Criterion Setting and the Dynamics of Recognition Memory

Gregory E. Cox (grcox@indiana.edu)
Richard M. Shiffrin (shiffrin@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. Tenth St., Bloomington, IN 47405 USA

Abstract

A major puzzle in recognition memory has been the process by which participants set reasonable old/new decision criteria when the study and test lists are comprised of items of widely varying types, with differing degrees of baseline familiarity and experience (e.g., words vs. random dot patterns). We present a model of the recognition process that addresses this issue. Its core assumption is that recognition decisions are based not on the absolute value of familiarity, but on how familiarity changes over time as features are sampled from the test item. We model recognition decisions as the outcome of a race between two parallel accumulators: one that accumulates positive changes in familiarity (leading to an "old" decision) and another that accumulates negative changes (leading to a "new" decision). Simulations with this model make realistic predictions for recognition performance and latency regardless of the baseline familiarity of study and test items.

Keywords: Episodic memory; recognition memory; memory models; reaction time.

Introduction

Recognition memory has been a focus of cognitive science for many decades, yet remains a rich source of intriguing results that challenge the many models that have been developed in this field (e.g., SAM, Gillund & Shiffrin, 1984; MINERVA2, Hintzman, 1988; TODAM, Murdock, 1982; REM, Shiffrin & Steyvers, 1997; subjective likelihood, McClelland & Chappell, 1998; BCDMEM, Dennis & Humphreys, 2001). These models share some similarities, including a reliance on "familiarity" as a major or sole component, where familiarity is a global signal of match between the memory probe and memory traces. During study of a list of items, memory traces are formed for the items and enter long-term memory, where individual traces may be stored separately (e.g., SAM, REM) or in composite (MINERVA2, TODAM, BCDMEM). At test, a probe item is presented and the task is to say whether or not the probe was on the most recently studied list. To make this decision, the probe is compared to the composite trace or separate traces, resulting in a familiarity value (summed across individual matches in the case of separate traces) used to make a decision, with higher values of familiarity leading to an "old" decision. Typically, the match depends both on context features (features defining the general list context) and content features (the features of the items on the list).

Generally, these models assume that only those traces from the most recently studied list enter into the matching process. In such models, the primary sources of noise in recognition are the traces of the other list items (called, for short, *item noise*). Another class of models, restricted to the usual case in which study lists are comprised of words (e.g., BCDMEM; Dennis & Humphreys, 2001), assumes that the noise inherent in recognition memory arises not from traces of the other studied words, but from traces of the test word formed during previous life history. In this case, the familiarity value is based on a comparison of the current context to the context stored with the probe. Target words, having occurred in the study context, tend to produce higher familiarity values than foils, with noise arising when probes have occurred in contexts similar to the study context—we term this *context noise*.

Both item noise and context noise models have been shown to account for myriad effects in the recognition memory literature, including manipulations of list length, list strength, and similarity between study and test items (for a review, see Malmberg, 2008). Each of these model types has its merits and demerits, but both suffer from a problem that arises most strongly when study and test items are a mixture of different stimulus types, strengths, and/or similarities. One can imagine, for example, a study list that mixes words (with many traces from history), random dot patterns (few or no traces from history), faces (known or unknown and in identical or similar views), and common objects, with different numbers of each of these types and items of each type studied for long or short times. It may be assumed that under these conditions, the absolute degree of match of different kinds of test items (both targets and foils) to the traces in memory will vary quite widely (e.g., a dot pattern test item will activate/match no history traces, but a word will). In such a case how does a criterion (or several criteria if graded judgments are required) get set? Different criteria would be needed for each stimulus type, lest certain types always be given "old" judgments and others always given "new" judgments. Thus far, the field has not provided a mechanism by which appropriate criteria could be established and used during testing of a single list of vastly different test types.

