ELSEVIER

Cognitive Science 27 (2003) 341–378

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Modeling memory and perception

Richard M. Shiffrin*

Psychology Department, Indiana University, Bloomington, IN 47405, USA

Received 21 May 2002; received in revised form 26 December 2002; accepted 28 February 2003

Abstract

I present a framework for modeling memory, retrieval, perception, and their interactions. Recent versions of the models were inspired by Bayesian induction: We chose models that make optimal decisions conditioned on a memory/perceptual system with inherently noisy storage and retrieval. The resultant models are, fortunately, largely consistent with my models dating back to the 1960s, and are therefore natural successors. My recent articles have presented simplified models in order to focus on particular applications. This article takes a larger perspective and places the individual models in a more global framework. I will discuss (1) the storage of episodic traces, the accumulation of these into knowledge (e.g., lexical/semantic traces in the case of words), and the changes in knowledge caused by learning; (2) the retrieval of information from episodic memory and from general knowledge; (3) decisions concerning storage, retrieval, and responding. Examples of applications include episodic recognition and cued and free recall, perceptual identification (naming, yes–no and forced-choice), lexical decision, and long-term and short-term priming.

Keywords: Bayesian techniques; Likelihood ratio; One-shot-of-context hypothesis

This article will focus on my models in their current incarnation, but I will also try to indicate how they emerged from a research program that now extends over 30 years. My research group has made much progress recently with a Bayesian approach, but I would not want to mislead readers into thinking I favor axiomatic approaches and derivations over data-determined modeling. Let me begin therefore with some meta-science perspectives. None of the models we use in psychology or cognitive science, at least for any behavioral tasks I find to be of any interest, are correct. We build models to increase our understanding of, and to slightly better approximate, the incredibly complex cognitive systems that determine behavior. We hope that

* Tel.: +1-812-855-4972; fax: +1-812-855-4691.
E-mail address: shiffrin@indiana.edu (R.M. Shiffrin).

0364-0213/03/$ – see front matter
doi:10.1016/S0364-0213(03)00027-2

even the relatively crude approximations to truth represented by present-day modeling will have utility in real world settings and applications, and will increasingly prove useful for predicting future results. Our models are developed first and foremost on the basis of experimental results, but not on this basis only, because we are drowning in data of varying reliability, and because the space of models is infinite and includes models of great complexity capable of predicting given data sets (almost) perfectly. Model development is therefore a complex process supported by vague intuition, selected experimental results, introspections, sub-models aimed to describe partially lawful empirical relations, testing of simple hypotheses and heuristics, clinical observations and tests carried out on such populations, neural experimentation, measurement of brain activity over time and by region, and considerations of consistency, and parsimony. We have found it critical to formulate each model in ways that allow formal derivations and quantitative predictions; because these have inevitably increased in complexity as our science progresses, we have increasingly found it necessary to instantiate them in computational terms and to carry out derivations by computer simulation. Depending on the purpose and application, the formal models are used to produce either qualitative or quantitative predictions.

Given the fact that data is the most important factor in model development, and given the increases in the amount and complexity of such data and the associated models, our recent use of Bayesian techniques to develop models is almost a throwback to earlier times. In addition, it has the flavor of an attempt to engineer the human cognitive system to conform to certain ideal prescriptions. Yet the approach has had remarkable successes in generating models that predict existing empirical findings, and in a few remarkable cases, making correct *a priori* predictions of data that could not have been intuited or anticipated. Thus, although this article is by no means an evangelical call to Bayesian modeling, I will be sure to emphasize the successes we have obtained with this approach.

In this article, I attempt to provide insight into storage and retrieval systems for human memory through examples of models we have applied to a variety of tasks. Space limitations prevent coverage of the enormous amount of related research by others, so I will focus almost exclusively on research recently published and in press for which I have been a co-author, and citations to my own earlier research. I will describe models, and fits of these models to data, for the storage in and retrieval from general knowledge and episodic memory, for the development of knowledge from experience, and for the influence of recent events on retrieval of well-learned knowledge. The models are tailored carefully to the demands and constraints of particular tasks, but have a common theme: Retrieval is based on the matching of probes to memory traces, and the evidence that a memory probe matches a particular stored memory trace is calculated as a likelihood ratio. Given that many of the recent articles of this type that I have co-authored are much longer than this report, some unpleasant exclusions and compromises have had to be made: In particular, almost all technical detail has been omitted, in favor of providing a conceptual overview. The reader interested in the more rigorous aspects of formal modeling is referred to the individual articles; these are generally replete with equations and modeling details.

The Bayesian approach is similar to ‘ideal observer theory’ (e.g., Green & Swets, 1966). It asks what model would produce optimal performance in a given task, conditional on external limitations upon the quality of data arriving from the environment and on internal limitations

of storage, retrieval, and decision-making. In our applications, the data from the environment is usually well specified, but internal limitations need to be assumed. We have assumed limitations on attention (e.g., number of type of features maintained in short-term memory and used as probe cues), storage ability (e.g., features are not always stored correctly), retrieval of knowledge (e.g., not all the information in long-term memory can be accessed in usable form at any given moment), perception (e.g., a variety of mechanisms limit the rate and quality of information processed from the environment) and feature selection and attribution (e.g., the features representing a perception or a cue are not only incomplete and error prone, but often contain irrelevant features that 'leak' from other sources in the environment nearby in time or space). These limitations are instantiated quantitatively, and we then ask the following question: If the memory probe (limited by any or all the mechanisms mentioned) is compared to all the information in long- and short-term memory (limited by any or all the mechanisms mentioned), what would be the calculations and decisions that would optimize performance? To make such an approach work in a given task setting, a good number of other implicit and explicit assumptions are required, whose choice goes beyond the Bayesian approach; motivation for these additional assumptions includes prior research in the field, neuroscience findings, consistency of assumptions within and between tasks, generalizability to potential other tasks, simplicity, and predictability.

Generally speaking, Bayesian approaches have become omnipresent in science and society in the latter part of the twentieth century, probably because whenever it is possible to measure success reliably, such approaches work better than the alternatives. Because success is harder to measure when developing theories in psychology, Bayesian approaches have caught on at a slower rate. Nevertheless, it is not hard to find examples of such approaches: They have been used with great success in sensory psychology since the 1950s, and more recently were the subject of a book on cognition by Anderson (1990). Recent years have seen an accelerating use of such approaches, one example of which is illustrated in the work of my research group over the last several years.

1. The REM framework for memory storage and retrieval

1.1. Overview

Events and memory traces are represented as vectors of feature values. The traces are stored separately, although some are formed through accumulation of multiple events. Traces of new events are stored in memory as incomplete and error prone vectors. Such traces are termed episodic, and contain context, physical, and meaning features. When an event is similar to one or more earlier events (e.g., a nominal repetition is taking place) there is a tendency to store a new episodic trace, but also a tendency to store additional information in the previously stored trace. Thus, similar events occurring at different times tend to accumulate in one trace, eventually producing a much larger, more complete, and more accurate trace termed lexical/semantic, or, more generally, knowledge.

Retrieval starts with the matching of a probe cue, again represented as a vector of feature values, to all memory traces in parallel, although only traces sufficiently similar are activated

(i.e., take part in further calculations or mental operations). Generally, closely matching lexical/semantic traces will be activated first (i.e., knowledge retrieval), and weaker episodic traces will be activated later, perhaps after some features from the lexical/semantic trace are added to the probe (i.e., explicit or episodic retrieval).

Matching of probe to a given trace consists of comparing corresponding feature values, and using the results to calculate a likelihood ratio: the probability that the trace is a copy of an earlier presentation of the present probe, divided by the probability that the trace is a copy of an earlier presentation of something other than the present probe. These likelihood ratios are the building blocks of retrieval. They are used to recognize (by averaging the likelihood ratios), to recall (by sampling traces in proportion to the likelihood ratios), and to categorize and recognize features (by averaging likelihood ratios across subsets of traces containing a specified feature value or values). Bayesian reasoning provides the justification for such retrieval operations.

Memory tasks sometimes require retrieval of recent events, and are termed explicit. Other tasks require retrieval of well known information, regardless of the recency or context of storage; this is termed knowledge retrieval. Storage of an event often affects later retrieval of knowledge involving that event (referred to as 'implicit memory,' and 'priming') even when the original event does not seem to be recallable explicitly. The term implicit is used because the occurrence of a recent event is demonstrated indirectly, through its effect on knowledge retrieval. REM predicts many implicit memory effects on the basis of its storage rules: Study of a prime causes not only storage of the prime in episodic traces, but also adds to the lexical/semantic trace of the prime new information such as current context. On a later test of general knowledge involving the prime, an increase in likelihood ratio is caused by the matching of context in the present environment to context recently added to the prime's lexical/semantic trace.

As I lay out the assumptions of the REM modeling framework, it will become clear that many of these are derived from earlier modeling efforts rather than Bayesian induction. The contribution from the Bayesian perspective is nonetheless quite important, providing the justification for the use of likelihood calculations, and leading to formal decision rules and calculations for these, in many task settings. The reader is referred to Shiffrin and Steyvers (1997) for the details of the Bayesian induction that provides the basic justification.

1.2. The assumptions of the REM framework

1.2.1. Short- and long-term memory

As in Atkinson and Shiffrin (1968), the control of the cognitive system resides in short-term memory. Long-term memory is a relatively permanent and passive repository for information (possibly corresponding to some base rate level of neural activity). Short-term memory is the active memory system (possibly corresponding to heightened levels of neural activity, for example as measured by studies using evoked potentials or functional magnetic resonance imaging). Short-term memory includes sensory memories with short residence times, and longer residence phonological and working memories. A variety of attention effects are governed in this system: Expectations concerning coming events are set in place here. Sensory information about new events enters short-term memory where attention selects attended portions for longer residence and further processing. In short-term memory, attention is used to generate probes for long-term retrieval, and used to select from the information retrieved. To

give just one common example, if a word is presented for study, its sensory/physical features enter short-term memory, join current context there, and an attended subset of both types of features is used to access the word's lexical/semantic representation in long-term memory. Of the semantic features retrieved from the long-term trace, attention governs which ones join the other features in short-term memory. Attention then selects certain of these features for further cognitive operations, such as episodic storage, or retrieval.

1.2.2. Event representation and features

Experience is continuous but is parsed into separate events by the operation of attention in short-term memory (by a process that is assumed but not yet specified in any of our models). The assumption that traces are stored in functionally separate fashion is not arbitrary; one source of evidence will be discussed shortly: the 'list-strength effect' first noted by Ratcliff, Clark, and Shiffrin (1990) and Shiffrin, Ratcliff, and Clark (1990).

Events in REM are represented somewhat neutrally as vectors of independent feature values. The number of feature values in an event is in principal flexible, within limits governed by attention and by attentional capacity limits. The positions of features are self-coded, so the memory traces of two events (say, one a probe cue, and one a trace) may be aligned: Corresponding feature values can be and are compared. In most of our work, the features and values are arbitrary and abstract, with no ties to features in the real world; at the end of this article I will mention some recent work by Steyvers that attempts to choose features and feature values on the basis of analysis of empirical data, and therefore links the assumed features to reality. For the main body of this article, however, the features are abstract representations.

There are two ways we have used to represent the values of features. The more natural method is probably to use continuous values; in one such scheme the population mean is (arbitrarily) set to zero, the population distribution (arbitrarily) assumed to be gaussian, and the population variance (arbitrarily) set to 1. In such a representation, the distance of a feature value from zero represents the base rate of that particular value in the population. Any given event consists of a particular sample of features and values of those features from the population. The other method (used most often because it is easier to simulate), uses discrete values. Unless specified otherwise, this article discusses research using the discrete representation.

In our discrete feature representation, the possible values (V) are zero and positive integers. 'Zero' represents no information stored. Positive integers have no meaning other than to code base rates in the population: '1' is most common, '2' less common, and so on. (We have used a geometric distribution, but this quantitative detail is not critical.)

Using the discrete representation, an illustration of a vector representing an event is given in Row A of Fig. 1 (let us imagine the event is presentation of a visual word). The zeros indicate features not relevant, and the positive values indicate the values of the features. There are more small integers, like 1, because these are the highest probability values in the population. The features are grouped into regions of visual form and context, for illustrative purposes. Of course there are too many features to attend in short-term memory. Row B depicts the smaller number of the features in the event that are attended, with zeros replacing those unattended. (Often it is convenient to delete some or all of the unattended and irrelevant features, so that the event is represented with few or no zeros.) Which features in an event are attended evolve over time within a brief study interval, even with no additional external stimuli, due in part to the

retrieval of additional information during study from lexical/semantic and episodic memory, but Row B shows the representation in short-term memory before this occurs.

The features in an event trace depend, of course, on the type of event (an auditory screech representing a bird cry will not be represented by features like those used to represent a picture of a horse). Thus, the physical and semantic features will differ across event types. However, it is key assumption of the framework that context is a part of all events and traces of those events. Context refers to those elements of the event that are not experimentally manipulated (and usually are not the explicit object of attention), including elements like one's internal physical signals, one's feelings and emotions, the locale and setting, the physical surround, and the situation.

