2088 | 0.2367 |
| $z_s$ | starting point for strong list | 0.0670 | 0.0686 |
| $z_w$ | starting point for weak list | 0.0615 | 0.0672 |
| $v_{\mathrm{HFst}}$ | drift rate for HF strong targets | 0.1675 | 0.2043 |
| $v_{\mathrm{HFwt}}$ | drift rate for HF weak targets | 0.0600 | 0.0576 |
| $v_{\mathrm{HFsf}}$ | drift rate for HF strong foils | −0.3126 | −0.3452 |
| $v_{\mathrm{HFwf}}$ | drift rate for HF weak foils | −0.2198 | −0.2436 |
| $v_{\mathrm{LFst}}$ | drift rate for LF strong targets | 0.2726 | 0.3568 |
| $v_{\mathrm{LFwt}}$ | drift rate for LF weak targets | 0.1549 | 0.1746 |
| $v_{\mathrm{LFsf}}$ | drift rate for LF strong foils | −0.4127 | −0.4906 |
| $v_{\mathrm{LFwf}}$ | drift rate for LF weak foils | −0.2650 | −0.3156 |

*Note.* HF = high frequency; st = strong target; wt = weak target; sf = strong foil; wf = weak foil; LF = low frequency.

leading edge is relatively stable. The RT data confirm prior findings of a larger drift rate for LF targets and foils than their HF counterparts (e.g., Ratcliff et al., 2004). This pattern of data requires differences in the functional drift rate for strong and weak targets and for strong and weak foils, with larger values of drift rate following a strongly encoded list. The drift rate changes can be attributed to changes in memory strength and/or changes in drift criterion; the SBME paradigm does not allow for discrimination of these two possibilities.

In both experiments the data were best fit by a mixed model in which starting point and drift rate change. In both cases the mixed model best fit the group data but received limited support from fits to individuals. Further, the gain from the additional parameter (drift criterion in Experiment 1 and starting point in Experiment 2) is not necessary for a very good fit that captures all qualitative patterns in the data; rather the additional parameter provided by the mixed model improves the intricate details of the fit. Whether the inclusion of an additional parameter in the mixed models provides a meaningful explanation for the underlying cognitive processes or merely fits noise in the data is unresolved. This remains a question for future research.

Two classes of memory models are being considered: differentiation and criterion shift models. First, consider Experiment 1, in which stimulus probabilities were manipulated. Both classes of models agree that the increase in HRs and FARs when 70% of the test items are targets is a criterion shift that does not interact with the evidence provided by the stimulus. This type of criterion shift represents bias in the decision process and is commonly attributed to the location of the criterion in signal-detection-like memory models (e.g., Rotello et al., 2006). Now consider Experiment 2 in which list strength and word frequency were manipulated. LF words are better remembered than HF words. One explanation comes from the REM model (Shiffrin & Steyvers, 1997). In REM, LF words tend to be composed of more distinctive features, and HF words tend to be composed of more common features. HF words tend to match many different words by chance because of these shared common features, accounting for the higher HF FAR. The amount of positive evidence provided by a matching feature is a function of the distinctiveness or diagnosticity of that feature, predicting a higher LF HR. Higher quality of evidence translates to drift rate in the RDM. How, then, might the differences in drift rate for strong and weak lists be interpreted?

In differentiation models, the encoding and retrieval processes predict an increase in subjective memory strength for targets and a decrease for foils as a function of list strength. This is consistent with interpreting changes in drift rate in Experiment 2 as differences in the quality of information provided by the stimulus, just as is assumed for the word frequency manipulation. The ordering of subjective memory strength predicted by default parameters of the differentiation models matches the ordering of drift rates generated by the diffusion model.

In criterion shift models, the subjective memory strength for targets increases as a function of list strength, corresponding to a higher drift rate for strong targets. However, the subjective memory strength for foils remains constant or increases as a function of list strength. Under this framework, the foil drift rate does not change as a function of strength; rather, the drift criterion changes. A change in the drift criterion between strong and weak lists means that the encoding conditions determine how evidence is accumulated. This type of criterion shift that affects the processing of the stimulus might be called evidentiary or perceptual bias (cf. Voss, Rothermund, & Brandtstädter, 2008). Thus, to account for the full set of data presented here, criterion shift models need two types of criteria changes: one that influences how evidence from the stimulus is treated, called evidentiary or perceptual bias, and one that operates at the level of the decision and is independent of evidence accumulation, called decision bias.