In this research, we propose a way by which recognition decisions could be made without reference to fixed criteria and in such a way that they are sensitive to the structure of both the study and test lists. The essential idea is that the test item itself can serve as its own reference point. It can do this because perception takes time. As the percept of the test item develops, perhaps over a few hundred ms, it is used in conjunction with context to probe memory, resulting in a familiarity value that evolves dynamically. Early in perception, this familiarity will be depend primarily on context matching, since while few item features have yet to be perceived, list context is more or less constant across testing and is known before the test item arrives. In these early moments of the test trial, then, familiarity will result mostly from list traces, since the item is not well enough perceived for historical traces of the test item to be evoked. For example, if the test item is a word, early matching will be based mainly on traces of list words, since these match on context and the test probe consists mainly of context features. As perception of the test item continues, the test probe will include more and more features of the test item itself, increasingly matching history traces of that test item while reducing the match with list traces (besides that of the test item itself). Thus, the time course profile of familiarity will have a characteristic shape that differs for targets and foils: target familiarity will gain as perception continues (because the list trace of the target will be adding activation) whereas foil familiarity will gain less or decrease (because there will be no list trace to add to the overall familiarity). The exact shape of the profiles will depend on the way matching is calculated. If we calculate activation with a likelihood ratio, as we do shortly, then the target profile tends to rise and the foil profile tends to fall.

The critical element of this approach, using the shape of the familiarity profile to make a decision, is the independence of the decision from overall level of activation: Some item types will have high familiarity and others low familiarity, but the profile shape will remain diagnostic of the target-foil discrimination. In this paper, we lay out the details of a system that will implement these ideas, and show how recognition decisions might be made sensibly despite wide variations in type of item studied and tested. Let us emphasize that there are many ways to estimate stimulus specific criteria if there is sufficient experience with a given item type, either during list study or during testing. Such tuning of criteria might indeed be superimposed on the model we are about to present. Here, our concern is recognition judgments that must be made without such extended experience. The development of this model is still in its infancy, so this paper serves primarily as a proof of concept. We have made several arbitrary and simplifying assumptions in the simulation results that follow that may need to be revised in applications to real data. Such simplification has the merit, however, that the basic concepts are implemented in a straightforward fashion.

The Model

The model described below owes much to the original REM model of Shiffrin and Steyvers (1997) and somewhat to the ARC-REM model of Diller, Nobel, and Shiffrin (2001), although it differs substantially from both of those models. The parameters of the model are summarized in Table 1.

Structure of Memory Traces

Memory traces for items are assumed to consist of a set of features, with equiprobable binary values, e.g., $[0,1,0,0,1,0,1,1,1,\ldots]$. Although we are agnostic as to exactly what these features might represent, we allow that some of these features arise from properties of the item itself, and others arise from properties of the context in which the trace for the item was encoded. Relevant item properties might include the item's physical features or semantic content. Contextual properties could include information about the time and place in which the item was studied. It is important to note that although features in the memory trace can arise from multiple sources (items, context), we do not assume that the "memory system" has access to the source identity of features in memory. That is, the memory system treats a context feature and an item feature identically when matching test probes to memory traces.

Name	*Value*	*Description*
K	*varies*	Number of history traces for a test item that have the potential to be activated.
N_c	30.	Number of item features stored in each memory trace.
N_x	30.	Number of context features stored in each memory trace.
c	0.85	Probability of correctly copying a feature to a memory trace.
T_s	*varies*	Study item presentation duration.
ρ	60.0	Rate (per unit time) at which feature values are sampled during encoding.
θ	1.0	Familiarity threshold for a trace to be activated.
α_{old}	13.0	Evidence threshold for making an "old" response.
α_{new}	-16.0	Evidence threshold for making a "new" response.

Table 1: Parameters of the model, along with the values used in the simulations reported in this paper. It should be noted that there is no compelling reason to set $N_c = N_x$; this is merely for convenience.