In the remainder of this article, unless specified differently, we will assume for concreteness of exposition that that the event being discussed is the presentation of a word on a list presented visually. Across lifetime experiences, knowledge about a particular word comes to be stored in a trace that we term lexical/semantic. Such a trace contains cumulative information about all the types of features in the attentive surround of each occurrence of that word, particularly including meaning and context. Row C of Fig. 1 depicts such a lexical/semantic trace. This tends to be a very large vector, containing much more information than can be attended and placed in short-term memory at any one moment.

The features already in short-term memory before the word presentation, such as context, combine with an attended subset of those initially provided by the study or test presentation,

REM Representation: Assume presentation of a word.

'0' represents nothing stored about a feature
Smaller integers represent higher environmental probabilities

A: Vector of context and sensory features for presented word
B: Attended vector of features held in short-term memory
C: The presented word's lexical/semantic vector in long-term memory.
D: The vector in short-term memory after retrieval of meaning features
E: Vector D stored in long-term memory as an episodic trace
F: Vector C after new features from D have been added

Visual Form Features	Context Features	Meaning Features

A: <..4,1,1,2,1,0,3,2,5,3,1,6,1.....1,2,0,1,1,4,5,1,1.....0,0,0,0,0,0,0,0,0,0....>

B: <..4,0,0,2,1,0,0,2,5,0,1,6,0.....1,0,0,1,0,0,5,0,1.....0,0,0,0,0,0,0,0,0,0....>

C: <..4,3,1,0,4,1,2,2,0,3,1,6,1.....3,1,1,0,3,5,0,2,5.....4,3,1,1,4,2,1,4,1,1....>

D: <..4,0,0,2,1,0,0,2,5,0,1,6,0.....1,0,0,1,0,0,5,0,1.....4,0,1,1,0,0,1,0,1,0....>

E: <..4,0,0,1,1,0,0,0,0,0,1,0,0.....1,0,0,1,0,0,4,0,0.....4,0,4,0,0,0,0,0,1,0....>

F: <..4,3,1,2,4,1,2,2,5,3,1,6,1.....3,1,1,1,3,5,0,2,5.....4,3,1,1,4,2,1,4,1,1....>

Fig. 1. Events and memory traces are represented as vectors of feature values (see text for discussion).

such as shape, to form a probe used to access long-term memory (the probe is depicted in Row B of Fig. 1). These features initiate retrieval from the corresponding lexical/semantic trace (i.e., Row B of Fig. 1 is compared to all memory traces, quickly converging on the correct lexical/semantic trace, represented by Row C); additional information is then retrieved from the lexical/semantic trace, the most important of which is meaning. This information from the lexicon is filtered by attention and combined with the presented information to produce another memory probe that includes meaning (depicted in Row D of Fig. 1). This set of features is then stored (see the mechanisms listed next) and matched to the episodic traces in memory (critical for explicit memory tasks).

1.2.3. Memory storage

Assume the system has retrieved from the lexicon, has selected a subset of context, meaning, and form features via attention, and is now ready to store the resultant event (i.e., is ready to store Row D of Fig. 1).

It is assumed that two kinds of storage can take place, episodic and lexical/semantic. Episodic storage involves an attempt to copy the study event (including its present context) into memory as a separate trace. This may occur through formation of a new trace, or addition of features to a very similar previously stored episodic trace (as might be the case if the present word had occurred earlier in the same list). The second of these episodic storage processes adds information to an item's episodic trace, much as if extra study time had occurred.

One effect of adding information to a trace is the *differentiation* of that trace from episodic traces of other words. If two traces encode events that are not very similar, then each feature added to one trace tends to increase the number of features that differ between the two. Thus, differentiation decreases the similarity of traces of different words. The decreased similarity between traces of list-words means that traces of list-words other than the test word are activated less strongly, and thereby cause less confusion, improving performance.

Differentiation and separate trace storage are the core of the SAM and REM accounts of the list-strength findings of Ratcliff et al. (1990). Our account built upon the facts that adding items to a list decreases recognition performance (the list-length effect), and repeating an item in a list improves its recognition (the strength effect). Models of these two common effects generally predict that repeating an item in a list would harm recognition of other list items (even more so than would adding different list items). However, we found that repeating some list items slightly improves recognition of other list items (the null list-strength finding). The separate storage account predicts these findings because strengthening an item differentiates it from other items. It has proven quite difficult to predict such findings with any model that merges different events into a common composite memory, probably because differentiation is very hard to implement within such a system. We have therefore used the list-strength results as evidence for separate event representation.

Storage occurs incompletely and with error. It is a key assumption of the REM approach that stored values normally do not change once stored. The exception occurs when attention is directed toward a feature value that feedback or other information reveals is stored incorrectly. Assuming a given feature is not yet stored in a given trace, what is stored depends on the nature and type of attention—i.e., strategies of encoding. Some storage occurs relatively automatically at the start of study of an event, and additional storage depends on which features are given

attention. In particular, Malmberg and Shiffrin (in press) demonstrated for word storage that context features tend to be stored automatically during the first one to two seconds of study, whereas semantic features continue to be stored during the full course of rehearsal—this was termed the *one-shot-of-context* hypothesis. Whether this storage pattern is due to a strategic decision awaits further research.

Let us suppose that at some moment a trace does not have a feature value stored, and an attempt is made to store in that trace the value present in the current study event. The probability of copying the feature value correctly is c; with probability $1 - c$, a random value is stored according to Eq. (1). Thus, after some period of study, an episodic trace will have been stored having zeros in some vector positions, and correct and incorrect values in the remaining vector positions. This is illustrated by the vector in Row E of Fig. 1 (copied incompletely and partially correctly from Row D). Longer study time or massed repetitions would produce more non-zero feature values in this trace. Spaced repetitions of a given word within a study list would sometimes produce a separate representation for two or more of the repetitions, but often would act as additional study time, increasing the number of non-zero feature values stored in one episodic trace.

In addition to episodic storage, features are stored in an item's lexical/semantic trace. This type of storage probably occurs more automatically and less subject to strategies. It is not yet known whether storage in such traces occurs at the time that the lexical/semantic trace is first accessed, later when episodic traces are stored, or at both times. Because an established lexical/semantic trace is rich with previously stored feature values (including both context features and semantic features), potential storage is limited to features that are in the current study event but are not already stored in the lexical/semantic trace. Such 'new' features tend to consist of a portion of the current context features, and some physical features unique to the present event (such as presentation of a word in an unusual visual font). The result is illustrated by Row F of Fig. 1, which is a copy of Row C plus some current context from D.

To preview the applications to implicit memory, the facts that only current context features tend to be stored in an item's lexical/semantic trace, and that both current context and semantic features tend to be stored in an item's episodic trace, produce a wide variety of dissociations that are commonly observed between explicit and implicit memory. To take one example, suppose a list of words is studied at time t_1. Later, at time t_2, a general knowledge test is given: Words presented visually are to be named as rapidly as possible. At t_1, current context is added to the word's lexical semantic trace. At t_2, both the visual test word and current context are used to probe lexical/semantic memory. The current context in the probe tends to match the current context that had been added to the lexical/semantic trace at study, speeding responding for studied words. On the other hand, studying words at a 'deeper' level (e.g., 'pleasantness' rating versus 'upper or lower case' rating) increases the semantic features added to the episodic trace, thereby improving explicit memory performance, but does not add semantic features to the lexical/semantic trace, and therefore does not speed naming time. Another prediction is based on the 'one-shot-of-context' results: Longer study time produces more semantic features in the episodic trace, improving explicit memory performance, but does not produce more current context features (in either the episodic or lexical/semantic trace) and hence speed of naming would not be related to the duration of study (see Malmberg & Shiffrin, in press).

Next consider how lexical/semantic traces come to be formed. When an item is first encountered during development, an episodic trace is formed. When next encountered, another episodic trace can be stored, information can be added to the previous episodic trace (when the two events are in close temporally and contextually), or both (when the trace of the earlier event is called to mind by the current presentation). This process continues over development, so that eventually memory is filled with many separate episodic traces for an item/event, and also contains a particular trace that has accumulated information from many of these developmental occurrences and therefore comes to contain many features. The latter trace is what is termed the lexical/semantic trace. It is important to note that the lexical/semantic trace contains not only semantic information, but also contexts accumulated over the lifetime of encounters. When such a trace is accessed, it therefore seems to be associated with no particular context, not because context information is not present, but because so many contexts are present that no one dominates. In fact, it is critical that context information does get added to lexical/semantic traces because in REM such added storage is the basis for many findings of implicit memory, particularly priming.

I am often asked whether it is necessary to assume separate lexical/semantic traces. Would it not be possible to assume only episodes, and let these be accumulated at the time of retrieval? Perhaps someone could formulate such a model, but this approach could not be used in the present framework without introducing conceptual and empirical inconsistencies. For example, explanations of the list-strength effect required us to assume traces accumulate in one stronger trace. Such accumulation if extrapolated over development seems to require the eventual formation of a lexical/semantic trace.

Finally, consider the rules governing whether storage of a current event occurs in a new episodic trace, or instead occurs by information addition to an already existing episodic trace of the same item. In particular, if words are repeated during one list, even at spaced intervals, what determines whether and how often are separate episodic traces formed for each occurrence, as contrasted with accumulation in one trace? Presumably, what happens depends among other factors on the degree to which it is noticed during study that the present item has occurred earlier in the list, and the previous trace recalled. The balance of these two storage possibilities is critical when predicting list-strength results (e.g., Shiffrin et al., 1990; Shiffrin & Steyvers, 1997; Malmberg & Shiffrin, in press). Storage accumulation in one trace makes that trace fill with more feature values, and thereby differentiates it from traces of other words (more feature values tend to differ)—this process improves memory for tests of other words. Formation of new traces means that each such trace has few features, and therefore tends to be less differentiated from the features of traces of other words—the process of storing many separate traces of a repeated word therefore harms memory for tests of other words. Because the list-strength results for recognition show that memory for a given word is slightly helped by additional spaced repetitions of other words, it is necessary in the REM framework that relatively few within-list repetitions form their own separate traces. Although we have described this implication in REM terms, it appears to hold for a large class of models.

It is interesting that Raaijmakers (in press) has shown in his article in this issue that basically these same conclusions concerning storage in separate traces versus storage in one trace allows an elegant analysis of spacing effects. In particular, the analysis allows one to understand heretofore mysterious findings concerning differences in spacing effects for: (1) retrieval of repeated words and (2) for retrieval of at least one of two different words each shown once.

1.2.4. Retrieval

Retrieval begins with a probe cue, generally consisting of some content and context information, and represented as a vector of feature values. The probe vector is compared in parallel to all traces in memory, but a threshold process eliminates most of the traces in memory from consideration: Only those that are sufficiently strong and sufficiently similar to the probe take part in further retrieval activity. This thresholding process is quite important, even though it is often omitted explicitly in applications by allowing only the traces in some relevant set, such as the recent list, to take part in calculations. For a variety of reasons, some quite fundamental, the Bayesian approach breaks down and becomes unusable if one assumes that all of the (perhaps trillions of) traces in memory (many very weak and many dissimilar to the probe) take part in likelihood calculations.

When the probe contains information matching that in a lexical/semantic trace (e.g., a word is presented), the lexical/semantic trace is accessed rapidly and with extremely high probability and information found there retrieved. If the task is one of retrieval from episodic memory, (some of) this retrieved lexical/semantic information is added to the probe and used for comparisons with the episodic traces. The focusing of episodic retrieval upon episodic traces could be accomplished by an emphasis upon context features, or alternatively, through temporary inhibition of the lexical/semantic trace after it has been retrieved (possibly determined by a process of synaptic depression), or both.

To carry out the Bayesian induction we took as basic data the set of feature matches and mismatches between the probe and every (above threshold) trace in memory. The comparison of probe to a given trace (either lexical/semantic or episodic) is calculated as a likelihood ratio, λ_j, where j represents the jth trace. The λ_j gives the probability that the jth trace had been stored as a result of study of an event containing the present probe, divided by the probability that the jth trace had been stored as a result of study of some other event. The λ_j is calculated by multiplying probability ratios for each feature position for which both probe and trace have non-zero values. This process is illustrated in Fig. 2 (the steps are illustrated in this figure, but for the justification the reader is referred to Shiffrin & Steyvers, 1997). A vector position with matching values contributes a ratio greater than one to the product. Because larger integers represent feature values with lower environmental base rates, larger integers produce large ratios. A vector position with mismatching values contributes a ratio less than one to the product (and this ratio does not depend on the values themselves). A trace of the probe tends to have many matching values, and hence a λ_j greater than 1.0; a trace of something else tends to have many mismatching values and hence a λ_j less than 1.0. These λ_j values assigned to the traces may be thought of as activation values, similarity values, or matching values. The Bayesian analysis provides the justification for the calculation and use of likelihood ratios, because it provides a basis for conditionally optimal decision making, as described below for episodic recognition.

