Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 55 (2006) 447–460
Journal of Memory and Language
www.elsevier.com/locate/jml

Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM) ☆

Amy H. Criss *, James L. McClelland

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

Received 29 January 2006; revision received 9 June 2006
Available online 2 August 2006

Abstract

The subjective likelihood model [SLiM; 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.] and the retrieving effectively from memory model [REM; Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin* & *Review*, 4, 145–166.] are often considered indistinguishable models. Indeed both share core assumptions including a Bayesian decision process and differentiation during encoding. We give a brief tutorial on each model and conduct simulations showing cases where they diverge. The first two simulations show that for foils that are similar to a studied item, REM predicts higher false alarms rates than SLiM. Thus REM is not able to account for certain associative recognition data without using emergent features to represent pairs. Without this assumption, rearranged pairs have too strong an effect. In contrast, this assumption is not required by SLiM. The third simulation shows that SLiM predicts a reversal in the low frequency hit rate advantage as a function of study time. This prediction is tested and confirmed in an experiment.

Keywords: Episodic memory; Recognition memory; Memory models; Associative recognition; Word frequency; Model comparison; Differentiation; Likelihood models

The goal of this paper is to compare and contrast two mathematical models of episodic memory, the Subjective Likelihood Model (SLiM; McClelland & Chappell, 1998) and the Retrieving Effectively from Memory model (REM; Shiffrin & Steyvers, 1997). Why compare REM and SLiM? These models share the core assumptions of a Bayesian decision process and the concept of differentiation (each will be discussed in greater detail later). They were initially developed to account for overlapping sets of empirical findings and were published around the same time. The combination of these factors resulted in the field essentially perceiving the two models as twins. This is evident in reading the literature as the two are almost universally cited together and assumed

☆ We thank E.J. Wagenmakers for helpful comments on earlier version of this manuscript. This research was supported by National Institutes of Mental Health Grant MH019983 to A.H.C. and MH64445 to J.L.M.
* Corresponding author.
E-mail addresses: acriss@andrew.cmu.edu (A.H. Criss), jlm@cnbc.cmu.edu (J.L. McClelland).

0749-596X/$ - see front matter
doi:10.1016/j.jml.2006.06.003

to make identical predictions. This belief perhaps grew out of the same impression by the authors of the models as obvious in the following quotes: "... it is striking how many similarities exist between the two models [REM and SLiM]" (McClelland & Chappell, p. 753). "It is a curious fact that this result makes the recognition models of McClelland and Chappell (1998) and Shiffrin and Steyvers (1997) extremely similar in both structure and parameterization, to an even greater degree than these set of authors may have appreciated heretofore." (Shiffrin & Steyvers, 1998, p. 90). In this paper, we evaluate the degree to which these models actually share assumptions and make identical qualitative predictions. To foreshadow, there are several important differences between the models: They provide different explanations for the same phenomenon and they make different qualitative predictions for the same experimental manipulations. Before comparing these models, we first take a brief look at the history of memory modeling that led to the development of both SLiM and REM.

All of the models we discuss, including SLiM and REM, are primarily concerned with the task of episodic recognition memory. Typically, this is investigated in a list memory paradigm where a list of items is studied followed by a short break typically including some irrelevant task (e.g., adding a list of digits). After the break a recognition memory test is administered where individual items are presented and the participant is simply asked to say whether or not each item was presented on the list. The dependent measures are the hit rate (HR) defined as probability that a studied item is correctly called "studied" and the false alarm rate (FAR) defined as the probability that an unstudied item is incorrectly called "studied." Early theorizing about episodic memory claimed that the decision about whether or not to call a test item "studied" was based on the overall strength or familiarity of the item (Flexser & Bower, 1974; Gillund & Shiffrin, 1984; Hintzman, 1988; Humphreys, Bain, & Pike, 1989; Humphreys, Pike, Bain, & Tehan, 1989; Murdock, 1982). Many early theories were in the tradition of signal detection theory and thus conceived of the comparison between the test probe and the contents of episodic memory as resulting in a single value often referred to as strength or familiarity. This subjective feeling was thought to arise from one of two normally distributed variables. The mean familiarity of the studied or target distribution was greater than the mean familiarity of the unstudied or foil distribution, allowing performance to rise above chance. If that subjective familiarity or strength value exceeded some criterion then the item was considered a member of the study list, otherwise it was considered new. This approach was fruitful for a number of years and is still a popular approach to modeling episodic memory (e.g., Banks, 2000; DeCarlo, 2002; Dunn, 2004; Rotello, Macmillan, & Reeder, 2004; Wixted & Stretch, 2004). There are numerous challenges to this approach including two empirical findings—the mirror effect and the null list strength effect that inspired the development of models such as SLiM and REM. Next we briefly consider why the mirror effect and the null list strength effect in recognition memory challenged a number of strength models and how REM and SLiM overcame these challenges.

The mirror effect and the null list strength effect

The mirror effect is simply the fact that when two (or more) classes of stimuli differ in discriminability (e.g., as measured by d-prime), the result is almost always a mirror pattern for the HR and FAR such that the more discriminable class has both a higher HR and a lower FAR than the comparison stimulus class (Glanzer & Adams, 1985, 1990; Murdock, 2003). This general pattern was a challenge to the models in use at the time, since items from the higher d-prime class seem to be both less familiar than those in the lower d-prime class when not on the study list (thus producing a lower FAR) but more familiar than the items of the lower d-prime class after study (thus producing a higher HR). The explanation for the mirror pattern in SLiM and REM depends on the specific conditions producing the effect. For example, words of different normative frequency typically result in a mirror pattern (but see Criss & Shiffrin, 2004a) as does comparison of a study list containing items repeated multiple times to a study list containing once presented items (Cary & Reder, 2003; Criss, submitted-a; Stretch & Wixted, 1998). While REM and SLiM can account for both types of mirror effects, the underlying reason differs. The mirror pattern for normative word frequency is the result of properties of the words themselves (e.g., in REM, low frequency words are composed of uncommon and thus diagnostic features and in SLiM high frequency words are encoded with more variability at study and at test). In contrast, the mirror pattern resulting from frequency in an experimental list is captured by the models for the same reason the models predict the null list strength, differentiation.

The null list strength effect in recognition memory is the finding that strengthening some items on a study list does not harm performance for other items on that list. This is typically demonstrated by showing approximately equivalent performance for a set of items each studied the same number of times (e.g. one time or five times) but in a list containing items studied the same number of times (pure lists) or containing items studied different numbers of times (mixed list). This null list strength effect contrasts with the predictions of strength based models: These models predicted that variance increases with strength. As the strength of the stored memory trace increases, so does the variance of the match between a test item and the stored traces. Thus for items studied a single time, performance should be worse on a mixed than pure list and the opposite was predicted for items studied multiple times (see Hirshman, 1995; Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990 for further discussion). For both REM and SLiM, the process of differentiation produces the strength based mirror effect and the null list strength effect.