Encoding

When an item is studied, a memory trace is created and added to memory. However, encoding is assumed to take place over time, and to be subject to noise. To create a memory trace for an item, features are sampled from the item one at a time. With probability c, the sampled feature is copied into the developing memory trace correctly and with probability $(1-c)$, a random value (0 or 1) is stored for that feature. Because each feature sampling event is presumed to be independent, it is possible that a feature may be sampled for which a value is already stored. In this case, the most recent sampled value (which, again, might be copied correctly or randomly) replaces any value that was already stored in the developing memory trace.

Memory traces also contain features from the context in which the item is experienced. In accord with findings that the amount of context stored is independent of study time (Malmberg & Shiffrin, 2005), and partly because one could imagine a trade-off between assumptions of probabilistic storage and relative number of context features, we assume that all context features have a value stored in the memory trace. Such storage is still noisy, so a context feature is correctly stored with probability c, otherwise a random value (either 0 or 1 with equal probability) is stored with probability $(1-c)$.

This encoding procedure is just a simple extension of the encoding process found in many memory models, especially

REM. It differs from REM in that features are sampled from items *with replacement* (i.e., the same feature might be sampled more than once), although this change is probably not consequential for present applications. Content features will be present in the final memory trace to a degree that depends on study time. As study time is reduced, fewer samples will be taken and fewer features will be stored, making it hard to differentiate between memory traces on the basis of their content features, resulting in reduced performance.

The notion that perception of an item can be described as a sequence of independent feature samples is not foreign to memory models (Brockdorff & Lamberts, 2000; Wagenmakers et al., 2004). As in such models, we assume that the distribution of times between samples is exponential with rate ρ, $f(t) = \rho \exp(-\rho t)$, meaning that feature sampling is a homogeneous Poisson process. Thus, for a given trial, the number of samples drawn, n_s, when studying an item for T_s time units is sampled from the following distribution:

$$\Pr(n_s \text{ samples}; T_s, \rho) = \frac{e^{-\rho T_s} (\rho T_s)^{n_s}}{n_s!}.$$

Recognition

We consider just yes-no recognition (i.e., not multi-alternative forced choice or confidence ratings). On each trial, a test item is presented. Just as in study trials, the test item consists of a set of binary features, which are sampled one at a time (with exponentially distributed times between samples) and added to a developing memory trace. The developing probe trace is compared in parallel to the traces in memory that are similar to the probe at that moment, producing a familiarity value that evolves as more features are sampled. The rate of change of this familiarity value as more probe features are sampled constitutes evidence for the recognition decision: a preponderance of negative changes is evidence that the item is new, while primarily positive changes are evidence that the item is old. Two non-interacting accumulators keep track of positive and negative familiarity changes and when one accumulator reaches a threshold, the corresponding response ("old" or "new") is made.

Feature Sampling The process of sampling features from the probe item is identical to the feature sampling process at study, with the exception that the sampling process is terminated not by stimulus offset (as with a fixed amount of study time) but by a signal from one of the "old" or "new" accumulators that it has reached a threshold.

Familiarity Calculation Familiarity is calculated as a likelihood ratio: the likelihood that the probe trace (given the current number of samples taken) and the memory trace encode the same event (an item and its associated context) versus the likelihood that the probe trace and memory trace encode different events (a different item or different context). Because encoding is probabilistic, if we assume that the memory system has some knowledge about the amount of noise in the system and about the possible feature values, a likelihood ratio is a natural way to represent the balance of evidence in favor of a match between the probe and the memory trace.

The probability of a feature matching between the two traces, given that the two traces encode the same item, is $\Pr(\text{Match}|\text{Same}) = c + (1-c)\frac{1}{2}$. That is, either the feature was encoded correctly or it was encoded incorrectly, but matches by chance. Similarly, the probability of a feature match between two traces that do *not* encode the same item is $\Pr(\text{Match}|\text{Different}) = \frac{1}{2}c + (1-c)\frac{1}{2} = \frac{1}{2}$; encoding might have been correct or incorrect, but in either case the match is purely by chance. Finally, $\Pr(\text{Mismatch}|\text{Same}) = (1-c)\frac{1}{2}$ (stored incorrectly *and* does not match by chance) and $\Pr(\text{Mismatch}|\text{Different}) = \frac{1}{2}c + (1-c)\frac{1}{2} = \frac{1}{2}$ (stored correctly and does not match by chance, or stored incorrectly and does not match by chance).