Unfortunately, modern recognition memory models do not account for RT distributions, and modification to the models would be required in order to do so (but see Diller, Nobel, & Shiffrin, 2001). Presumably, one could combine a differentiation memory model with a diffusion decision model by taking values of subjective strength from differentiation models and feeding them as drift rates into a diffusion process. The current data would seem to follow directly from such a model. Existing fixed-strength criterion shift models may have conceptual difficulties, for they would need to incorporate two types of criterion shift, whereas there is currently only one. In signal detection models there exists a single type of criterion shift, typically interpreted as decision bias (e.g., Green & Swets, 1966). In contrast, much of the RDM literature equates the drift criterion, which represents evidentiary or perceptual bias, with the criterion in signal detection theory (cf. Ratcliff & McKoon, 2008). An alternative interpretation is that the drift criterion serves as a likelihood transformation.[6] One avenue for further research is to explore likelihood-based models as one type of change in response probabilities based on expectancy (e.g., evidentiary or perceptual bias) along with a criterion for responding "studied" or "not studied" based on the memory strength resulting from that likelihood computation (e.g., decision bias).

Changes in decision bias and evidentiary or perceptual bias have similar (or identical) outcomes in accuracy but substantially different RT distribution profiles (and are therefore governed by different parameters in the RDM). Ignoring RT in the memory literature has led to the current unviable situation in which different types of shifts in criteria are treated as identical and accounted for by the same parameter (e.g., the criterion in signal detection). Further, empirical manipulations leading to the different types of criterion shifts are not obvious, and nothing inherent in extant models provides insight into this question. One solution to this problem is to develop models that describe the processes underlying episodic memory. Currently, there are a number of measurement models (e.g., various forms of signal detection theory) that are treated as process models. The differentiation models suffer none of these consequences, as they attribute changes in drift rate to the quality of evidence provided by the stimulus rather than bias and changes in decision bias to the location of the criterion.

## Evidence for Differentiation

Evidence for the differentiation of perceptual and semantic knowledge and the developmental course of differentiation of broad categories first, followed by more subtle distinctions, is plentiful (cf. E. J. Gibson, 1940, 1969; J. J. Gibson & Gibson, 1955). Loss of semantic knowledge follows the reverse pattern, with fine distinctions between similar items lost before the dis-

[6] I thank Andrew Heathcote for this suggestion.

tinction between more general categories (e.g., McClelland & Rogers, 2003; Rogers & McClelland, 2004). Evidence for differentiation in episodic memory is growing. Criss (2006) reported simulations from differentiation models predicting that a similar foil matches a strong memory trace better than a weak memory trace (opposite to the pattern for an unrelated foil). These simulations considered the case in which a foil is similar to one strong or one weak target and the predictions were confirmed empirically. Criss (2009) used the direct ratings method of Mickes, Wixted, and Wais (2007) to collect distributions of subjective memory strength for targets and foils following strong and weak lists. The data supported a priori predictions of differentiation models, namely, that the distribution for foils following a strong and a weak list differ, as do distributions for strong and weak targets. In a second experiment, Criss manipulated response bias (using the same stimulus probability manipulation used in Experiment 1) and found no change in the distributions of memory strength despite finding the expected changes in HR and FAR. 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." All the episodic memory data just reviewed and the data reported here are predicted by differentiation models. Criterion shift memory models can also explain such data, but doing so requires a number of post hoc assumptions about the presence (and absence) of shifts in criteria for each experimental condition.

## Likelihood Ratio Models

In their seminal work on mirror effects, Glanzer and Adams (1985, 1990) demonstrated mirror effects for a large number of experimental manipulations that resulted in one class of items that were better remembered than another class of items. They developed the attention likelihood model, a fully informed likelihood ratio model in the original form, in which the decision about whether 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. In the original SBME article, 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). If the decision rule took into account the information about the test item, then a within-list mirror effect should have been observed. Fully informed likelihood ratio models have also been disconfirmed in other paradigms (e.g., Balakrishnan & Ratcliff, 1996; Hintzman, 1994). On this basis, Stretch and Wixted ruled out fully informed likelihood ratio models as an explanation for this class of mirror effects, and such models have largely been dismissed in the interim.

However, recent experiments are consistent with partially informed likelihood ratio models that take into account some (but not all) properties of the experimental situation. For example, a within-list mirror effect is observed for items studied immediately prior to the list and those studied in the more distant past (Singer, Gagnon, & Richards, 2002; Singer & Wixted, 2006). Starns (2009) recently applied one such model, the bind cue decide model of episodic memory (Dennis & Humphreys, 2001; cf. Criss & Shiffrin, 2004), to the SBME by adjusting the amount of evidence required for a "studied" response based on the expected difficulty of the test list (which was provided by the experimenter). Memory models with likelihood ratio decision rules (which include the differentiation models) are consistent with the data presented here and are a promising avenue for further theoretical development.