Differentiation

Differentiation is the idea that additional study of an item (e.g., through repetition or study time) results in further clarification of a single memory trace representing that study item. This is in contrast to the idea popularized by some strength based models that additional study results in storage of additional copies of the item. For example, suppose a word is presented for study three times. Many strength based models (e.g., Hintzman, 1988; Murdock, Smith, & Bai, 2001; Nosofsky, 1984) assume that three noisy copies of the concept are stored in episodic memory. In contrast, SLiM and REM assume that a single trace will be stored and that trace will be updated with each repetition, forming a more complete and more accurate copy of the concept than if the word had been presented just once.

At test, the REM and SLiM models (and most extant models) assume that the test item is compared to all items in episodic memory and the item is considered "studied" to the extent that it matches the contents of memory. What impact does differentiation have on this comparison process? Differentiation results in two factors that combine to produce both the strength based mirror effect and the null list strength effect. First, the match between the test probe and the memory trace stored during study of that item (if in fact that item was studied) is greater following multiple opportunities for encoding that study item. Because a single memory trace is updated with each study opportunity, that trace is a more complete and more accurate representation of the studied item. The more accurate a memory trace, the more it will match the features of its own representation when presented at test. Second, the match between any test probe (target or foil) and the memory traces stored during study of *other* items will have more opportunities to mismatch and thus match less well following multiple opportunities for encoding the studied items. For the case of a target, these two factors are in competition. As each item on the list receives additional study, the match between the target and its own memory trace grows but the match between the target and the remaining memory traces shrinks. However, the match between the target and its own memory trace tends to be large and thus dominates the overall match to memory, producing an increase in the hit rate with increases in study time. If the test probe is a foil, only the second factor applies. The degree to which a foil probe matches the contents of episodic memory decreases as study time increases, producing a lower FAR. Thus differentiation naturally produces a strength based mirror effect. The same reasoning applies to the null list strength effect. Strengthening some items on the study list results in those items being more dissimilar from other test item. This is true for both the match between a target item and the memory traces stored following study of other targets and for foil items which have no corresponding memory trace. Thus while HR and FAR may change slightly, they do so in tandem and overall discriminability does not change.

All of this is simply to make clear how differentiation can account for the mirror effect and the null list strength effect. REM and SLiM were developed as a means of accounting for these (and other) data and did so in part by incorporating the idea of differentiation. Now we turn to a more detailed description of the similarities and differences between SLiM and REM highlighting those that lead to an empirical test between the two models. The following is not a tutorial on how to implement each model. We refer a reader interested in this level of detail to the original sources (i.e., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997).

Details of REM and SLiM

Both REM and SLiM represent an item as a vector of feature values. The feature values are drawn from some distribution with some parameter(s). When an item is presented for study, an episodic memory trace is stored in the form of a vector that is a 'noisy' copy of the true complete vector representing the item. Additional study (either by means of massed or spaced presentation) results in additional storage in the same vector leaving the memory trace less 'noisy.' The definition of noisy differs for the two models. First consider REM. REM assumes that the features of an item are drawn from a geometric distribution where some features (i.e., those with low numerical values such as 1 or 2) are more common than others (i.e., those with high numerical values such as 10 or 18) and common features are less diagnostic than uncommon features. In REM, episodic traces contain two sources of noise or error as outlined in the top panel of Fig. 1. Prior to study, the vector is empty (i.e., the value of zero represents the absence of information). Study is guided by two parameters, one for the probability that a feature is stored (u) and one for the probability that the correct feature value

REM						
Study Stimulus	[1	5	1	3	1	9]
Test Stimulus	[1	5	1	3	1	9]
Memory (pre-study)	[0	0	0	0	0	0]
Memory (post-study)	[1	2	0	3	0	0]

SLiM						
Item Generator	[.86	.01	.01	.86	.01	.01]
Study Stimulus	[1	0	0	0	0	0]
Test Stimulus	[1	0	0	1	0	0]
Memory (pre-study)	[.15	.26	.18	.15	.19	.41]
Memory (post-study)	[.68	.26	.04	.15	.04	.08]

Fig. 1. A schematic of the steps involved in generating and storing a stimulus in episodic memory for both REM and SLiM. See the text for additional details.

will be stored given that a feature value is stored (c).[1] The storage of features is discrete and independent. For example, if the feature value of the actual stimulus is 5 and the feature is stored correctly, a 5 will be stored. If an incorrect feature is stored, it will be a value drawn anew from the geometric distribution. Once a feature is stored, its value is fixed and will not change during the course of the experiment. Additional study results in the storage of additional features. Thus, the two sources of noise in REM's storage process are the absence of information about a feature and an incorrect value being encoded. Absent features do not contribute to the decision process.

In SLiM, noise or error arises in both the perception of the stimulus itself and in the storage process as illustrated in the bottom panel of Fig. 1. The features of the actual stimulus are binary with some relatively small proportion of features taking the value of 1 and the remaining taking the value of 0, governed by the parameter f. The momentary perception of the stimulus is subject to variability such that a given feature of a given stimulus might produce different feature values at different moments in time. This is modeled by assuming that each perceived stimulus is produced by an underlying item generator. There are two types of features - those likely to be active and take the value of 1 (*p1* features) and those likely to be inactive and take the value 0 (*p0* features). In practice an item generator is simply a vector designating the probability that each feature will take the value of 1 during a presentation. The same process of stimulus generation is repeated at both study and test so the same feature of the given stimulus may produce a 1 during encoding but a 0 at test or vice versa. Thus, one source of noise in SLiM is the variability in encoding between study and test. Another source of noise in SLiM occurs during storage of the memory trace. Prior to study, initial feature values are taken from the logistic-normal distribution with a mean equal to the average probability of any feature having the value of 1. This initialization process is uninformative in that the values are randomly chosen and centered at the expected value of a feature. However, in contrast to REM, the initialization does allow each feature of each memory trace to participate in the decision process. During encoding, a subset of the features are learnable and the remainder are frozen at the initial uninformative value. Given that a feature is learnable, there is a learning rate parameter governing the degree to which the value stored in memory will approach the value of the study stimulus. Learning is graded in that a stored feature is an estimate of the actual value of the stimulus and the estimate becomes more accurate with additional study. Thus, another source of noise or error in SLiM arises from the features that are not updated and from the noisy encoding of those features that are updated.