1.2.5. Neural implementation of Bayesian calculations

I would not want to claim that our neural machinery actually calculates likelihood ratios in the way suggested by Bayesian induction. However, it is plausible that pressures to adapt behavior to optimize performance would have led to systems that approximate such decision rules. Such systems would adapt to take advantage of regularities in the environment, a common

Trace j: <3, 0, 0, 1, 4, 0, 5, 1, 0, 2,.....>
Probe: <3, 3, 1, 1, 2, 1, 5, 3, 1, 2,.....>
↓ ↓ ↓ ↓ ↓ ↓
Ratios: r_1 r_4 r_5 r_7 r_8 r_9 ....
>1 >1 <1 >1 <1 >1

λ_j = P(trace j matches probe)/P(trace j mismatches probe)

$= (r_1)(r_4)(r_5)(r_7)(r_8)(r_9)$

P(r given the value of match is j) = $[c + (1-c)g(1-g)^{j-1}]/[g(1-g)^{j-1}] > 1.0$
P(r given mismatch) = $(1-c) < 1.0$

c is the probability that a feature is copied correctly when a feature value is stored.

g determines the base rate distribution of feature values:
$P(\text{value } j) = g(1-g)^{j-1}$

NOTE: The average likelihood ratio times the prior odds gives the odds that the trace of the test probe is in the set of activated traces:

$$\Phi = \Phi_0(1/n)\Sigma\lambda_j$$

NOTE: The REM model for recognition assumes an 'old' response is given when the odds is greater than 1.0. (When the numbers of new and old items tested are equal, the prior odds is assumed to be 1).

Fig. 2. Calculation of likelihood ratio from the comparison of probe to trace vectors. The likelihood ratio gives the probability that trace *j* was stored when the item in the probe had been studied divided by the probability that trace *j* was stored when some other item had been studied. Bayesian induction tells us that the odds the test probe is in the activated set of traces is just the prior odds times the average of these likelihood ratios. Shiffrin and Steyvers (1997) proposed use of the odds as a model of recognition.

theme in recent years (see the article by Geisler, in press, in this issue). Of course, when such systems are applied to situations in which the regularities are violated, they can lead to peculiar results. Although I focus on cases where the Bayesian approach takes advantage of regularities, some very interesting results illustrate how the same approach can predict diagnostic failures. To mention just one example, certain visual illusions can be understood in Bayesian terms as a result of rational inferences (Geisler & Kersten, 2002).

1.2.6. Quantitative and qualitative modeling

For the very simplest REM models applied to simple data sets it is possible to understand the structure of the model and the tasks well enough to allow qualitatively correct intuitive predictions to be made, and to provide confidence that quantitatively correct predictions could be obtained. The simplest model of episodic recognition in Shiffrin and Steyvers (1997)

provides one example, although simulations were provided to bolster the qualitative intuitions. For more complex models applied to more complex sets of data, quantitative fits to data become critical as an existence proof of model 'sufficiency.' (I place sufficiency in quotes because in almost all cases the quantitative fit allows the model to be rejected on strict statistical grounds; the sufficiency of models is a complex process going well beyond statistical accuracy.) There are those who argue that quantitative fits are not useful, believing that it is always possible to add assumptions and parameters to achieve success. It is my experience, however, that such a shotgun approach seldom makes it possible to produce an acceptable fit of a poorly chosen theory to data, especially if the theorist demands that the conceptual assumptions of the theory are integrated, coherent, and potentially generalizable to other tasks. Thus, quantitative fits have an important role, and this article will for this reason display several quantitative fits. Although it is desirable to establish sufficiency with an acceptable quantitative fit, it is usually a necessity of quantitative modeling that a number of auxiliary assumptions be imposed. In a well-conceived model, as we hope is the case with REM, these auxiliary assumptions ought to be much less important than the basic conceptual structure. In the text of this article I have therefore chosen to emphasize the conceptual structure and coherence of the model, rather than details of the fits to data.

2. Episodic memory retrieval

2.1. Recognition tasks

In Shiffrin and Steyvers (1997), we made a number of simplifying assumptions enabling Bayesian analysis to be applied to single-item episodic recognition—i.e., presentation of a list of words followed by a list of test words, half of which were on the list and half new (a very similar episodic recognition model was developed in parallel by McClelland & Chappell, 1998). It was assumed that all study and test words were randomly generated vectors chosen according to a geometric distribution with parameter g (as specified in Fig. 2). In a process not explicitly modeled, it was assumed that a first retrieval operation using only list context features served to activate all the list traces, and no others. The key assumption stipulated a parallel comparison of the content features in the test probe (both low level visual features and meaning) to each of the list traces. The optimal decision is found by applying Bayes theorem: Let the observed data, D, be all the matches and mismatches found from the comparison of test probe to all the activated traces in memory. We want the odds of 'old,' and calculate this by dividing the probability of D given the probe was indeed 'old' by the probability of D given that the probe was indeed 'new.' Bayesian analysis shows that this odds is just the average of the λ_j values for each list trace. Fig. 2 diagrams the process (justification is found in Shiffrin & Steyvers, 1997). Thus, the optimal decision (everything else being equal) is to respond 'old' if the average likelihood ratio is greater than 1.0.

This application of REM has a natural mirror effect (e.g., Glanzer & Adams, 1990), because 1.0 acts as the default odds for the system: For many factors that can be varied to improve performance, the hit rate (p['old'|old], or p(H)) rises and the false alarm rate (p['old'|new], or p(F)) falls. Fig. 3 depicts data and predictions of REM for three factors that affect performance

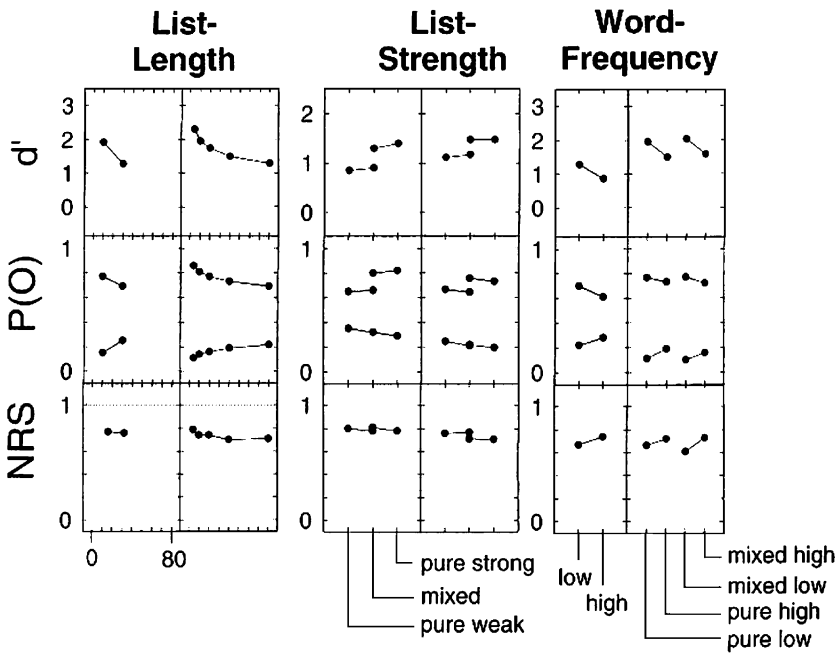

Fig. 3. Predictions of REM (right side of columns) for episodic recognition data (left side of columns). The top row gives recognition performance measured as d'; these are calculated from the hits (top) and false alarms (bottom) in the second row. The bottom row gives the slope of the normal ROC curve, termed NRC (bottom row)—NRC less than 1.0 indicates a higher variance of odds for targets than foils. The length, strength, and frequency manipulations are described in the text. The predictions were based on the following REM values and parameters: vector length = 20; probability of storage on each storage attempt = .04; 10 storage attempts for strong words and 7 for weak words; g_H = .45 for constructing high frequency words, g_L = .325 for constructing low frequency words; g = .40 for retrieval calculations; c = .7 (from Shiffrin & Steyvers, 1997).

(measured by d'), and the changes in p(H) and p(F) that give rise to those performance levels. The left panel depicts list length. The middle panel depicts strength (comparing pure lists of constant strength) and list-strength (comparing words of equal strength between pure lists and lists with words of other strengths). The right panel depicts the effects of natural language word frequency (in pure and mixed frequency lists).

Why does REM make these predictions? The proper analysis is couched in signal detection terms, based on one distribution of odds when a target is tested, another distribution of odds when a foil is tested, and a decision criterion (in the simplest REM case at 1.0). For length predictions note that extra non-target traces each add a constant average likelihood ratio (below 1.0) to both the target and foil distributions, so that the difference before the means of these do not change. However, each such trace adds variance to both distributions; this causes the distributions to overlap more and decreases performance. The centering of the criterion causes p(H) to fall and p(F) to rise. The second example involves study time, and item repetitions: Suppose items are each presented for the same time, or same number of repetitions in a given list. We assume that more study time and more repetitions produce a stronger single trace (most of the time) rather than multiple traces, so that both manipulations produce traces with more non-zero

feature values. For foil tests, differentiation will reduce similarity and lower p(F); for target tests the number of matching features rises, overcoming the differentiation effect and causing p(H) to rise. Differentiation also causes negative list-strength predictions—strengthening other list items decreases their similarity to the test item, reducing the ability to cause confusions. Finally, consider REM's account for the finding that low frequency words produce higher performance than high frequency words. The model assumes that one factor correlated with word frequency is feature frequency (evidence for this assumption has been obtained by Malmberg, Steyvers, Stephens, & Shiffrin, 2002): high frequency words are assumed to have more common features (see Fig. 2), increasing their inter-item similarity, thereby increasing confusability and lowering performance. The system is assumed to carry out calculations in all cases with an intermediate value of feature frequency. A mirror effect is the result. (For discussion of the NRS, see Shiffrin & Steyvers, 1997.)

Of course, this REM model is too simple to generalize to more complex variants of the standard paradigms. Shiffrin and Steyvers (1997) assumed successively more realistic and more complex sets of assumptions to govern what traces are stored and what traces join the activated set. Simulations demonstrated qualitatively unchanged patterns of predictions if decisions are based on calculated likelihood ratios as in the simple model. All of the simulations in that article were used to produce qualitative predictions, rather than fits to a particular set of data.

For quantitative fits of REM to episodic recognition data we refer the reader to Diller, Nobel, and Shiffrin (2001), Criss and Shiffrin (submitted, a), Malmberg, Holden, and Shiffrin (submitted), and Malmberg, Zeelenberg, & Shiffrin (accepted). These studies varied list length, category length, study time, word frequency, similarity of targets to foils, presence or absence of targets and foils in lists preceding the critical list, time to respond, and the effect of an amnesic drug.

Given that it has been our typical practice to assume activation only of traces of items from the most recent list, it is worth saying a few more words about the study of Criss and Shiffrin (submitted, a): The words tested after a given study list, whether targets or foils, were sometimes present also in either or both of the two lists preceding the critical study list. Clearly, this setting requires a more sophisticated set of activation assumptions. We therefore assumed that context changed from list to list, assumed that traces from all three relevant lists took part in comparisons, assumed that context and item cuing took place jointly, and assumed that the most recent context was used in the test probe (the likelihood calculations being adjusted appropriately for these assumptions). Fig. 4 shows some of the data and model predictions. Although not shown in the figure this study also varied number of category members within list(s), and this manipulation had an effect, but a much smaller one than those shown in the figure. Thus, contextual confusions in this setting were a much greater source of noise (and therefore much more responsible for lowering performance) than were the number of similar items within the category of the test word, though both factors play a role (see Dennis & Humphreys, 2001).