Because features are encoded independent of one another, we can directly multiply the probabilities of matches and mismatches for either the "same" or "different" alternatives. Letting N_m and N_n be the number of feature matches and mismatches, respectively, the final match value between a probe trace and a memory trace is given by:

$$\begin{aligned}\lambda &= \frac{\Pr(N_m, N_n|\text{Same})}{\Pr(N_m, N_n|\text{Different})}\\ &= \left[\frac{\Pr(\text{Match}|\text{Same})}{\Pr(\text{Match}|\text{Different})}\right]^{N_m} \left[\frac{\Pr(\text{Mismatch}|\text{Same})}{\Pr(\text{Mismatch}|\text{Different})}\right]^{N_n}\\ &= \left[\frac{c+(1-c)\frac{1}{2}}{\frac{1}{2}}\right]^{N_m} \left[\frac{(1-c)\frac{1}{2}}{\frac{1}{2}}\right]^{N_n}\\ &= (c+1)^{N_m}(1-c)^{N_n} \qquad (1)\end{aligned}$$

and any features where either the probe or the memory trace have nothing stored (i.e., no value was sampled for that feature) do not enter into the calculation.

Selection of Active Traces We have postulated that the probe activates only those memory traces that are similar enough to the probe. If for no other reason, such an assumption is warranted by the fact that the number of episodic traces in memory is virtually uncountable. Similar traces could include those formed during list study (a target trace matches well in both content and context; a foil matches only on context features) as well as some of those stored in memory prior to list study (i.e., from prior life history). For simplicity, we only consider history traces that are stored during prior occurrences *of the test item*, since although these fail to match well in context, they will match well on their content features.

The rule for activation is simple: There is an initial threshold θ for the familiarity value (as calculated in equation 1), and any trace exceeding that threshold is activated and takes part in further calculations. We assume that all N traces from the study list as well as K historical traces of the probe item are available to be activated. The choice of K is somewhat arbitrary, but should be fairly large if the probe item is something with which one is expected to have had prior experience (e.g., a word or picture of a common object). Here, we sim-

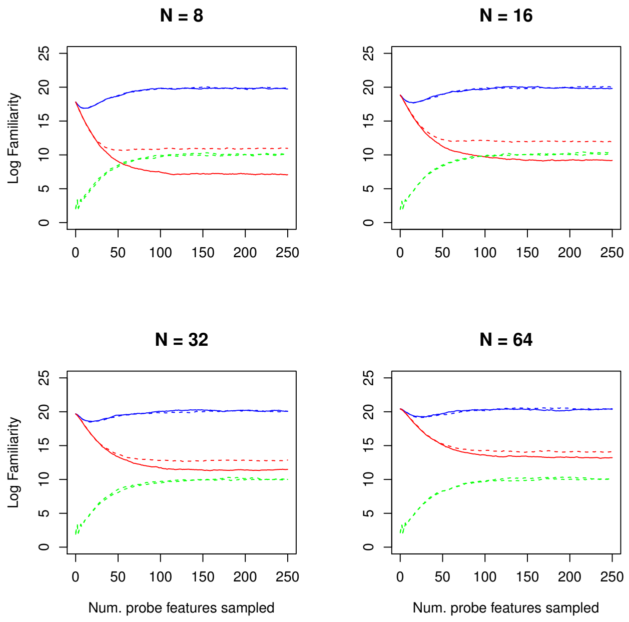

Figure 1: Familiarity over time as probe features are sampled, for a variety of study list lengths N. Profiles for target (studied) items are shown in blue, for foils in red. The cases where $K = 0$ (no available history traces) are shown in solid lines, while dashed lines depict $K = 200$. Green lines reflect the amount of activation from history—as opposed to list—traces. Averaged over 1000 simulations.

ply let $\theta = 1$, meaning that if the balance of evidence favors a potential match to the probe (which could be a result of matching item features, context features, or both) it is activated.