As just described, many details of the stimulus generation and storage processes differ for the two models under consideration. On the other hand, the comparison between a test probe and the contents of memory is fairly similar. First note that both general theories of memory make a distinction between memory traces for episodic memory and those for general knowledge (see McClelland, McNaughton, & O'Reilly, 1995; Schooler, Shiffrin, & Raaijmakers, 2001). Neither theory implies separate or independent systems for different 'types' of memory. Rather this distinction simply implies the ability to access a subset of information without a requirement to access all instances of that information across the lifespan. The remaining discussion of the comparison process assumes that the decision process is accessing only recent episodic memory (i.e., the study list).[2]

During a recognition memory test, the complete set of features representing the test probe is compared to each of the noisy memory traces stored as described above. The comparison process results in one value for each stored memory trace indicating the likelihood that the memory trace resulted from the current test stimulus. The value is simply the probability of the data (e.g., the discrepancy between features of the stimulus and features stored in memory) given that the current stimulus generated the memory trace, divided by the probability

[1] Note that we report the parameter u which is the combination of two parameters in the original REM paper (Shiffrin & Steyvers, 1997). In the original paper, there was a probability of storing a feature ($u*$) for each time step (t). These two parameters are redundant and can be reduced to a single parameter u as follows: $u = 1 - (1 - u*)^t$. In the original paper, $u* = .04$ and $t = 10$ thus $u = .335167$ and this is the value we use.

[2] This is an idealization of the slightly less complete selectivity that would be provided by assuming that items are stored with a representation of the study context and that a joint context-plus item probe is used to probe memory.

of the data given that the current stimulus did not generate the memory trace. The parameters that are used in the calculation of this likelihood ratio differ slightly for the two models but neither model assumes full knowledge of all parameters. The main difference between the comparison process of each model is the information used to make a decision. In REM, the decision of whether to call the test item "studied" is based on the mean of the likelihood ratio for all memory traces whereas only the maximum likelihood ratio is used in SLiM. If the value (mean or max depending on the model) exceeds some criterion, then the item will be called "old" otherwise it will be called "new."

There are many approaches one could take to compare models (e.g., see Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006). The strategy we adopt is very simple. We compare qualitative predictions for a limited set of commonly employed variables including study time, normative word frequency, list length, and the similarity between targets and foils. We note that the original models have been extended to account for a host of new empirical findings (e.g., Criss & Shiffrin, 2004b, 2004c, 2005; Malmberg & Shiffrin, 2005; Malmberg, Holden, & Shiffrin, 2004). However, for simplicity we compare the original formulations using the parameter values from the original papers (i.e., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). The parameter values are reported in Table 1 and are taken from page 149 of Shiffrin and Steyvers and Table 10 (for Simulation 3), or Table 7 (for all remaining simulations) of McClelland and Chappell. As we now describe in detail, the models make different qualitative predictions even for the basic variables we employ, and they provide different explanations for the same empirical findings.

Simulation 1

Both SLiM and REM are similarity-based models, in that the match between the test probe and the contents of memory (i.e., the decision variable) grows as a function of the overlap between the study and test stimuli. Similarity effects are pervasive in the empirical literature. In studies where careful controls are employed to eliminate encoding strategies, robust similarity effects are found. For example, following incidental study of a list containing categories of similar items, the false alarm rate to unstudied items and the hit rate for studied items both increase as the number of similar studied items increases (Criss & Shiffrin, 2004b; Shiffrin, Huber, & Marinelli, 1995). The natural ability of both such models to account for these similarity effects has already been shown. Of interest in the current simulation is whether there is differential influence of similarity on the performance of the models.

To address this issue, we ran a simulation where a single item is stored in memory following the procedures for storage described earlier. Test items are constructed to vary in similarity to the stored item. In REM, similar items are constructed by generating two vectors with the standard procedure (i.e., drawing from the geometric distribution with parameter g). One of those vectors is deemed the target item and one the foil item. Each

Table 1
A list of the parameters of REM and SLiM and the values used for the current simulations

		Simulations 1 and 4	Simulation 2	Simulation 3
	REM			
	Number of features per item	20	10	
g	Geometric distribution parameter	.40	.40	
u	Probability of storing a feature	.335167	.335167	
c	Probability of storing the correct value given that a feature is stored	.70	.70	
	Criterion	1	1	
	SLiM			
	Number of features per item	73	37	313
	Probability of a feature being updated	.29	.29	.32
f	Probability of being a $p1$ feature	.25	.25	.04
$p0$	Probability the feature will be active for $p0$ features	.0005	.0005	HF = .03, LF = .001
$p1$	Probability the feature will be active for $p1$ features	.86	.86	.96
	Learning rate	.82	.82	.01, .04, .07, .10, .13, .16
	Mean, standard deviation for the distribution of initial feature values	−1.45, 0.75	−1.45, 0.75	−2.95 , 0.405
	Criterion	−2.51	−2.51	−1.4
ρ	Overall probability of an active feature ($=fp1 + (1-f)p0$)	0.2154	0.2154	0.0533

Note that ρ and the mean of the distribution are computed based on values of other parameters; they are not free to vary. HF refers to high frequency words and LF to low frequency words.

feature of the target item is independently copied to the foil item with some probability determined by the similarity parameter. If the parameter is equal to 1 then the two vectors are identical and if the parameter is equal to 0 then the two vectors are randomly similar. Note that randomly similar items do share features, the probability of sharing a given feature value, v, is determined by the geometric distribution,

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

In our simulations, we use a value of $g = .40$ resulting in an overlap between any two randomly generated vectors of approximately 25%. Thus, the actual overlap between the two stimuli is greater than the value of the similarity parameter and is governed by both the similarity parameter and the parameter for the geometric distribution. The procedure just described operates on the stimuli themselves, prior to encoding.

The original procedure for generating similar vectors in SLiM is different than that just described.[3] To be sure that any difference between the behavior of the models is due to the theoretical mechanisms rather than the arbitrary procedure used to produce similar vectors, we ran a simulation using the SLiM model but substituted the REM procedure for generating similar items. This simulation is identical to the original SLiM model in all details with the exception of the method for generating similar vectors. We simply borrow the procedure just described and generate two item generators, then copy the features of the study item generator to the test item generator with some probability determined by the similarity parameter.