The articles by Malmberg, Zeelenberg, and Shiffrin (in press, accepted) are also worth a few words. Hirshman, Fisher, Henthorn, Arndt, and Passannante (2002) compared groups given either a saline injection or an amnesic drug at the time of study of a list of words varying in natural language word frequency. The amnesic group showed a normal word frequency false alarm effect, but a reversed word frequency hit effect (compared with the data shown in Fig. 3, for example). Malmberg, Zeelenberg, and Shiffrin showed that the usual simple REM

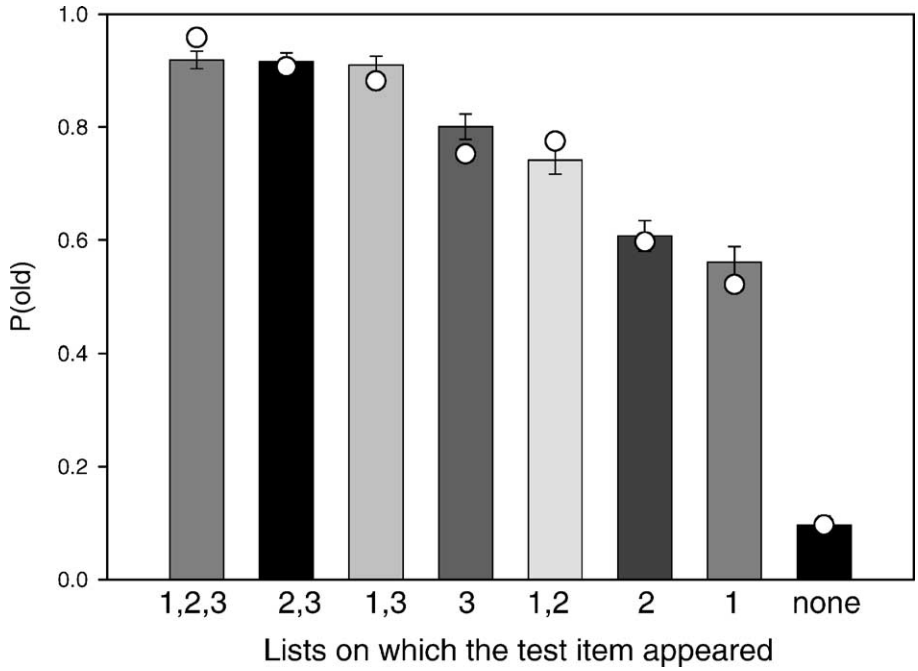

Fig. 4. Data and predictions from REM for the probability of an 'old' response from an episodic recognition study reported by Criss and Shiffrin (submitted, a). Test words were completely new (right most column), or in various combinations of the three most recent lists. List 3 is the most recent list, so correct response probabilities are indicated in the three left most columns. Foils were not in list three, but columns four through seven show that foils that had appeared in lists one or two produced high false alarm rates. The REM model predicted this pattern because the context was similar across successive lists.

recognition model could predict all the findings from the normal and amnesic groups, with only one added assumption—that the value of c (the parameter governing correctness of storage) was lower for the amnesic group. Lowering c reduces the number of matching features in the target trace due to veridical storage, and increases the number due to chance matching; because chance matching favors high frequency words, a cross-over point is reached at which high frequency words come to have a higher hit rate. This result raises the possibility that certain kinds of hippocampal disfunction may produce less accurate storage rather than less storage.

2.2. Recall tasks

It is a basic assumption of the REM framework that all retrieval begins by comparison of the probe cue to memory traces. The basic issue is the way such comparisons should be calculated. A Bayesian analysis applied to recognition led us to a particular formulation in which a likelihood ratio is calculated based on the observed matches and mismatches. We would like to assume that the system carries out comparisons with one approach, in all task settings, including recall. Thus, the comparison process is calculated as likelihood ratios for each trace, as the first step in all retrieval operations. These likelihood ratios serve as the building blocks of retrieval, even for recall.

To model recall, we need to make a few representational assumptions, particularly about the ways that groups of items (such as pairs) are encoded. Until recently, we assumed: (1) study pairs are encoded in one trace containing a set of context features plus two sets of item features, one for each word; (2) a context probe activates pair-traces from the list; (3) the test item is compared to each trace, and two likelihood ratios are calculated, with the higher ratio assigned to each trace.

It seems natural to posit a near-optimal recall decision, in the spirit of the Bayesian approach. For one approach, assume that the pair-trace with the highest likelihood ratio is selected, and that the features of the other item in that trace are used to select the best matching trace from lexical/semantic memory. A similar approach (borrowing from Hintzman, 1988) sums over the features values found in the response halves of the activated traces, weighted by the likelihood ratios assigned to the other halves. The resultant feature values would then be matched to the lexicon with an appropriate likelihood calculation. Both approaches implement a parallel model of cued recall that operates in two steps.

Although these approaches are desirable because they are quite consistent with the Bayesian approach, we have not been able to produce an acceptable form of such a near-optimal recall model. One problem is too high a rate of recall relative to recognition (see Shiffrin & Steyvers, 1998). As discussed in Shiffrin and Steyvers (1998) this can be a problem even if one assumes sampling in proportion to the likelihood ratios. One way to deal with this problem is to assume sampling in proportion to a compressive transformation of the likelihood ratios. This approach was used by Diller et al. (2001); McClelland and Chappell (1998) also employed a form of compression. A greater problem is data from Nobel and Shiffrin (2001). We showed that response times in cued recall, in a task carefully matched to recognition, had distributions much slower and more skewed than those for recognition. Similar results were found in signal-to-respond variants of the paradigm. Model explorations and additional studies demonstrated to our satisfaction that different retrieval procedures operated in recognition and cued recall. We have therefore been led to assume a less than optimal recall process, in the form of a search process. A search process does, however, have the advantage of consistency with the SAM model of Raaijmakers and Shiffrin (1980, 1981).

Recent research (Criss & Shiffrin, submitted, b) has led us to augment this basic approach by extending the representation of pairs to include a set of pair features distinct from the features for the words making up a pair. For present purposes, this augmentation is primarily useful for explaining results we obtained using an associative recognition task (requiring discrimination of intact from rearranged pairs), but leaves basically unchanged the model for cued recall.

2.3. Cued recall tasks

Assume the pair representation stipulated in the preceding section, and that retrieval begins with a context probe producing a set of activated pair-traces (mostly the pairs from the recent list). Next the visual form of the test item is used to retrieve meaning and other features from its lexical/semantic trace, and these features are combined with the presented features to form an episodic probe. The probe is compared in parallel with each of the two sets of single-item features in each of the traces in memory, producing a maximum likelihood ratio assigned to each trace.

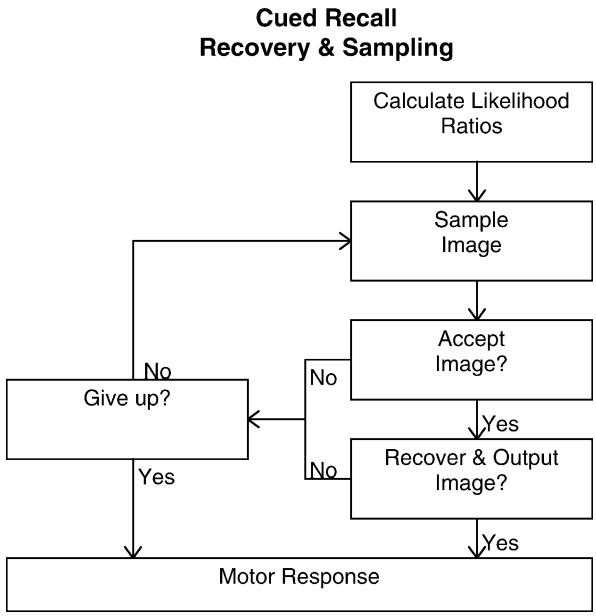

Fig. 5. The most important steps in one cycle of the REM search process for cued (and free) recall. Note that 'image' in this figure has the same meaning as 'trace.' The cycles that tend to take place before a response is emitted produce slower and more skewed response times than in recognition testing.

At this point, recall diverges from recognition. A search process begins, consisting of cycles of sampling and recovery. The sequence is illustrated in Fig. 5: At each step, a trace is sampled in proportion to the likelihood ratios (or some monotonic function of the likelihood ratios). Some of the features in the sampled half-trace and some of the features in the paired half-trace are recovered; the number recovered depends on both the size of the likelihood ratio and the number of features stored in the trace. The recovered information from the sampled half-trace is assessed to decide whether the correct pair-trace has been sampled. Given a positive decision, then an attempt is made to produce an answer by using the recovered information from the other paired half of the same trace (or in the augmented model, also from the pair feature set). This latter process consists of matching the recovered information to the lexicon (in a process described more fully in the discussion of knowledge retrieval later in this article). If this matching to the lexicon does not produce an output, then a decision is made whether to continue searching or to stop. If the search continues, another probe cue is selected (typically the same one just used) and a new sample taken (with replacement). The search continues until an answer is produced, time to search is exceeded, or a decision is made to give up. A carefully worked out model of cued recall is applied to a variety of accuracy and response time data in Diller et al. (2001).

Although a search model is by its nature less than optimal, one can ask why our model assumes proportional sampling, when optimal sampling would select traces in order of largest to smallest likelihood ratios. Shiffrin and Steyvers (1998) discuss this issue and give some justifications for the assumption of proportional sampling. One additional justification might be the need to add variability to the search process. Especially if probe construction is imprecise,

it might be a poor idea for a system to continually resample the same strong trace, when the strongest trace is not the one desired.

2.4. Free recall tasks

The cued recall model is considerably more complex than that for recognition. The even more complex REM model for free recall is also borrowed from the SAM model (see many examples of predictions for free recall and related paradigms in Raaijmakers & Shiffrin, 1980, 1981). In free recall, a list of words is studied, and subjects attempt to recall them in any order. To oversimplify quite a bit, the process can be described as follows: It is assumed that the initial probe cue is context. Then search and sampling begins. When a word is retrieved and output, that word is the basis for the next probe cue. The search continues in this fashion until the current probe cue fails to produce a new output after some number of cycles; then the probe cue reverts to context. The overall search continues until too few new outputs occur to justify continuing.

The switch back and forth in free recall between item-plus-context probe cues and context-only probe cues is critical for predicting certain aspects of list-strength findings. Ratcliff et al. (1990) observed that strengthening some list items didn't harm and may have helped recognition and cued recall of the other list items—this was a negative list-strength result. The explanation, as described earlier, is based on differentiation: Stronger traces produce smaller likelihood ratios when the memory probe includes some other item. However, these authors also observed that strengthening some list items harmed free recall of other list items—this was a positive list-strength result. In the theory, this finding is due to those search cycles that use context cues without item cues. Stronger traces are more strongly similar to the context cue and therefore have higher likelihood ratios. Such stronger traces therefore are sampled preferentially and reduce free recall of other list items.

This prediction was used by Malmberg and Shiffrin (in press) to test the 'one-shot-of-context' hypothesis. If massed repetitions of an item do not increase context storage, but spaced repetitions do, then the positive list-strength effect for free recall should occur only for spaced repetitions. These authors demonstrated just this result in several studies. Representative results are shown in Fig. 6 and qualitative predictions of a simple REM free recall model are shown in Fig. 7.

Although these results may have some inherent interest for specialists in memory, they have been included in this article for another reason, in order to demonstrate a kind of 'web of consistency' for the modeling framework. The one-shot studies were developed because we had wanted to understand why the magnitude of long-term priming (long-term priming is discussed later in this article) depended on the number of presentations of the prime only when the presentations were massed, and not when they were spaced. Because we attributed such priming to the storage of context in lexical traces, we were led to hypothesize that massed study did not increase context storage. But this hypothesis simply restated the finding; how could it be tested independently? If massed study also produced one shot of storage in episodic traces, then a test could be devised. This test, however, depended on the validity of both the SAM model for free recall (e.g., Raaijmakers & Shiffrin, 1980, 1981), the Shiffrin et al. (1990) account of list-strength effects and differentiation, and the model for priming. It would probably be difficult to come up with another reason to expect the one-shot findings, even after the fact,

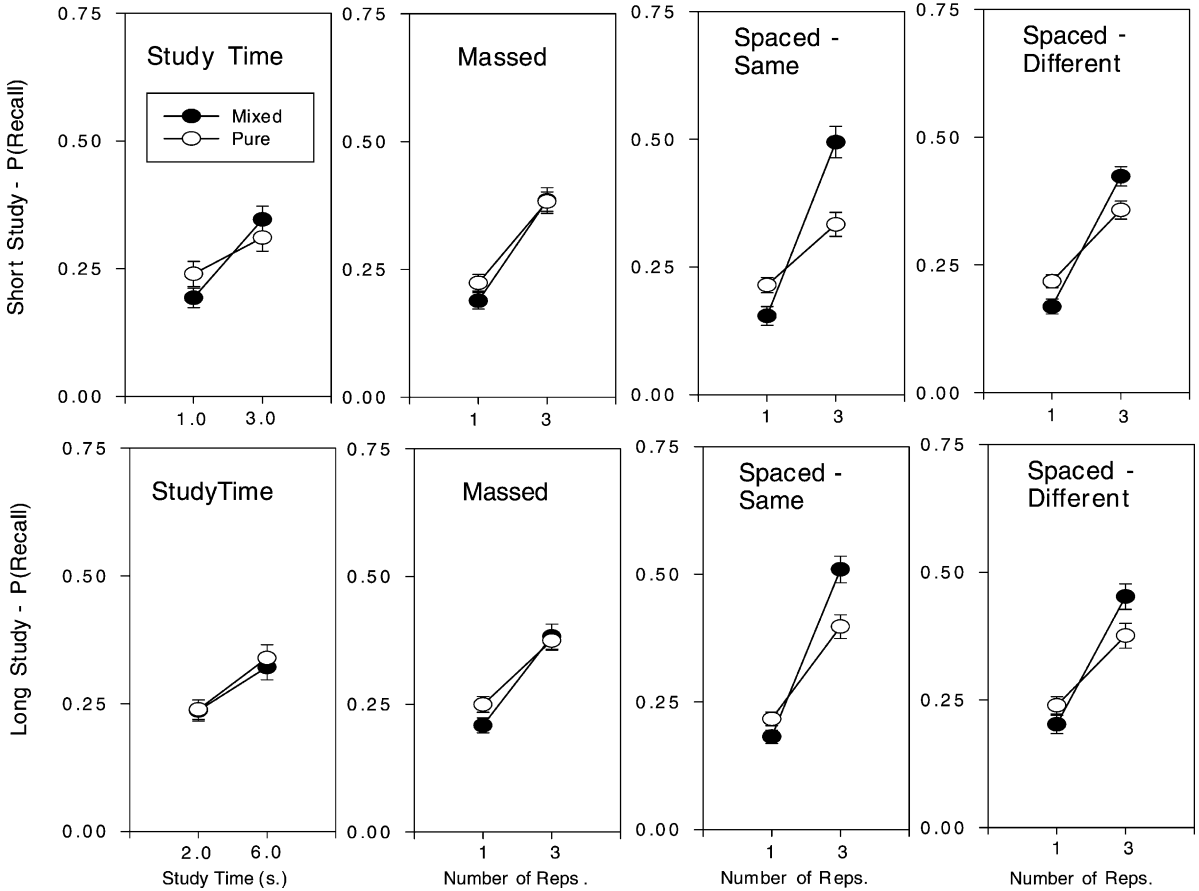

Fig. 6. Data for list-strength effects in free recall. If the black and white dots lie atop one another (as generally true in the left panels), this indicates a null list-strength effect; if the black and white points show the crossover of the right hand panels, this indicates a positive list-strength effect. The four left-side panels illustrate the case when items are strengthened by massed study (extra time or extra massed repetitions); they therefore get only 'one shot' of context storage, regardless of study duration, and show a null list-strength effect. The small positive list-strength effect in the top left panel is attributed to the short presentation time (1 s), which may not have allowed a full shot of context to be stored. The right panels illustrate the case when items are strengthened by spaced study; they therefore get multiple 'shots' of storage of context, and show a positive list-strength effect (spaced-same means the words surrounding each repetition were also the same). Adapted from Malmberg and Shiffrin (submitted).