Computing Total Familiarity Computing the summed familiarity ϕ over the activated memory traces is straightforward, since a familiarity value was already computed for each trace to determine whether it passes threshold. To put it formally:

$$\phi = \sum_{i \in \text{Memory}} \begin{cases} \lambda_i & \text{if } \lambda_i > \theta \\ 0 & \text{if } \lambda_i \leq \theta \end{cases} \qquad (2)$$

where λ_i is the familiarity match of memory trace i to the probe trace, as given by equation 1.

Making a Recognition Decision Figure 1 shows the average time course of total familiarity (specifically, the natural logarithm of the result of equation 2) as more features are sampled and added to the probe trace. The "activation profiles" for both targets and foils across list lengths approach an asymptotic value, reflecting the saturation of the probe trace when all (or nearly all) features have a sampled value[1]. Note, however, that although target familiarity is greater than foil familiarity, the absolute values are not diagnostic on their own; they vary with list length, number of available history traces, and other factors to be discussed shortly, including study time, category size, and similarity. Simulations with fixed criteria operating on asymptotic familiarity (omitted due to space limitations) systematically fail to produce appropriate performance predictions.

The shape of the familiarity profiles for targets and foils are systematically different, regardless of list length or number of available history traces. Although the shape of the profiles arises in a complex fashion, the slope of target familiarity profiles is usually positive, while that of foils is predominantly negative. Thus, rather than make a recognition decision on the basis of raw familiarity, we can do so on the basis of this slope information. The slope of the log-familiarity[2] profile at time t, $\log \phi(t)$ can be estimated by taking the difference between times t and $(t-1)$, $\nabla[\log \phi(t)] = \log \phi(t) - \log \phi(t-1) = \log \frac{\phi(t)}{\phi(t-1)}$. We posit that this slope information accrues in two independent, racing accumulators: Positive values of $\nabla[\log \phi(t)]$ are added to an "old" accumulator, while negative values are added to a "new" accumulator. When either of these accumulators reaches its respective threshold, α_{old} or α_{new}, feature sampling stops and the corresponding decision is made.

Dynamics The process by which features are sequentially sampled from the test item to form a probe trace is presumed to be the same process at work when studying list items—a homogeneous Poisson process. Given that the decision procedure terminates after taking a certain number of feature samples from the test item (call this number of samples q), we can predict a trial's response time (RT) as a sample from a q-stage Gamma distribution with rate ρ:

$$f(RT; q, \rho) = \frac{\rho^q}{(q-1)!} RT^{q-1} e^{-\rho RT}.$$

Applications

In all of the following applications of the model, context is represented by 30 random binary features, the values of which are fixed across all study and test items. Each simulated data point reflects an average of 1000 simulations. Test lists are unbiased, consisting of an equal number of old and new items. No forgetting is posited, nor are any new traces formed during testing. These assumptions are made for the sake of simplicity, to better demonstrate the properties of the recognition mechanism, which is our primary focus. We also note that the model's parameters were chosen in a more-or-less arbitrary fashion, and were not fit to data; the qualitative trends shown here hold across a variety of parameter settings. History traces are created by the same encoding process as list traces (with study time $T_s = 2$), only the context features of each history trace are randomized.

[1] Because of probabilistic noise in the sampling process, this value will never remain completely stationary, since a sampled feature value might later be replaced with a different value.

[2] We work with the logarithm solely to transform the domain to all reals, but the logic remains the same if working with untransformed familiarity.