Fig. 2 shows the probability an item is categorized as "studied" as a function of the similarity parameter used to generate the test item. Note that the leftmost (randomly similar foils, similarity parameter = 0.0) and rightmost (target items, similarity parameter = 1.0) points are approximately equal for each model, indicating that the overall level of discriminability is approximately equal. What differs is the response of the models to items that are similar but not identical to the studied item. The line labeled SLiM_original procedure in McClelland and Chappell (1998) used to generate similar items (described in Footnote 3). The line labeled SLiM_alternate is the SLiM model combined with the REM procedure for generating similar items. Regardless of how similar vectors are generated in SLiM, the REM model always produces a greater P(old) for similar foils than does SLiM, across the entire range of similarity. For any level of similarity, REM results in a higher match for similar foils than SLiM even when the discrimination between targets and randomly similar foils is approximately equivalent.

Why is this the case? In REM, the features take any value greater than zero and the value indicates diagnosticity (e.g., higher values are less common and thus more diag-

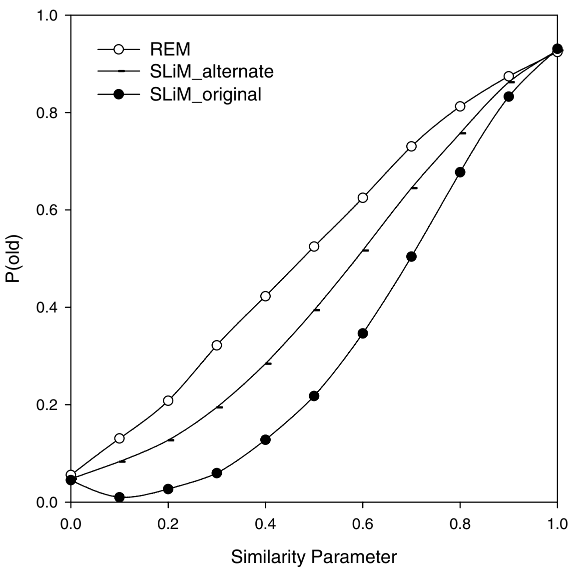

Fig. 2. The predicted probability of calling an item old (P(old)) as a function of similarity between the single item stored in memory and the test item. SLiM alternate is the SLiM model using the REM method for generating similar vectors.

[3] The original procedure used to generate similar items in SLiM follows. Recall that there are two types of features in SLiM: those likely to be active (*p1*) and those likely to be inactive (*p0*) when the stimulus is experienced. The study item generator is constructed using the standard procedures. Then, a similar test item is generated as follows: the test item generator is constructed by independently copying each *p1* feature of the study item generator with some probability governed by the similarity parameter. If the *p1* feature is not copied, the test stimulus generator takes the value *p0*. It is important to maintain the same overall probability of *any* feature in any generator being a *p1* feature. Therefore, the probability that each *p0* feature of the study item generator is copied to the test item generator is computed as follows:

$$(1 - \text{Similarity Parameter})\left(\frac{f}{1-f}\right), \tag{2}$$

where f is the expected proportion of features that are *p1* features. This procedure guarantees that any and all overlap between two item generators is dictated by the similarity parameter. This is particularly evident in the points for similarity parameter .10 and .20 of the SLiM_original predictions in Fig. 2. These two points actually dip below the value for similarity parameter of 0. The distribution used to generate item vectors in SLiM results in approximately 21.54% overlap between any two randomly chosen vectors for the parameters used here. Thus, a similarity parameter of less than .2154 creates vectors that are less similar than if the vectors had been randomly generated.

nostic). When matching the test stimulus to the contents of episodic memory, diagnosticity is taken into account. Thus matching a diagnostic feature (e.g., a value of 6) provides more evidence that the test stimulus is old than matching a less diagnostic feature (e.g., a value of 1). On the other hand, all mismatching features, regardless of diagnosticity, provide the same amount of evidence that the stimulus is new as can be seen in Eq. (3).

$$\lambda_{(i,j,k)} = (1-c)^{nq_{(i,j,k)}} \prod_{v=1}^{\infty} \left[\frac{c + (1-c)g(1-g)^{v-1}}{g(1-g)^{v-1}} \right]^{nm(v,i,j,k)}. \quad (3)$$

This is the REM equation for the likelihood ratio that test stimulus j matches memory trace i for stimulated subject k. The number of features with a non-zero value that mismatch between test vector j and memory trace i is labeled nq. The number of features that match and have the value v is labeled nm. Note that only matching features take into account the actual value of the feature. If there is no information stored for a given feature, indicated with a zero in the memory trace, that feature does not play a role in the calculation. The decision about whether an item is old or new is based on comparing the average of the likelihood ratios to some criterion.

SLiM is similar in that one type of feature is less common (features with value one) and thus provides more evidence in favor of an "old" response. However because the features stored in memory in SLiM are some continuous value between 0 and 1, the evidence from a match and the evidence from a mismatch both vary as a function of the distance to the value of the test stimulus, as shown below.

$$\lambda_{(i,j,k)} = \frac{\prod_{d=1}^{\#\text{-featurs}} (M_{di})^{S_{dj}} (1-M_{di})^{(1-S_{dj})}}{\prod_{d=1}^{\#\text{-featurs}} \rho^{S_{dj}} (1-\rho)^{(1-S_{dj})}}. \quad (4)$$

This is the SLiM equation for the likelihood ratio that the test stimulus j matches memory trace i for stimulated subject k. S_{dj} is the value of feature d in test stimulus j. M_{di} is the value of feature d stored in memory trace i and ρ is the estimate of any individual feature having the value of 1. A test stimulus is called "studied" if the maximum likelihood ratio exceeds some criterion.[4]

[4] The value used in the decision is actually the maximum log odds. The odds is simply the likelihood computed via Eq. (4) multiplied by the prior probability that the test item is item i. The prior term is constant for all items in a given list length and thus does not differentially contribute to any one experimental condition more or less than another condition (provided that the list length for each condition is equal). SLiM uses the log of the odds value, rather than the odds itself. The log is a monotonic transformation and thus provides the same information as the untransformed odds. Neither the use of the prior or the log transformation contribute to any difference between REM and SLiM.

Thus, the difference in similarity functions comes from the more limited range of possible matches in SLiM compared to REM and the graded evidence provided by mismatching features in SLiM but not REM.

How might this difference play out in empirical data? Measuring similarity in the world in a way that translates directly to model parameters is perhaps not yet feasible. We know similarity is flexible, context-dependent, and subject to bias (e.g., Goldstone, 1994; Tversky, 1977) but our models have not yet incorporated such processes. Thus, we cannot rule out either REM or SLiM based on an empirically measured similarity value. Instead, our approach is to look at a task where stimuli are pairs of items and similarity is simply the number of items shared between stimuli.