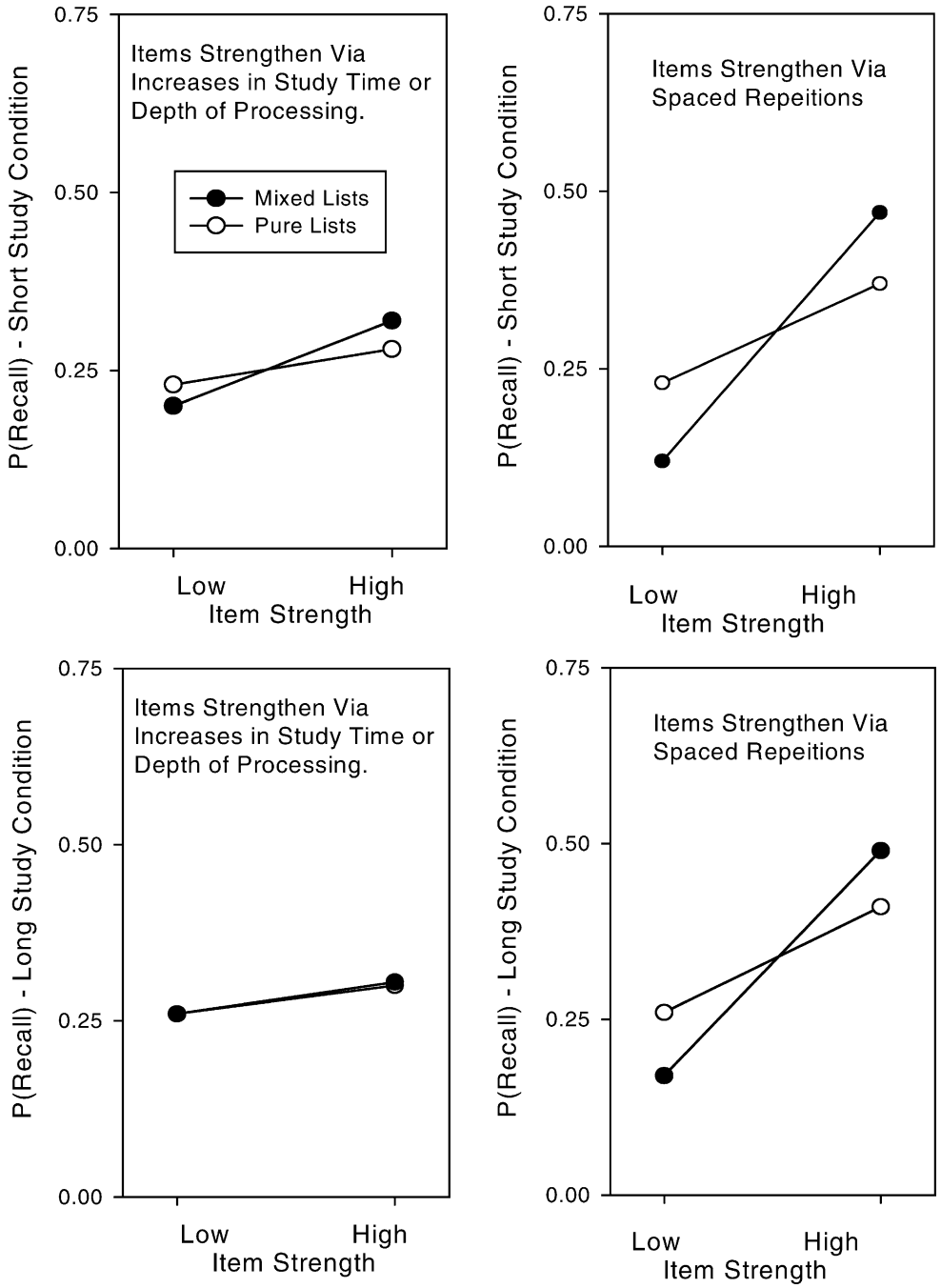

Fig. 7. Simulated list-strength predictions of the REM model for free recall incorporating the 'one-shot-of-context' hypothesis. The model predicts most of the effects depicted in Fig. 6, including the null-effects for massed study and the positive list-strength effects for spaced study (and also the small effect for rapid presentations, attributed to insufficient time to store a full shot of context).

so I would like to suggest that the match of the results to those predicted is testimony to the validity and consistency of the REM memory framework.

2.5. Recall in recognition

There is nothing in the REM framework to exclude the possibility that recall occurs during a recognition task. In fact, the logic of the system essentially guarantees that recall will occur—otherwise one would have to posit some process by which sampling and recovery could be suppressed in certain tasks (like recognition), and there is no reason to posit such a mechanism. One might think to rule out recall because the recall process seems slower (as illustrated by the Nobel & Shiffrin, 2001, results). However, one could assume that only the first sample of a memory trace contributes to recognition, and this first sample could be fast enough to allow recall to contribute to recognition without conflicting with response time data.

Suppose then that recall does occur in recognition tasks. This could be true without implying that such recall would appreciably affect recognition decisions. The reason is that the two routes might almost always produce the same decision. Recognition is based on the average likelihood ratio, which is usually dominated by the single best matching trace. Such a strongly matching trace would also tend to be the one sampled, producing a high correlation between responses produced via global familiarity and responses produced after recovery of information from a sampled trace. This reasoning would apply to many of the usual recognition tasks, but in other types of tasks, the recall route might produce information that would lead to a different decision than that based on familiarity. For example, the use of recall might alter remember/know judgments, or confidence ratings. For another example, the use of very similar foils might make recall useful for the making of fine discriminations.

The last idea was studied by Malmberg, Holden, and Shiffrin (submitted); they chose foils to be words differing in plurality from the same word types at study. In this 'registration-without-learning' study (e.g., Hintzman, Curran, & Oppy, 1992), participants probably were led to use recall to reject test items that were very familiar but in the wrong plurality. A dual process REM model including this assumption was fit to the data: We assumed that test words with likelihood ratios less than 1.0 trigger a 'new' response. Likelihood ratios greater than 1.0 lead to an attempt to use recall. It is assumed that the chance of recalling stored plurality information rises with study repetitions, and if recall does succeed, the response is based on the recalled information (whether or not that information had been stored correctly). If stored plurality information is not recalled, then a guess is made. Data and the model's fit are illustrated in Fig. 8.

As indicated in this figure, this study also varied word natural language word frequency, and it is noteworthy that the model correctly predicts the reversal of the usual false alarm findings for very similar foils (and does so whether or not a recall-to-reject process is added to the model). The explanation is simple: Dissimilar low frequency foils match traces less often by chance because their features are rare. However, similar low frequency foils match target traces well, and the rarity of their matching features gives them higher likelihood ratios compared with high frequency similar foils.

In this study, participants also gave frequency judgments for words judged to have been 'old.' These judgments rose roughly linearly with actual frequency, for both

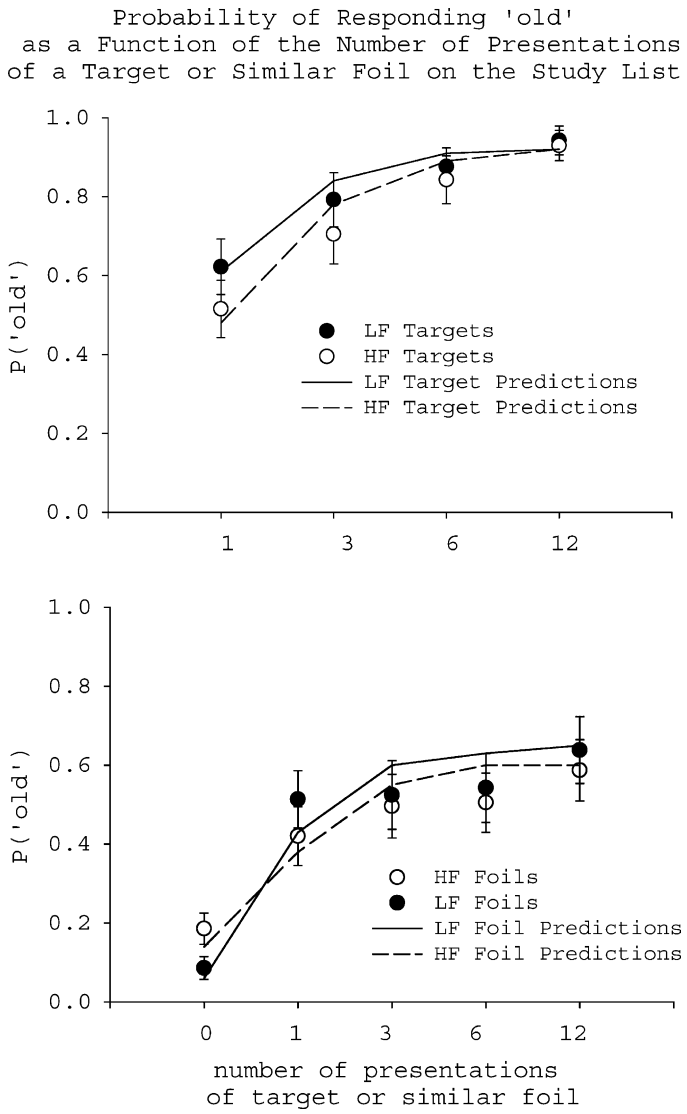

Fig. 8. Episodic recognition data (points) and predictions (lines) of a simple dual-process REM model, from Malmberg, Holden, and Shiffrin (submitted). Top panel: *p*(old) is uniformly higher for low than high frequency targets, across repetitions. Bottom panel: Foils not studied in either plurality show *p*(old) lower for low than high frequency words (the usual mirror effect, indicated by the point labeled zero presentations); foils differing in plurality from studied words show *p*(old) uniformly higher for low than high frequency words (the opposite of the usual mirror pattern, indicated by the points labeled 1, 3, 6, and 12). The REM model predicts this pattern on the basis of foil similarity.

targets and similar foils. The model fit this data well by assuming that frequency judgments were based on degree of familiarity (i.e., the 'odds'). Data and fits of the model to the frequency judgments (using the same parameters) are shown in Fig. 9.