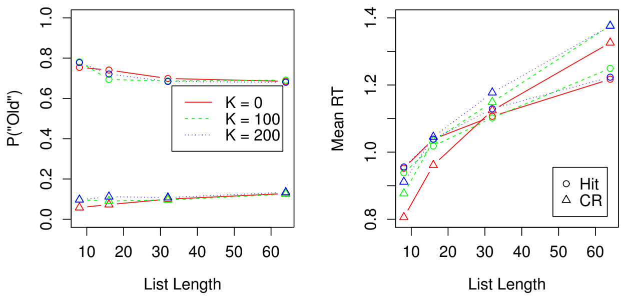

Figure 2: Simulation results for a variety of list lengths (N) and number of available history traces (K). The left plot shows hits (○) and FAs (△) while the right plot shows mean predicted RT (in arbitrary units) for hits and CRs.

List Length

Before addressing more complicated situations, we consider the most basic episodic recognition paradigm in which participants study a single list of items of the same stimulus type (e.g., words) one at a time for $T_s = 1$ time unit. Study of each item results in the encoding of a trace in long-term memory. The content features for each item are represented by 30 binary features, assigned randomly such that the items do not bear any systematic similarity to one another. At test, the content features of foil items are also assigned at random.

The model's predicted hit and FA rates, as well as mean correct RTs (hits and CRs), for varying list lengths are shown in Figure 2. In addition, we varied the number of history traces for each test item (K) that were available during recognition (i.e., the number that could contribute to the summed familiarity in equation 2). Across varying values of K, the model predicts a standard list length mirror effect—decreasing hit rate and increasing FA rate with list length. The model's predicted response time distributions also conform to what is usually found in recognition memory: increased overall correct RT with list length (e.g., Ratcliff & Murdock, 1976).

Mixed Lists

A particularly problematic situation for models with fixed decision criteria is the case where study and test lists are comprised of varying numbers of items of different types. Different item types may have different raw familiarity values by virtue of having many history traces (as, perhaps, with words or well-known objects) or by being studied for a longer time (an increase in memory "strength").

Varying Strength Study time is governed by the model parameter T_s, which reflects the amount of time available to draw samples from an item at study (at rate ρ). For this simulation, we created a study list with 40 items, twenty of which were studied for $T_s = .5$ time units, the others for $T_s = 2$ units. Further, ten items from each strength level were presumed to be "novel" ($K = 0$) while the other ten had been previously encountered in other contexts ($K = 200$). Foils could also be "novel" ($K = 0$) or previously seen ($K = 200$). Consistent with results on increased study time (Criss, 2006; Ratcliff & Murdock, 1976), more study is predicted to result in a higher hit rate and faster correct responses, regardless of the number of history traces available for activation, although a foil with history traces may be endorsed more readily than one without (see Figure 3).

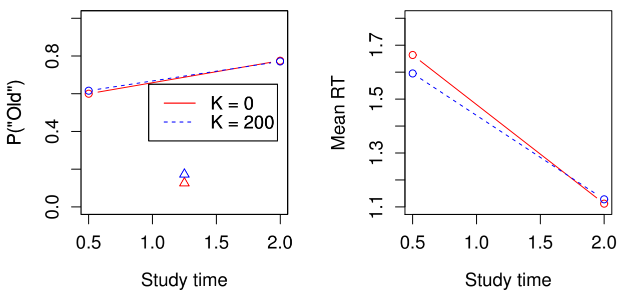

Figure 3: Simulation results for a list with varying study time (T_s) and history traces (K) per item. In the left plot, upper lines (○) are hits, lower lines (△) are FAs (which, strictly, have a study time of zero). The right plot shows mean model RTs for hits (in arbitrary units).

Varying Knowledge Finally, we turn to a version of the case described in the introduction in which a single list contains varying numbers of items of varying familiarity. The study list in this simulation consists of items drawn from eight categories, four of which are novel (and thus their items have no history traces, e.g., random dot patterns) and four of which are familiar (200 available history traces, e.g., everyday objects). The list itself contains a single exemplar from each of two of the novel categories and two of the familiar categories and eight exemplars from the remaining four categories (two novel and two familiar). Within-category similarity was modeled by randomly generating 30 binary features to serve as a category "prototype" while the features of each exemplar could be copied from the prototype with 50% probability, or assigned at random.