Simulation 2

We now turn to the associative recognition paradigm. In this paradigm, pairs of items are studied (e.g., AB, CD, and EF) and the memory test requires discriminating between intact pairs (AB) and rearranged pairs (AD) constructed by combining individual items from two different study pairs. This paradigm allows another comparison of the behavior of the models in the face of considerable overlap between stimuli without relying on the similarity parameter. In this simulation, pairs were studied followed by testing intact (AB), rearranged (AD), and novel pairs (XY). Novel pairs are constructed from two items that were not presented on the study list. Fig. 3 shows the performance of the models for these three types of test pairs as a function of list length. Performance for intact and novel pairs was approximately equal across models but the performance on rearranged

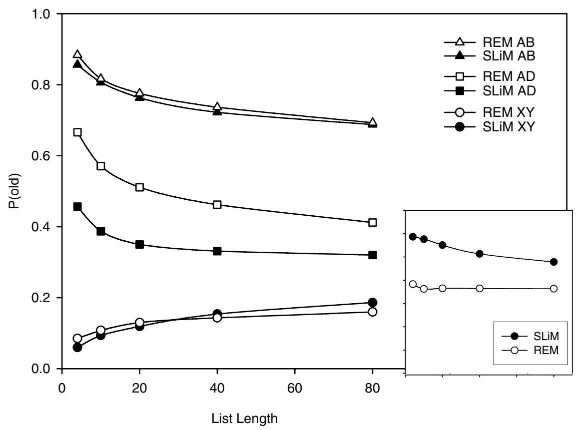

Fig. 3. The predicted probability of calling a pair "old" (P(old)) as a function of list length. AB refers to intact test pairs, AD refers to rearranged test pairs, and XY refers to novel pairs. The inset plots d-prime for intact and rearranged pairs as a function of list length.

pairs differs. Consistent with the first simulation, REM has a greater tendency to false alarm to foils that overlap with targets, the rearranged pairs in this case, than does SLiM. The inset contains d-prime for intact and rearranged pairs and shows that discriminability in associative recognition is much better in SLiM than REM and that SLiM but not REM predicts a list length effect for associative recognition. Empirical data consistently show list length effects in associative recognition when list length is manipulated between-list or within-list (Nobel & Shiffrin, 2001; Criss and Shiffrin, 2004c, 2005). As discussed in Criss and Shiffrin (2004c; 2005), REM cannot predict a within-list list length effect for associative recognition and a between-list list length effect is only obtained with a change in the criterion parameter. This and other findings lead Criss and Shiffrin to adopt an emergent features account of associative recognition. Following Murdock (1982; 1997), they assumed that study of a pair results in the storage of item and association information that contain independently generated features. Thus, the similarity between an association and the items from which is was generated is no greater than the similarity between two randomly selected items. SLiM is able to predict a list length effect for associative recognition without assuming emergent associative features. Ongoing work is assessing whether or not the original formulation of SLiM can account for the full set of findings presented in Criss and Shiffrin that lead to the emergent features extension of REM.

A related issue is the relative level of performance for associative recognition (AB vs. AD) and pair recognition (AB vs. XY). Nobel and Shiffrin (2001) found that following an identical study list of 20 pairs, d-prime for a pair recognition test was 0.58 higher than d-prime for an associative recognition test. Shiffrin and Steyvers (1998) simulated a variety of strategies for performing associative recognition and found no satisfactory strategy that produced performance close to the behavioral data, in part due to the high level of FAR to rearranged pairs. According to the inset in Fig. 3, at a list length of 20, SLiM predicts that the difference in d-prime for pair recognition and associative recognition is 0.77 while REM predicts a difference of 1.31 (both showing better performance for pair than associative recognition).

Both Simulations 1 and 2 suggest that regardless of whether similarity is manipulated by constructing similar vectors or by rearranging items in a pair, REM is more likely than SLiM to commit false alarms to test stimuli that overlap with studied stimuli. This translates into REM predicting a null list length effect and generally poor performance in associative recognition, neither of which are supported by empirical studies. SLiM fares better in that it does predict a list length effect and the relative level of performance is closer (though not identical) to empirical data. Further simulations are required to assess whether SLiM can account for within-category but not between-category list length effects in associative recognition demonstrated in Criss and Shiffrin, (2004c, 2005).

Simulation 3

In this simulation, we examine the impact of study time on the word frequency mirror effect. The impact of normative word frequency (WF) on episodic memory performance is perhaps one of the most studied topics in memory research. The standard finding is better performance for words of low normative frequency (LF, e.g., barge) than words of high normative frequency (HF, e.g., drive), manifest in both a higher HR and lower FAR (e.g., Glanzer, Adams, Iverson, & Kim, 1993). Both REM and SLiM attribute the effect of WF to a difference in the underlying properties of the words themselves. REM assumes the features of high and low frequency words differ in diagnosticity. LF words have a lower *g* parameter for the geometric distribution than HF words. The result is that HF words have more common and thus less diagnostic features than LF words. Further, because HF words have common features, they also tend to be more similar to one another than LF words. SLiM assumes that HF words are more subject to encoding variability relative to LF words (i.e., *p0* features of HF item generators have a higher value than LF item generators). Both assumptions have received empirical support. Words with distinctive features are better remembered than words with less distinctive features, consistent with REM (Criss, submitted-b; Landauer & Streeter, 1973; Malmberg, Steyvers, Stephens, & Shiffrin, 2002; Zechmeister, 1972). Words present in few semantic contexts in the environment are better remembered than words experienced in many different contexts consistent with SLiM (Steyvers & Malmberg, 2003). Rather than dispute the underlying cause of the WF effect, the current simulations are concerned with a case where the mirror pattern is disrupted.

Recently, there have been numerous reports of disruptions of the mirror pattern (e.g., Balota, Burgess, Cortese, & Adams, 2002; Criss, submitted-b; Criss & Shiffrin, 2004a; Hirshman & Arndt, 1997; Hirshman, Fisher, Henthorn, Arndt, & Passannante, 2002), questioning the assumed ubiquitous nature of the effect. We focus on the case of drug induced amnesia in part because REM has already been applied to the relevant data. When participants are under the influence of the benzodiazepine Midazolam during study of a list of words, the resulting pattern of data is a higher HR and a higher FAR for HF than LF words. The same participants show the standard pattern of a higher HR and lower FAR for LF than HF words when given saline (Hirshman et al., 2002, see also Mintzer, 2003). Original reports of these findings argued that the data ruled out a single process model of memory and favored a dual process with one component (i.e., recollection) disrupted by Midazolam. Malmberg, Zeelenberg, and Shiffrin (2004) showed that REM, a single process model, accounted for the pattern of results. Recall that two parameters govern storage in REM: u, the probability of storing a feature and c, the probability that the correct value will be stored given that a feature is stored. The u parameter is typically assumed to vary as a function of study time and the c parameter is assumed to be a fixed parameter of the system, not subject to experimental manipulation. Malmberg et al., showed that varying the parameter c at study allowed REM to account for the reversal in the HR effect. Specifically, the Midazolam condition was modeled as having a lower c during encoding than the saline condition (the value of c at test was the same for both conditions). Briefly, the idea is that Midazolam-induced amnesia disrupts the normal processing of the memory system causing noise (e.g., incorrect features) to be stored. In the REM account, the amount of information stored (i.e., number of features) is approximately the same for saline and Midazolam; Midazolam simply results in more cases where the value of the stored feature does not match the value of the studied stimulus. This is in contrast to allowing the parameter u to vary as a consequence of Midazolam, resulting in the storage of less information (e.g., fewer features).