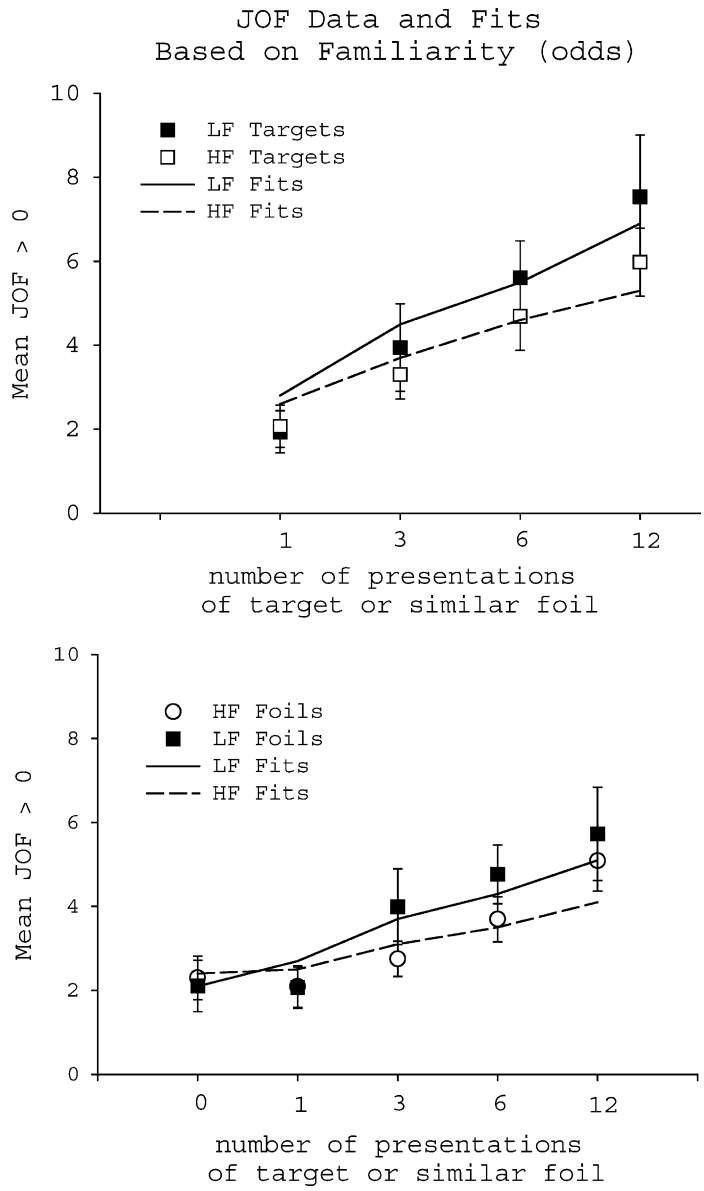

Fig. 9. Frequency judgments for the words judged 'old' in the study depicted in Fig. 8. Predictions are based on familiarity of the word (i.e., the 'odds'), using the same parameter values, and are given as lines.

2.6. Combined use of context and content cues

Episodic memory tasks by definition require joint use of context and content information. Fig. 10 illustrates the logic of the situation for a system assuming separate memory traces. This figure shows the array of traces relevant for a typical episodic memory task: the current list (top

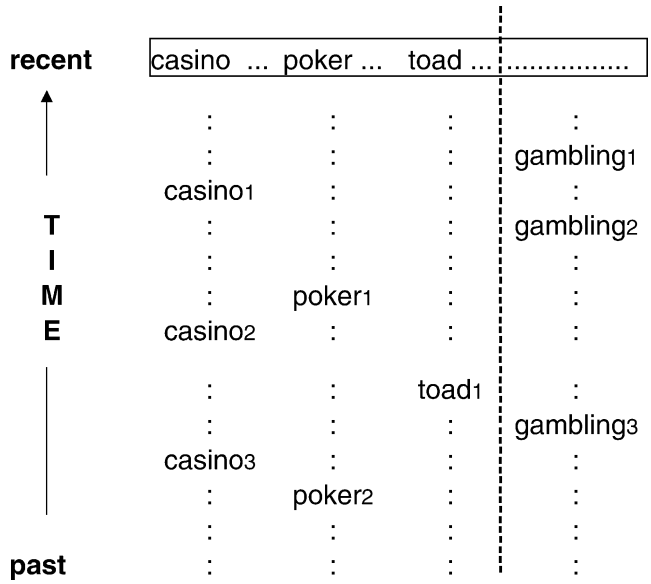

Fig. 10. Episodic memory traces that must be the objects of comparison in episodic tasks. The traces in the box in the top row are those for words just studied in a list, and are therefore most recent. Past traces of those words are shown below. To the right of the dashed line are traces of words not studied in the recent list. Episodic recognition requires discrimination of words (right versus left) and of time (top versus bottom). Hence both context and item cuing is required (from Criss & Shiffrin, submitted).

row, outlined by a box), the past traces of the current list words (rows below the top), and traces of words not on the current list (right column). Only the traces in the box are targets, and clearly needed are both context cueing (to discriminate the top row from the others) and item cueing (to discriminate the list columns from the others). The order in which these context and content cues are used differs among current models. The simplest REM model for recognition (Shiffrin & Steyvers, 1997) assumed a first probe used only context features. This probe activated traces from the recent list, and a second probe with only content features was then matched to the activated traces. The second phase produced the likelihood ratios used to make a recognition decision. The SAM model for recognition (Gillund & Shiffrin, 1984) assumed simultaneous use of context and item cues in the probe. Dennis and Humphreys (2001) in effect assumed a first probe with content information followed by a second probe with context that is matched to the information retrieved due to the first probe.

Criss and Shiffrin (submitted) analyzed this situation and argued that existing data do not allow these different orders of cue utilization to be discriminated. Certain details of assumptions that accompany these model types can often be tested. For example, in the REM approach of Shiffrin and Steyvers (1997) the context-based likelihood ratios for traces that join the activated set are discarded when the second stage context-matching begins. It ought to be possible to test empirically whether this assumption is distinguishable from an alternative assumption that the context-based likelihood ratios serve as priors for the second stage calculations. As another example, Dennis and Humphreys (2001) assume that their stage-one retrieval, based on

content information, never accesses traces of any word other than that tested. Criss and Shiffrin (submitted) presented data and logical arguments that such an assumption is too strong.

However, these model variants and their tests do not bear on the issue of order of cue use: This issue may not be amenable to empirical test, at least for words as stimuli, because the accuracy of response cannot rise above chance until both stages are completed, whatever order is used. I think tests of this issue might require the use of novel stimuli, such as new faces, pseudowords, or unknown objects. For such stimuli, context might not be needed to produce above chance performance, and an empirical test of the issue in question might be able to make use of this possibility. However, directly relevant tests have not yet been carried out.

3. Retrieval of knowledge and priming

The REM system can and ought to be applied to the retrieval of knowledge: A memory framework can hardly be complete without modeling knowledge retrieval, both in its own right, and because such retrieval is critical to explain episodic storage and retrieval. By knowledge we refer not only to facts, and verbal knowledge, but perception generally, motor programs, and everything else that is learned over development. Furthermore, our view is that such learning lies on a continuum: Although we have distinguished episodic traces and lexical/semantic traces for convenience, our view is that the latter grows from the former through a process of accumulation of information, and there must be a continuum between the two. Given this to be the case, then our model needs to address the relation between the two (typically studied as 'implicit memory,' or 'priming').

To have a concrete example to clarify concepts, let us consider word knowledge as stored in lexical/semantic traces. Examples of empirical tests of such knowledge include lexical decision (word/non-word decisions), naming of above threshold words, identification of words presented near perceptual threshold, animacy decisions, and production of associates. In such tasks, the test stimulus is generally a string of letters presented visually (usually a word). A probe is therefore formed from the visual form features plus context. Before describing the REM model we used in such situations I should admit at the outset that a generally acceptable model for retrieval of general knowledge almost certainly needs to consider an enriched representation of knowledge. In particular, one probably needs to represent knowledge in terms of layers or stages of processing, the output of one being the input to the next (e.g., visual form $\rightarrow$ letters $\rightarrow$ sublexical units $\rightarrow$ lexical/semantic trace) with the additional possibility of parallel routes (e.g., phonological processing). There would be memory traces at each of these levels or stages, each matched to the relevant part of the input, and each producing a likelihood ratio. One approach would use the likelihood ratios from one stage as prior odds for the subsequent matching calculations at the next stage. We have not yet implemented such a scheme, so will simplify for now by assuming the existence in memory of separate lexical/semantic codes, and no other relevant units.

The probe is therefore compared in parallel to the various lexical/semantic traces, and a likelihood ratio calculated for each. The way to generate models is now specific to each task. For threshold presentations and accuracy measures (e.g., perceptual identification) one can use a model very much like that for accuracy in episodic memory. However, for tasks that

use response time as the primary measure of performance (e.g., naming times), it is necessary to model the evolution over time of the likelihood ratio and the way in which this evolving measure is used to produce responses.

Much of our information about the way events accumulate to form knowledge comes from long-term priming studies. Priming is the most typical implicit memory effect: The test word is studied in a separate earlier task, and such study affects general knowledge performance (often in a way that appears to facilitate performance, though that issue is not as clear as it might seem). The REM approach to long-term priming is similar across general knowledge tasks that involve perceptual (rather than conceptual) processing: At the time of earlier study, current context (and unique and new visual form information, such as font) is added to the lexical/semantic trace of the studied word. The test involves visual form information plus test context (which is similar to the study context). Thus, primed words match their lexical traces better at test due to the extra matching context (and possibly extra matching form features).

3.1. Perceptual identification

Let me start with a perceptual task in which accuracy is measured: *perceptual identification*. A word is flashed briefly and masked. The subject is asked either (1) to identify the word (termed *identification*), (2) to say whether the flashed word matches a subsequently presented word (either a target or foil; termed *yes–no*), or to choose between two alternatives, one a target and the other a foil (termed *forced-choice*). Ratcliff and McKoon (1997) carried out such studies; their studies manipulated long-term priming because they were aimed to distinguish between two accounts of priming: one in which priming improved perception and one in which priming acted as a source of decision bias. In their studies, therefore, each of the target and foil choices had sometimes been studied in an earlier list.

Schooler, Shiffrin, and Raaijmakers (2001) presented a Bayesian inspired model for the Ratcliff and McKoon results (Ratcliff & McKoon presented their own model, but this article will not attempt to compare the two models). Our model was termed REMI (for REM-implicit). We assumed that the presentation of the flashed word plus mask results in a visual percept represented as a vector filled with accurate and inaccurate visual features (the proportion of accurate ones rises with flash time). These are added to some current context features to form a probe that is compared to the lexical/semantic traces, producing likelihood ratios. This is perhaps the simplest form of a Bayesian model.

A plausible alternative model would assume that the perceptual vector is compared directly to the visual forms of the one or two choices provided in ‘yes–no,’ or ‘forced-choice.’ Such a model has trouble explaining why priming effects found by Ratcliff and McKoon are much more pronounced when the two choices are visually similar. In addition, we have several lines of evidence (that space does not permit detailing) showing that the visual masks degrade visual form information, making form level comparisons less useful and leading to the use of higher level information for the decisions.

Let us then turn to REMI applied to the forced-choice task. When the choices appear, the choice word that has the lexical/semantic trace with the higher likelihood ratio is chosen. Note that features in common to the two choice words (termed non-diagnostic) must match the flash percept equally well, and therefore do not affect the choice. This fact means that similar choices

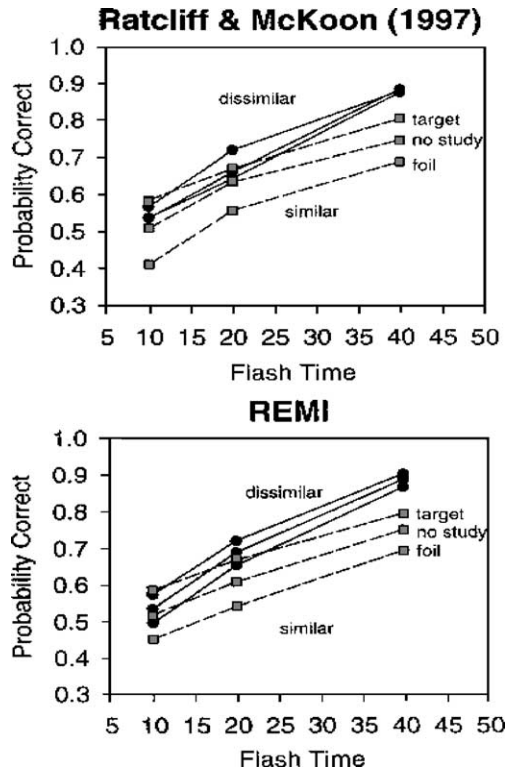

Fig. 11. Priming effects in forced-choice perceptual identification. Data given as p(C) (top panel, from Ratcliff & McKoon, 1997), and REMI model predictions (bottom panel, from Schooler et al., 2001). Flash time is on the horizontal axis, choice words are orthographically similar (dashed lines) or dissimilar (solid lines) and for each group of three lines, the top line is for priming of the target, the middle line for no priming, and the bottom line for priming of the foil.

lower performance (there are fewer diagnostic features to be perceived) and implies that similar choices should increase priming (the number of context-based matches don't change when the number of diagnostic features does change). These predictions are illustrated by the fit of the Schooler et al. (2001) REMI model to the forced-choice results of Ratcliff and McKoon (1997) shown in Fig. 11. As an aside, we note that the data of Ratcliff and McKoon (1997) show that long-term priming acts like a bias, causing a tendency to select whichever choice had been studied earlier. Both their model and our model predict this finding.

For yes–no tasks, a variant decision rule is used: A 'yes' response (i.e., a match response) is given if the likelihood ratio for the test word's lexical/semantic trace exceeds a criterion. This model, using key parameter estimates borrowed from the forced-choice fit, does quite well, as illustrated in Table 1.