## Conclusions

A number of other studies have investigated the nature of criterion changes within a single test list (e.g., S. Brown & Steyvers, 2005; Hockley & Niewiadomski, 2007; Singer et al., 2002; Singer & Wixted, 2006) and whether criteria are set based on the perceived difficulty of the study list or the experienced difficulty of the test items (e.g., Benjamin & Bawa, 2004; J. Brown, Lewis, & Monk, 1977; Hirshman, 1995; Verde & Rotello, 2007). The majority of these experiments assumed a fixed-strength model and did not consider differentiation models. Indeed, prior to Criss (2006), differentiation was not recognized as an explanation for the SBME. Since then a number of studies have reported evidence consistent with a priori predictions of differentiation models (e.g., Criss, 2009, and the current data; Hockley & Niewiadomski, 2007). At this point, the literature does not unambiguously favor a criterion shift or a differentiation account, and this domain is fruitful ground for future research and theory development.

Surprisingly few studies of recognition memory report reaction item, and even fewer evaluate RT distributions. This article illustrates a case in which identical patterns of accuracy are accompanied by qualitatively different patterns of RT distributions. Analysis of accuracy in the absence of RT fails to capture these differences in performance. Developing diffusion models based on the assumptions of memory models proved to be a useful tool for evaluating competing memory models.

## References


Akaike, H. (1978). On the likelihood of a time series model. *Statistician, 27,* 217–235.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika, 66,* 237–242.

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., & Treiman, R. (2002). *The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords.* Retrieved from http://elexicon.wustl.edu

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31,* 297–305.

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

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

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., & Heathcote, A. (2003). QMLE: Fast, robust, and efficient estimation of distribution functions based on quantiles. *Behavior Research Methods, Instruments & Computers, 35,* 485–492.

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer-Verlag.

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

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. (2009). The distribution of subjective memory strength: Foils and response bias. *Cognitive Psychology, 59,* 297–319.

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

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478.

Diederich, A. (2008). A further test of sequential-sampling models that account for payoff effects on response bias in perceptual decision tasks. *Perception & Psychophysics, 70*(2), 229–256.

Diederich, A., & Busemeyer, J. R. (2006). Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis. *Perception & Psychophysics, 68*(2), 194–207.

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC–REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 414–435.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology, 2*(2), 312–329.

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

Gibson, E. J. (1969). *Principles of perceptual learning and development.* New York, NY: Appleton-Century-Crofts.

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

Glanc, G. A., & Greene, R. L. (2007). Orthographic neighborhood size effects in recognition memory. *Memory & Cognition, 35,* 365–371.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & 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.

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

Heathcote, A., & Brown, S. (2004). Reply to Speckman and Rouder: A theoretical basis for QML. *Psychonomic Bulletin & Review, 11,* 577–578.

Heathcote, A., Brown, S., & Mewhort, D. J. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review, 9,* 394–401.

Heathcote, A., Ditton, E., & Mitchell, K. (2006). Word frequency and word likeness mirror effects in episodic recognition memory. *Memory & Cognition, 34,* 826–838.

Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(1), 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 & Cognition, 35,* 679–688.

Hoshino, Y. (1991). A bias in favor of the positive response to high-frequency words in recognition memory. *Memory & Cognition, 19,* 607–616.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Laming, D. R. J. (1968). *Information theory of choice-reaction times.* Oxford, England: Academic Press.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74,* 100–109.

Malmberg, K. J., & Nelson, T. O. (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. *Memory & Cognition, 31,* 35–43.

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.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4,* 310–322.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14*(5), 858–865.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(5), 855–874.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204.

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. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88,* 552–572.

Ratcliff, R. (1985). Theoretical interpretations of speed and accuracy of positive and negative responses. *Psychological Review, 92,* 212–225.

Ratcliff, R. (1987). More on the speed and accuracy of positive and negative responses. *Psychological Review, 94,* 277–280.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

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,* 163–178.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873–922.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333–367.

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.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438–481.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261–300.

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.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember–know judgments using signal detection and threshold models. *Memory & Cognition, 34,* 1598–1614.

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

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

Singer, M., Gagnon, N., & Richards, E. (2002). Strategies of text retrieval: A criterion shift account. *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 & Cognition, 34,* 125–137.

Starns, J. (2009, January). *Enhancing lure rejection by strengthening studied items: Contrasting encoding- and retrieval-based mechanisms from REM and BCDMEM.* Paper presented at the Context and 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.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91*(1), 68–111.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14,* 1011–1026.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40,* 61–72.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35*(2), 254–262.

Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology, 44*(4), 273–282.

Voss, A., Rothermund, K., & Brandtstädter, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology, 44*(4), 1048–1056.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32,* 1206–1220.

Voss, A., & Voss, J. (2007). Fast-Dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39,* 767–775.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192–196.

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.

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14,* 3–22.

Xu, J., & Malmberg, K. J. (2007). Modeling the effects of verbal- and nonverbal-pair strength on associative recognition. *Memory & Cognition, 35*(3), 526–544.

Received March 3, 2009
Revision received November 16, 2009
Accepted November 23, 2009 ■