Recall that in SLiM, every feature in a memory trace begins as noise and those features able to be learned move toward the actual value of the stimulus as a function of study time. In contrast to REM, SLiM can account for the reversal in the HR without assuming that the processes underlying memory are different for an intact and an impaired memory system. Instead, a SLiM account can be given under the simple assumption that Midazolam-induced amnesia reflects a reduced learning rate. Fig. 4 shows the consequence of varying the learning rate parameter on the WF effect in SLiM. In this simulation, WF and learning rate are varied in a single study list containing 60 items equally divided among conditions. Test items include targets from all learning rate by WF conditions as well as unstudied LF and HF words. The leftmost points in Fig. 4 are the false alarm rates (learning rate = 0) and we see that SLiM predicts HF FAR > LF FAR, consistent with empirical data. The remaining points are targets with various amounts of study and we see that SLiM predicts a pattern in which low learning rates result in a HF HR advantage and high learning rates result in a LF HR advantage. As described earlier, HF words have a higher value for *p0* which results in more spurious features with the value of 1. Because active features are more diagnostic than inactive features, matching a 1 provides more evidence that the test item is old than matching an inactive feature. Thus, the higher FAR for HF words is due to spurious matching of active features. A moderate amount of encoding is required to overcome the effect of this spurious matching, thus the advantage of LF targets is not immediately present.

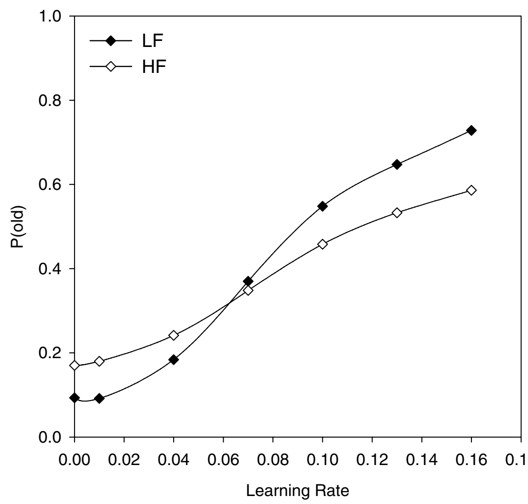

Fig. 4. The predicted probability of calling an item "old" (P(old)) as a function of the learning rate parameter in SLiM. HF refers to high frequency words and LF to low frequency words.

Experiment

If the HR advantage reverses with reduced learning rate, then it should be possible to demonstrate this pattern of data in participants with unimpaired memory. Recall that study time manipulations are modeled by the learning rate parameter. Thus SLiM predicts that very short study times should result in a HF HR advantage that reverses with additional study time. In other words, Midazolam induced amnesia is simply a point on a continuum between low and high learning rates. Other studies found disruptions of the WF mirror pattern with Korsakoff and geriatric participants, both perhaps consistent with a deficit in learning rate explanation (Huppert & Piercy, 1976 & Balota et al., 2002, respectively). If the learning rate explanation is viable, we should be able to mimic the effect by reducing learning rate in an otherwise unimpaired memory system. In contrast, REM required introducing different processing assumptions (storing noise rather than nothing) to account for performance of participants who have taken Midazolam. In fact, Malmberg et al. (2004) explicitly ruled out the possibility that Midazolam induced amnesia could be accounted for by varying the u parameter, the parameter used to model study time.[5] The goal of the following study is to test the predictions of SLiM. Three different study times were used in a mixed-list design followed by single item yes–no recognition memory test.

Methods

Participants

Thirty-five people from the Carnegie Mellon University community received $7 for participation.

Materials

The LF word pool consisted of 242 words with a frequency between 1 and 10 per million ($M = 3.36$, $SD = 2.51$). The HF word pool consisted of 242 words with a frequency of at least 50 per million ($M = 110.22$, $SD = 86.52$; Coltheart, 1981; Kucera & Francis, 1967).

Procedure

The study list was constructed anew for each participant and contained 60 HF and 60 LF words randomly selected from their respective word pool. An equal number of LF and HF words were assigned to each of three different study time conditions: 150, 300, or 600 ms. All 120 study items were randomly intermixed and presented individually on a computer monitor for the designated study time. A blank screen was presented for 250 ms between consecutive study trials. Between the study and test lists, participants played a game containing no words for 35–50 min. Following the break, participants made a yes-no single item recognition memory decision for 240 test items including all of the targets and an equal number of HF and LF unstudied foils.

Results and discussion

Hits, false alarms, and standard errors are reported in Table 2. FARs follow the standard pattern with higher FARs for HF than LF words but the HRs show a reversal as a function of study time. The LF FAR is lower than the HF FAR $t(34) = -3.613$, $p = .001$. A planned comparison on the interaction between WF and study time for target items supported what is evident in Table 2: a reversal in the HR as a function of study time $F(1, 34) = 4.16$, $MSE = .010$, $p = .049$. Consistent with a priori predictions of SLiM, we see that the LF HR advantage does not emerge until a sufficient amount of study time has accumulated. No doubt, the empirical effect is small. The bottom half of Table 2 shows the data from the Midazolam condition of Hirshman et al. (2002).[6] The best performance in the Hirshman et al study is approximately equivalent to the worst performance in our experiment and is the only condition where we find a larger HR for HF words. Thus, it is likely that larger effects can be found by pushing performance even lower than we were able to do here. Further reduction of study time is possible but becomes somewhat tenuous because of the need to ensure that participants have sufficient time to read the words. Alternative potential strategies to reduce the learning rate and thus performance (e.g., masked presentation, low contrast presentation, dual task encoding, etc) await further investigation.