For naming, the lexical/semantic trace with the highest likelihood ratio is emitted, as long as the ratio exceeds a criterion. This naming model also fared well, but I refer the reader to Schooler et al. (2001) for model fits, and extensions of the model to deal with word frequency and study repetitions (also see Wagenmakers, Zeelenberg, & Raaijmakers, 2000, 2002).

Table 1
p('yes') in yes–no primed perceptual identification: Data from Ratcliff and McKoon (1997, Experiment 8)

Relation of decision word to flash	Decision word primed?	Data	REMI
Target-same	No	.70	.70
Target-same	Yes	.77	.74
Foil-similar	No	.51	.52
Foil-similar	Yes	.55	.57
Foil-dissimilar	No	.21	.14
Foil-dissimilar	Yes	.23	.18

REMI predictions from Schooler et al. (2001). Similarity and priming both increase *p*('yes').

The pattern of results obtained by Ratcliff and McKoon (1997) and modeled by REMI is not the only pattern obtained in forced-choice perceptual identification. With slight changes in instructions, one can obtain symmetrical priming that is roughly equal for similar and dissimilar choice words (Bowers, 1999; McKoon & Ratcliff, 2001). I can only speculate concerning these striking differences but it is possible that a different strategy is adopted with different instructions. For example, in the Bowers' paradigm some trials may be based on pure visual matching, bypassing lexical matching, thereby exhibiting no priming, and other trials might have too little visual information in the percept to justify visual matching, so a guess is made based on episodic familiarity.

One interesting issue related to this speculation, concerns forced-choice priming when there are no lexical/semantic entries to which the flash can be compared. Suppose instructions are used that do not induce episodic retrieval (like those in Ratcliff & McKoon, 1997); suppose also that study of the prime cannot add current context to the prime's lexical/semantic trace because the prime has no such trace. Then there should be no priming in forced-choice. Ratcliff and McKoon (1995) used line drawings of complex geometric objects, and similar 'impossible' drawings of the same objects, and observed no priming in forced-choice.

3.2. Lexical decision

The task in lexical decision is to classify a test string of letters as a word or non-word. Most versions use free response, and measure response time. In a variant, a signal-to-respond is given at various times after presentation of the test string, and accuracy of decision is the measure. There is a very large literature concerning lexical decision, possibly because the task is believed to provide insight concerning the structure of the mental lexicon. The lexical decision task is quite amenable to analysis based on matching a probe to lexical/semantic traces, calculating likelihood ratios, and responding word if the summed likelihood ratios are higher than a criterion. This approach is of course closely analogous to that used for recognition memory, and this is no accident—the two tasks have a similar structure. A recognition decision is based on how well a test item matches a set of episodic traces activated by context; a lexical decision is based on how well a test item matches a set of lexical/semantic word traces. To illustrate the approach, I will describe briefly a REM model for signal-to-respond data developed by Wagenmakers, Steyvers, Raaijmakers, Shiffrin, van Rijn, and Zeelenberg (accepted).

It is assumed as usual that the probe cue is visual form of the test string plus context, though the context features are probably de-emphasized relative to episodic recognition tasks. The probe is matched to the traces in the lexicon. As in recognition, it is assumed that only the most similar traces in the lexicon take part in calculations, those with likelihood ratios higher than a threshold value. The activated set may be thought of as a set of (mainly orthographic) lexical neighbors of the test string. The model assumes the response 'word' is made if the average likelihood ratio exceeds a criterion near 1.0.

To predict signal-to-respond data, we assume that it takes time to activate probe features and compare them to the lexical traces. The probability of activation of a feature increases monotonically from a minimum time to 1.0. Only activated features participate in comparisons (i.e., the probe vector grows in size as time passes).

To predict effects of natural language word frequency, it is assumed that the percept derived from the flash and current context tends to match the test word's own lexical trace better for a high frequency word than for a low frequency word. This is a natural consequence of our assumptions concerning the way lexical/semantic traces form through experience: High frequency words are encountered more often and hence come to have more complete and more accurate lexical/semantic traces. To predict effects of repetition priming, we make the usual REM assumption that study and test of a word tends to add current context to that word's lexical trace. We note that non-word repetition priming is observed in various studies, including our own. Negative non-word priming is found in speeded tasks like signal-to-respond (but see Wagenmakers, Zeelenberg, Steyvers, Shiffrin, & Raaijmakers, accepted, for other results in non-speeded tasks). We need to augment the basic model to accommodate non-word priming. We therefore made the arguably plausible assumption that study and test of a non-word tends to add current context to the lexical trace of some near orthographic neighbor of that non-word. The addition of current context to a lexical neighbor of the non-word induces a tendency to respond 'word' to a later test of that same non-word. This logic suggests that the effect of priming will be less for non-words than that for words: For word tests, additions of context can be made to the word's strongly-matching lexical/semantic trace; for non-words, additions of context can only be made to the non-word's weakly matching lexical semantic neighbors. The same justification suggests that more addition of context to neighbor(s) occurs for more word-like non-words than less word-like non-words.

These assumptions produce qualitatively correct predictions for a variety of findings from the signal-to-respond lexical decision paradigm. Fig. 12 gives data and fits of the model using these assumptions.

3.3. Short-term priming

The previous sections have presented models for knowledge retrieval in the form of word perception (word identification and decisions concerning whether a presented string of letters is a word), and also for the effect of episodic storage upon such knowledge retrieval (i.e., long-term priming). I turn now in greater depth to the 'pure' perceptual task of word identification. The paradigm is termed 'short-term priming,' because it involves a 'prime' presented just prior to the target. Historically, this paradigm was thought of as a window on the structure of the mental lexicon, because it was thought that the prime initiated activation in the lexicon that spread to

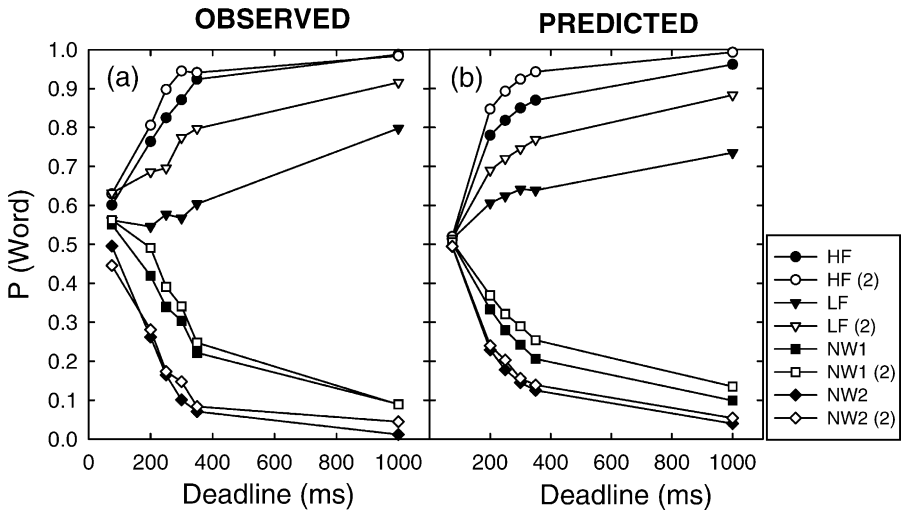

Fig. 12. Probability of responding 'word' in a lexical decision study using a signal-to-respond technique (Wagenmakers, Steyvers, Raaijmakers, Shiffrin, van Rijn, & Zeelenberg, accepted): data in the left panel, and predictions of a REM model in the right panel. Words were high (HF) and low (LF) frequency, and non-words were 'one-letter replaced' (NW1) and 'two-letter replaced' (NW2). The digit '2' in parentheses indicates the second presentation of each item (note that repetition for any item increases the probability of responding 'word').

neighbors; if the target was such a neighbor, it would benefit from the additional activation. To give one example, suppose in lexical decision the target is DOCTOR; when immediately preceded by the prime NURSE the correct response 'word' is speeded, compared to controls. The research I will describe in this section paints a rather different picture of the processes occurring in this situation. To preview the results slightly, we found that priming could produce either positive or negative effects, a result difficult to reconcile with spreading activation as usually conceived. We were led to model this situation instead as a perception/knowledge retrieval task, using a Bayesian approach (Huber, Shiffrin, Lyle, & Ruys, 2001). The discussion that follows will be limited to repetition and orthographic priming (as opposed to semantic priming, as in the NURSE–DOCTOR example). The Huber et al. (2001) article contains data concerning semantic priming, but the effects are much smaller in magnitude than those for orthographic priming, and I will omit discussion of these.

Let me precede discussion of short-term priming with a remark that this research may provide the single best example of the power of the Bayesian approach. For example, the model we proposed proved effective in making *a priori* predictions of findings that could not have been intuited. This research underlies the observation I made at the outset of this article that Bayesian modeling has proved of greatest utility and power in applications to sensory processes.

We used a forced-choice perceptual identification paradigm like that described above for long-term priming: The test consisted of a briefly flashed and masked word followed by two choices. Just prior to the presentation of the target flash, two prime words are presented (for brief or long durations). (The primes, flashed target, and choices were all in different temporal and spatial locations). The use of two primes allowed us to prime the target choice, the foil choice, both, or neither.

The key to our findings and model development was our manipulation of prime processing: In one condition, termed *passive priming*, the primes were presented for 500 ms, with instructions that they were to be ignored, because they were merely a warning that the trial was to begin (the most typical paradigm found in the literature). In the other, termed *active priming*, the primes were better attended because a decision about them was required (e.g., a decision concerning their relative animacy, requiring about three seconds; in fact in other of our studies, we have shown that simply presenting primes for longer durations produces the same pattern of results as a decisional task).

The key data from the first experiment are shown in Fig. 13. There are several noteworthy findings: (1) Both-primed performance was always inferior to neither primed. (2) Processing the primes passively/briefly produced a bias to choose the primed choice, but processing the primes actively/extensively produced a bias NOT to choose the primed choice.

These findings are however quite consistent with a REM model based on Bayesian analysis (Huber et al., 2001). The model builds upon the REMI model we used to explain two alternative forced-choice data collected by Ratcliff and McKoon (1997): The 'higher level' features extracted from the flash are compared to corresponding features of the two choice

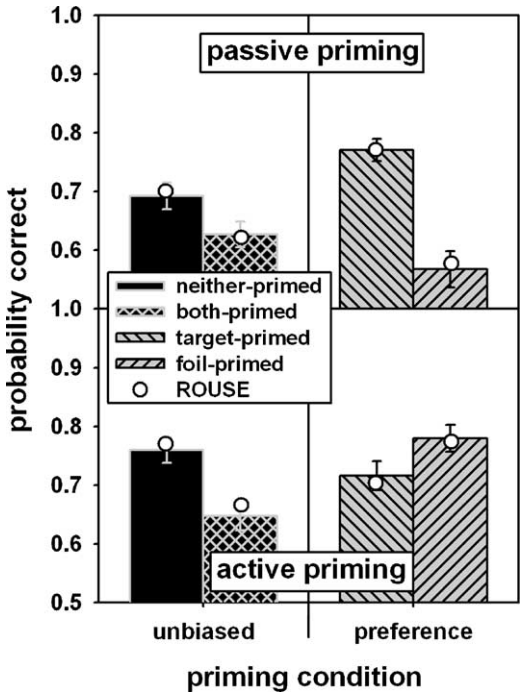

Fig. 13. Forced-choice perceptual identification as a function of short-term priming: Data are given as bars and fits of the ROUSE model are given as points. Passive versus active priming was a between-participants manipulation. The primes were identical to neither choice (column one), both choices (column two), the target choice (column three), or the foil choice (column four). The reversal of direction of preference for the conditions in the right hand panels, for active versus passive priming, was the key result underlying the development of the ROUSE model (from Huber et al., 2001).

words, producing likelihood ratios. The choice word with the higher likelihood ratio is chosen. Because this research developed originally along a different track than that used to formulate the REMI model, we used a slightly different representation and language. In particular, we described this process in terms of 'activated' features in each choice, where activation referred to features in a choice word that matched those in the percept.

The model was termed ROUSE, for Responding Optimally with Unknown Sources of Evidence. ROUSE has two key assumptions: (1) Source Confusion: Primes produce features that are confused with those deriving from the flashed target. (2) Discounting: The evidence provided by a matching feature is discounted if it is associated with a feature known to be in a prime—the evidence should be lower in such a case because the source of the match might have been the prime rather than the flash. In our studies, the primes processed both passively and actively were well above perceptual threshold, so the relevant features were available to be discounted.

To be more specific, it was assumed a feature in a choice word could be activated with probability γ by visual noise, with probability β by the flash (if the flashed word contained that feature), and with probability α by one of the primes (if the primes contained that feature). To calculate evidence, estimates γ', β', and α' of these parameters are required. The accuracy of the estimates of γ and β turns out not to matter much, so we set $\gamma' = \gamma$ and $\beta' = \beta$. However, the accuracy of the estimates of α turned out to be critical: To predict the data, it was necessary for passive priming that α' be a little lower than the actual value, α, and for active priming that α' be a little higher than the actual value, α. This is a good time to point out that an optimal model would set $\alpha' = \alpha$. However, the ROUSE departure from strictly optimal decision making is well founded conceptually: The system can only discount evidence from primes when the system has the information that the observed evidence might have arisen from the primes. The degree to which this information is available to the system must be related to the attention paid to the primes (e.g., if the primes were completely unattended, then it would be sensible that $\alpha' = 0$, something we assume occurs in so-called 'subliminal' priming).