Once again, performance of the model (shown in Figure 4) is generally robust to the introduction of history traces. The model correctly predicts an increase in FA to categorically related foils with the number of studied exemplars with no concomitant rise in unrelated FAs or fall in hit rate (Dennis & Chapman, 2010; Shiffrin, Huber, & Marinelli, 1995). The presence of history traces has its main effect on RT, where more history traces result in faster hits and slower FAs (as the presence of history traces tends to flatten the familiarity profile for foils, as can be seen in Figure 1).

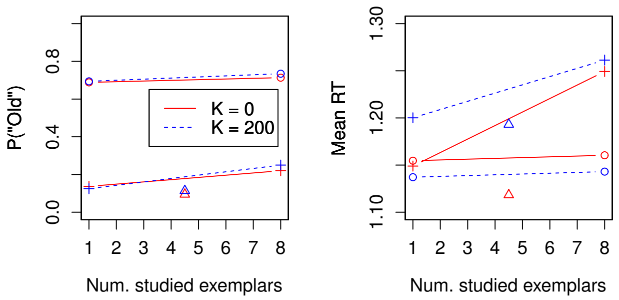

Figure 4: Simulation results for a list with varying numbers of category exemplars and history traces (K). In the left plot, upper lines (○) are hits, lower lines are FAs to unrelated (△) and related (+) foils. In the right plot, ○'s show mean RT for hits, +'s show mean RT for CRs of related foils, △'s show mean RT for CRs of unrelated foils. RTs shown in arbitrary units.

Discussion

The recognition memory model presented in this paper is a proof of concept that sensible recognition decisions are possible even with variations in task, study factors, stimuli, and various other factors that affect overall familiarity of a test probe. Using two racing accumulators to make a decision provides a way to take advantage of the expected difference in the time course profile of familiarity growth for targets and foils: Noting that the successive familiarity changes tend to be positive for targets and negative for foils, we let one accumulator add up positive changes, and the other negative changes. The present results show that such an approach has promise to solve a long time puzzle in recognition memory: how recognition can occur when tasks and stimulus types vary widely and there is insufficient experience with any one type to learn appropriate criteria. Traditional models treat this problem by adjusting criteria as any of these variables change. Our approach still involves criteria (α_{old} and α_{new}) but these do not need to change across stimulus types in order to produce reasonable predictions. In a sense, this can be seen as re-framing the problem of criterion setting such that items themselves can serve as their own reference, rather than relying on assumed distributions of absolute familiarity.

We believe the approach presented here can shed light on some recent results that apparently demonstrate criteria adjustments that are problematic for many existing memory models (e.g., Dennis & Chapman, 2010; Starns, White, & Ratcliff, 2010), although we have yet to apply our model directly to their data. Our preliminary results also suggest that the source of noise may change during recognition, with other list items serving as the primary source of noise early in recognition, while historical traces of the test item intrude later in the process (note the rising activation of history traces in Figure 1). This possibility could be examined within a signal-to-respond paradigm that interrupts recognition at various points (as in Brockdorff & Lamberts, 2000; Hintzman & Curran, 1994). Finally, the ability of our model to make simultaneous predictions about both accuracy and RT underscores the need for future research into the dynamics of long-term recognition memory, so as to place tighter constraints on theory development. While the model presented in this paper is clearly preliminary, we hope that it may suggest new avenues of research and new ways of conceptualizing problems in recognition memory.

References

Brockdorff, N., & Lamberts, K. (2000). A feature-sampling account of the time course of old-new recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 77–102.

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

Dennis, S., & Chapman, A. (2010). The inverse list length effect: A challenge for pure exemplar models of recognition memory. *Journal of Memory and Language*, *63*(3), 416–424.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(2), 452–478.

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*(2), 414–435.

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

Hintzman, D. L. (1988). Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, *33*, 1–18.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*, 335–384.

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 322–336.

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), 724–760.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(3), 609–626.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190–214.

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*(2), 267–287.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM–retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language*, *63*(1), 18–34.

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