Table 2
The probability of calling an item old as a function of condition for the experiment and for the Midazolam condition of Hirshman et al. (2002)

Study time (ms)	High frequency	Low frequency
0	.317 (.032)	.227 (.024)
150	.460 (.035)	.436 (.033)
300	.469 (.040)	.489 (.033)
600	.490 (.037)	.536 (.036)
	Hirshman et al. (2002)	
0	0.38	0.25
1500	0.39	0.32
1200	0.43	0.39
2500	0.46	0.41

When available, standard errors of the mean are reported in parentheses.

Simulation 4

In our final simulation, we consider the decision rule used to judge whether a test stimulus is old or new. In REM, the decision is based on the average of the likelihood ratios while in SLiM the decision is based on the single maximum likelihood ratio (scaled by the list length, see Footnote 3). This difference has been pointed out before (Shiffrin & Steyvers, 1998) with the conclusion that the mean is the optimal decision rule because

[5] Note that in the simulations reported by Malmberg et al. (2004) there is no crossover in the HR when the u parameter (i.e., study time) is varied. We conducted additional simulations not reported here and show that in fact REM can produce a reversal in the HR as a function of study time. One difference between the Malmberg et al simulations and our own is the range of u under consideration. The Malmberg et al simulations begin with $u = .20$ and our simulations show a HF HR advantage for values of $u \leqslant .05$.

[6] For convenience, the distractor time between study and test we used is somewhat less than the break of 70 min required to minimize the sedating effect of Midazolam in the Hirshman et al. (2002) study. Otherwise, the details of our experiment (e.g., list length) are similar to Hirshman et al. with the obvious exception that no drugs were administered.

it uses all available information. Indeed, for simulations with REM, Shiffrin and Steyvers showed that the mean rule produces superior performance in terms of d-prime than the max rule but otherwise similar patterns of data. To examine this issue in SLiM, we simply ran a simulation with 40 study items followed by testing of targets and randomly similar foils. We recorded the odds values generated for each test item and divide them into four categories, the maximum odds value ($n = 1$) and all the remaining odds ($n = 39$) for cases where a target is tested and cases where a foil is tested. Histograms of the distributions are plotted in Fig. 5. As expected, the distribution with the highest mean is the distribution for the maximum value when a target item is tested, followed by the max when a foil item is tested, followed by the distributions for all values other than the max. The distributions of all of the odds values other than the max are approximately identical regardless of whether a target or a foil is tested. Indeed for SLiM, the max and the mean appear to provide the same information and the same level of performance, as indicated from the ROC analysis shown in Fig. 6.

In fact, we obtained similar plots for REM and suspect that the small benefit in d-prime for the mean rule reported by Shiffrin and Steyvers (1998) was the result of using slightly different location for the criteria for the max and mean rules. When the target and foil distributions have unequal variance, as is the case for both REM and SLiM, a change in criterion can result in a change in d-prime (Green & Swets, 1966; Macmillan & Creelman, 1991). Note however, that when performance is very low, there is a small benefit for the mean rule over the max rule. Thus based on our simulations with both REM and SLiM, we conclude that there is little practical difference between using the max or mean rule except under conditions with very low performance. The mean rule does have the user-friendly quality of centering the distributions so that the target and foil distributions intersect at an odds value of 1 thus producing an optimal criterion of 1 for all experimental manipulations (though note that participants do not necessarily use the optimal criterion). Otherwise, the use of the mean or max rule is not a critical difference between the models.

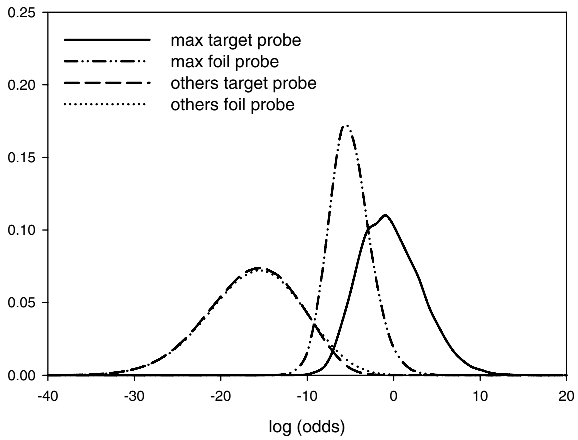

Fig. 5. The distribution of log(odds) for the maximum value and the remaining values when a target item is tested and when a foil item is tested.

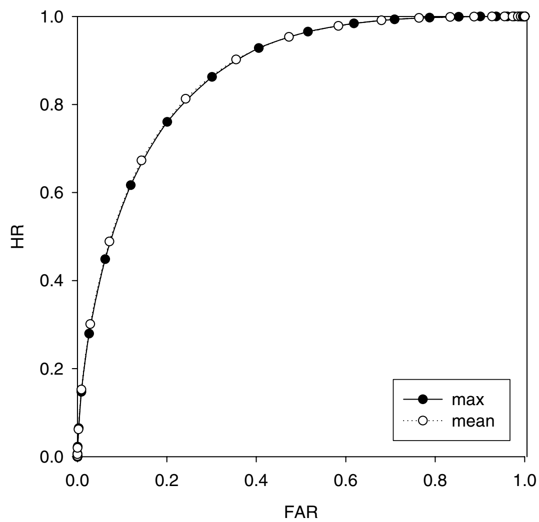

Fig. 6. A ROC plot for the max and mean rules in SLiM. The plot was generated by taking the cumulative sum across successive bins of the odds distributions.

Differentiation models are not fully informed likelihood models

The following concerns a property the two models share that is often misunderstood. Both models compute a likelihood ratio as part of the decision process and are often referred to as likelihood models. Unfortunately, this same term is also used to refer to another class of models in the tradition of signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 1991). In these latter models, which we will call fully informed likelihood models, it is assumed that the memory system knows the full statistical properties of the distribution of familiarity values associated with both the old and the new stimuli used within each test condition of the experiment. The likelihood of obtaining the familiarity value associated with a given test item under each of the two distributions is then used as the basis for categorizing the item as old or new and for assigning a confidence rating.

While it is true that REM and SLiM make use of what both models call likelihood ratios, these ratios are (as the name of SLiM indicates) *subjective* likelihoods based on the quality of match between the test item and memory. These subjective likelihoods correspond to the familiarity values used in the fully informed likelihood ratio models, but crucially, unlike fully informed likelihood models, no knowledge of the distribution of such values specific to the experimental condition is assumed. REM and SLiM both make use of some parameter values assumed to reflect knowledge of item properties in calculating the likelihood ratio [see Eqs. (3) and (4)]. However, fixed values for these parameters (c and g in REM and ρ in SLiM) are used across all conditions of an experiment. Further, McClelland and Chappell make clear that all values used in the likelihood ratios are assumed to be estimates of the true values based on long-run averages of item properties. Similarly, the convention in REM is that the parameter values are accurate long run averages across the system's life. There is some empirical support for the approach taken in REM and SLiM. Balakrishnan and Ratcliff (1996) demonstrated that fully informed likelihood ratio models necessarily predict that cumulative confidence curves will cross (at least for the case where the subjective response to targets and foils are normally distributed variables). They conducted a range of empirical studies, and never found a case where cumulative confidence curves cross, consistent with the idea that full condition-specific distributional knowledge is not used to compute likelihood ratios. Both REM and SLiM are supported by this finding because they do not make use of different parameter values for different experimental conditions when computing the likelihood ratio. This was demonstrated for SLiM in the original McClelland and Chappell paper and was confirmed for REM in simulations not reported here. Because of these differences between REM and SLiM on the one hand and fully informed likelihood models on the other, we encourage authors to abandon the use of the label 'likelihood models' when discussing any of the models so as to avoid confusion. Some further modifiers are required so that the models can be properly distinguished.

General discussion

To summarize the analyses presented above, despite popular belief that the REM and SLiM models make parallel predictions, the models sometimes behave quite differently. In simulations we showed that (1) REM is more likely to false alarm to items that share any degree of similarity with a studied item (2) SLiM a priori predicts a reversal in the HR pattern for HF and LF words as a function of study time and this prediction was confirmed empirically. In contrast, the study time parameter was explicitly ruled out by Malmberg et al. (2004) as a basis for this reversal in REM. Interestingly, the one previously acknowledged difference between the models appears not to have serious consequences: (3) Using either the max or the mean of the likelihood ratios as a basis for decision provides virtually identical information.

The differences between the models are important to keep in mind for at least two reasons. Perhaps most importantly, they indicate that the shared core assumptions of the two models are not enough to ensure that they make identical predictions. This means that, as with other models, each one's successes and failures apparently depend on additional assumptions—ones that are unlikely to have been central to the design goals of either model's inventors. Thus, efforts to confirm or disconfirm the core assumptions must be conducted with care within either model, taking into account the ancillary assumptions.

A second important point is that the models often account for the same data in different ways, thus leading to alternative ways of thinking about what has been learned from a particular set of experiments. Consider two cases in point: First, the finding that REM and SLiM differ in their response to similarity between studied items. This difference may be especially important for our understanding of the representation of pairs in memory. Certain results from studies on associative recognition cannot be modeled in REM without assuming that pairs of items give rise to additional emergent features not present in the members of the pair (e.g., see Criss, 2005; Criss & Shiffrin, 2004c, 2005). The same may not be true if these results are modeled with SLiM. Indeed, we are currently exploring the possibility that associative recognition can be explained by SLiM without recourse to the notion that pairs contain additional emergent features. Second, consider the effects of Midazolam on memory. In modeling work based on REM, the absence of a mirror effect under the drug could not easily be explained as a simple reduction of learning; instead it was concluded that the drug resulted in a reduction in the accuracy of encoding item properties. But SLiM accounts for the effects of Midazolam in terms of a simple reduction of learning rate. These quite different explanations of amnesia merit further testing.

In summary, it appears that the apparently innocuous differences between REM and SLiM can have quite dramatically different implications for the conclusions one may draw from the results of experiments. However, it must be noted that the above simulations used parameter values reported in the original papers. Since the behavior of both models have a complex high dimensional non-linear dependence on parameter values, the results presented here represent a very local observation of what each model may be capable of predicting and may not be generally true for all parameter settings. Based on our own experience with the models spanning several years and a variety of parameters and experimental designs, we are fairly confident that the findings reported here are not arbitrary or confined to the particular parameter values. However, this can only be addressed with more formal methods for addressing model mimicry (e.g., Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). Further research is clearly called for so that we may better understand the generality of the differences between the models . In the meantime, we urge others to keep in mind that the models are indeed quite different, and that the behavior of one should not be assumed to be fully representative of the behavior of the other.

References

Balakrishnan, J. D., & Ratcliff, R. (1996). Origins of the feeling of confidence in classification data. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 615.

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: Evidence for two processes in episodic recognition memory. *Journal of Memory and Language, 46*, 199–226.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science, 11*(4), 267–273.

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

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A*, 497–505.

Criss (submitted-a). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect.

Criss, A. H. (submitted-b). The contribution of letter distinctiveness to the word frequency effect.

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

Criss, A. H., & 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.

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

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

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

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

Flexser, A. J., & Bower, G. H. (1974). How frequency affects recency judgments: a model for recency discrimination. *Journal of Experimental Psychology, 103*, 706–716.

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

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & 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., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*(3), 546–567.

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

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

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

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.

Hirshman, E., & Arndt, J. (1997). Discriminating alternative conceptions of false recognition: The cases of word concreteness and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(6), 1306–1323.

Hirshman, E., Fisher, J., Henthorn, T., Arndt, J., & Passannante, A. (2002). Midazolam amnesia and dual-process models of the word-frequency mirror effect. *Journal of Memory and Language, 47*, 499–516.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: a theory for episodic, semantic, and procedural tasks. *Psychological Review, 96*, 208–233.

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

Huppert, F. A., & Piercy, M. (1976). Recognition memory in amnesic patients: Effect of temporal context and familiarity of material. *Cortex, 12*, 3–20.

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

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior, 12*, 119–131.

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

Malmberg, K. J., Holden, J., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 319–331.

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.

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30*(4), 607–613.

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 322–336.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419–457.

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.

Mintzer, M. Z. (2003). Triazolam-induced amnesia and the word-frequency effect in recognition memory: Support for a dual process account. *Journal of Memory and Language, 48*, 596–602.

Murdock, B. B. (2003). The mirror effect and the spacing effect. *Psychonomic Bulletin & Review, 10*(3), 570–588.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609–626.

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

Myung, I. J., Forster, M. R., & Browne, M. W. (Eds.), (2000). Model selection [Special issue]. *Journal of Mathematical Psychology, 44*, 1–39.

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 384–413.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 104–114.

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

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: a two-dimensional signal detection model. *Psychological Review, 111*, 588–616.

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

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

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

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

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95). Oxford, England: Oxford University Press.

Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5), 760–766.

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.

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

Wagenmakers, E. J., & Waldorp, L. (Eds.), (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology, 50*, 99–214.

Wagenmakers, E-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*, 28–50.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal-detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11*, 616–641.

Zechmeister, E. (1972). Orthographic distinctiveness as a variable in word recognition. *American Journal of Psychology, 85*, 425–430.