As shown in Fig. 13, ROUSE predicts the pattern of results quite well with this pattern of parameter values: ROUSE predicts a both-primed deficit because priming both choices randomly activates features in the two choices, sometimes favoring one, sometimes the other. The direction of priming switches between passive and active priming because the degree of discounting relative to actual source confusion (the relation of α' to α) differs between these conditions.

This research on short-term priming proceeded to explore a number of other factors, such as similarity of choices to each other, similarity of primes to choices, flash time, and a number of other factors. The ROUSE model with the parameter values from the first study was used to make *a priori* qualitative predictions for these studies. Quite remarkably, the model correctly predicted a large number of highly non-intuitive results. (The results in each case were fit quantitatively with new parameter values, but these did not differ in important ways from those estimated from the first study.)

Perhaps the three most surprising of these follow (I mention the results but refer the reader to the cited articles for details): (A) As the orthographic similarity of the primes to the choices was decreased, in passive conditions the preference for primed targets over primed foils monotonically decreased, but in active conditions the preference for primed foils over primed targets

first reverses, then decreases (Huber et al., 2001, Experiment 2). The correct *a priori* model predictions depended only on the logically necessary assumption that feature overlap between prime and target rises with similarity. (B) As the choices were made more orthographically similar to each other, in passive conditions the preference for primed targets over primed foils remained stable, but in active conditions the roughly equal target and foil priming changed to a large preference for primed targets (Huber et al., 2001, Experiment 3; see also Huber, Shiffrin, Lyle, & Quach, 2002, introduction). Again, these predictions depended only on the assumption that feature overlap rises with similarity. (C) When the duration of the target flash was decreased in steps to zero, in passive conditions the preference for primed targets over primed foils increased, but in active conditions the preference for primed foils over primed targets reversed (Huber, Shiffrin, Lyle, & Quach, 2002, Experiment 1). This prediction depended only on the logically necessary assumption that fewer features are acquired from the flash at shorter durations. In addition to these surprising confirmations of the theory, several other more straightforward predictions were also confirmed (Huber, Shiffrin, Quach, & Lyle, 2002). In the most recent research, Weidemann, Huber, and Shiffrin (submitted) showed that the same word used as primes in different spatial and temporal locations produced independent priming effects that would sum or compete, as predicted by ROUSE.

The ROUSE model for short-term visual priming provides an excellent example of the Bayesian approach: The limitations built into the system are limited perception (indexed by β), visual noise (indexed by γ), and concatenation of irrelevant information into the percept (source confusion, indexed by α). An optimal decision system is then derived, but to fit the observed findings a slight retreat from optimal decision making is needed—a not quite accurate estimate of the amount of source confusion ($\alpha' > \alpha$ for active priming, and $\alpha' < \alpha$ for passive priming). The result was a model that proved remarkably successful, generating a long series of *a priori* and non-intuitive predictions each of which was confirmed by experimentation.

4. Extensions

The studies and modeling described in this article, of course, represent only a current and personal snapshot of a continuing journey of research that is just beginning and will not near completion for many lifetimes. I hope and believe this theory represents systematic progress that builds on the great mass of collected data and many earlier theories, including my own (e.g., Atkinson & Shiffrin, 1968; Shiffrin, 1970; Raaijmakers & Shiffrin, 1980, 1981; Gillund & Shiffrin, 1984; Shiffrin et al., 1990). However, the work yet to be done far outweighs the progress thus far. Neither I nor anyone else could possibly predict the rest of the journey, and this section makes no pretense of mapping the path to come. I will simply mention a few of the directions I expect research will take me in the next few years.

4.1. Feature representations

The features used in modeling in the above applications are arbitrary. Each feature codes base rates in the population and is otherwise without meaning. However, the features are meant to be stand-ins for meaningful information, so it would be a large step forward in the theory if the

features could be provided real interpretations. As a first step toward this goal Steyvers (2000) and Steyvers, Shiffrin, and Nelson (draft) used a method described below to replace with more meaningful entries the part of the REM vectors that are used to represent meaning. We used the resultant vectors to predict performance differences across individual words and lists in episodic memory tasks. Several methods were tried. In one, the associative norms collected by Nelson, McEvoy, and Schreiber (1999) were analyzed by single value decomposition to place each of (about) 5,000 words in a space of several hundred dimensions. The words are placed in the space so that distance represents dissimilarity, and so that the inter-word distances produce a best fit to the structure found in the associative norms (on the assumption that words are more similar if they have a higher entry in the table of associative productions). The dimensions can be thought of as features, and the values on each dimension for a given word can be used as the meaning feature values (although doing this required us to switch REM from an integer to continuous representation). The result is a REM model that can be used to predict individual performance for particular words in particular lists. Steyvers (2000) found that he could predict some of the variance in false memory recall across categories, and some of the variance in other memory tests. He could do so also when the meaning values were produced by an LSA analysis of a document data base (e.g., Landauer & Dumais, 1997), or were produced by HAL (e.g., Burgess & Lund, 2000). This research is still in its infancy.

4.2. Categorization and judgment of feature values

It would be a step forward if a REM-based approach could be used to justify models of categorization. In a sense, this has already been done by Anderson (1991) who proposed a 'rational' model for categorization with many similarities to those I have described, a model that when simplified and adjusted properly reduces to a version of the very successful exemplar model, GCM (Nosofsky, 1986). However, Anderson's approach does not map closely enough onto the REM modeling we have been using to be entirely satisfying. I have therefore started a project to explore ways to generalize the REM memory models to categorization. Such approaches are likelihood-ratio based, and therefore also have close similarities to categorization models based on distributions of labeled exemplars in a psychological space (e.g., Ashby & Perrin, 1988; Ashby & Maddox, 1993).

Let me mention just a few of our preliminary investigations. In one, we suppose that category labels (A or B) are stored with exemplars (to simplify, assume storage of these labels is always correct), that the test item would be one of the studied exemplars, and that the system is unaware that the two categories have higher within-category than between-category similarity. In such a case, the optimal decision strategy is to average likelihood ratios over the traces labeled Category A, and choose A if this average is higher than a similar average over the traces labeled B. We carried out simulations to show that such a model is virtually indistinguishable from the standard exemplar model (e.g., Nosofsky, 1986). To give a second example, if it is assumed that the exemplars are stored with correct labels, that the A exemplars are similar to each other and dissimilar from the B exemplars which are similar to each other, and that a new exemplar from A or B (i.e., non-studied) is tested, then the optimal decision strategy is to choose A if the average log-likelihood ratio is higher for A traces than B traces. Simulations have shown that this model also is hard to distinguish from the standard exemplar model.

The idea of summing or averaging likelihood ratios across traces with some defined feature (like a category label) has many other possible uses. For example, one can use Bayesian analysis to produce a posterior probability distribution for the value of any missing feature(s) in a probe cue: This turns out to be a matter of averaging likelihood ratios across all traces found with a given value for the missing feature, and doing this for each value. One application was to priming data collected by Pecher and Zeelenberg (draft): They required participants to make animacy judgments for presented words (e.g., BOAT). If an orthographic near neighbor of the test word had been studied earlier and matched in animacy (e.g., COAT), then response time was speeded, and conversely if the previously studied near neighbor mismatched in animacy (e.g., GOAT) then response time was slowed. According to our priming model, study adds current context to the lexical/semantic trace. Suppose at test the participant makes an animacy decision by comparing summed likelihood ratios across lexical/semantic traces of near-neighbors labeled animate to summed likelihood ratios across lexical/semantic traces of near-neighbors labeled inanimate. A primed trace in the neighborhood would match a bit better due to matching of current context, and would increase the tendency to respond with the label attached to that trace.

5. Final note

This article was meant to summarize the present status of a long-term research project through a conceptual overview. Space limitations required omissions such as almost all formal analysis, even though the article's genesis was the award of the David E. Rumelhart Prize for contributions to the 'the formal analysis of human cognition.' I want to assure the reader that formal analysis is available in the typically much longer publications cited. The present overview emphasizes the role of Bayesian analysis in model development, but will surely disappoint those who would like to see the model derived directly from first principles and optimal adaptation. In fact, the Bayesian approach may contribute only 5 or 10% of the theoretical assumptions, the rest deriving from many factors, the most important of which are surely empirical findings. It would be nice if the theory had progressed to the point where it could be used to provide *a priori* predictions for a wide variety of tasks, but the day when that will be possible is far in the future. In the present stage of development, quantitative modeling must be designed to handle the specifics of given paradigms, and generally must be based on assumptions tailored to the observed outcomes for each task. The successful *a priori* predictions of the ROUSE model for short-term priming of perceptual identification is a surprising and unexpected exception to this general rule. For the most part, the best I can claim is that the general framework provides a coherent and plausible structure having consistent assumptions from which models can be built for many current and (hopefully) future tasks. I am pleased by the results to date, but this pleasure is tempered by the realization that my understanding of memory (and the understanding of the field as a whole) is at such a primitive level that current models will in the not too distant future be viewed as comically oversimplified. For those entering this field, however, the present state of affairs can be viewed as providing a grand opportunity for significant new contributions; this prospect is as exciting and enticing to me as to any new graduate student.

Acknowledgments

This research was supported by NIMH Grants 12717 and 63996, and numerous students, postdoctoral visitors, and research collaborators.

References

Anderson, J. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*, 124–150.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York: Academic Press.

Bowers, J. S. (1999). Priming is not all bias: Commentary on Ratcliff and McKoon (1997). *Psychological Review*, *106*, 582–596.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines*. Hillsdale, NJ: Erlbaum.

Criss, A. H., & Shiffrin, R. M. (submitted, a). Context noise and item noise jointly determine recognition memory: A comment on Dennis & Humphreys. *Psychological Review*.

Criss, A. H., & Shiffrin, R. M. (submitted, b). The lack of interference for pairs and single items in episodic memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*, 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*, 414–435.

Geisler, W. (in press).

Geisler, W. S., & Kersten, D. (2002). Illusions, perception, and Bayes. *Nature Neuroscience*, *5*, 508–510.

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

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

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

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

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

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.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Quach, R. (2002). Mechanisms of source confusion and discounting in short-term priming 2: Effects of prime similarity and target duration. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*, 1120–1136.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review*, *108*, 149–182.

Huber, D. E., Shiffrin, R. M., Quach, R., & Lyle, K. B. (2002). Mechanisms of source confusion and discounting in short-term priming. 1: Effects of prime duration and prime recognition. *Memory & Cognition*, *30*, 745–757.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (submitted). Effect of repetitions, similarity, and normative word frequency on judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Malmberg, K. J., & Shiffrin, R. M. (in press). The 'one-shot' context hypothesis: Effects of study time in explicit and implicit memory. *Psychological Bulletin & Review*.

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

Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (in press). Modeling Midazolam's effect on the hippocampus and recognition memory. *Advances in Neural Information Processing Systems*.

Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (accepted). An analysis of single- and dual-process accounts of the nature of memory impairment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 724–760.

McKoon, G., & Ratcliff, R. R. (2001). The counter model for word identification: Reply to Bowers (1999). *Psychological Review*, *108*(3), 674–681.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1999). *The University of South Florida word association, rhyme, and word fragment norms*. http://www.usf.edu/Free Association.

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. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Pecher, D., & Zeelenberg, R. (draft). Animacy priming.

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

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 14, pp. 207–262). New York: Academic Press.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

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

Ratcliff, R. M., & McKoon, G. (1995). Bias in the priming of object decisions. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 754–767.

Ratcliff, R. M., & McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review*, *104*, 319–343.

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. (1970). Memory search. In D. A. Norman (Ed.), *Models of memory* (pp. 375–447). New York: Academic Press.

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

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. *Psychonomic Bulletin and Review*, *4*(2), 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, UK: Oxford University Press.

Steyvers, M. (2000). *Modeling semantic and orthographic similarity effects on memory for individual words*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (draft). Semantic spaces based on free association that predict memory performance. *Triple Festschrift honoring Lyle Bourne, Walter Kintsch, and Tom Landauer*.

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

Wagenmakers, E. J. M., Zeelenberg, R., Steyvers, M., Shiffrin, R. M., & Raaijmakers, J. G. W. (accepted). Non-word repetition in lexical decision: Evidence for two opposing processes. *Quarterly Journal of Experimental Psychology*.

Wagenmakers, E. J. M., Zeelenberg, R., Huber, D., Raaijmakers, J. G. W., Shiffrin, R. M., & Schooler, L. (2002). REMI and ROUSE: Quantitative models for long-term and short-term priming in perceptual identification. In J. Bowers & C. J. Marsolek (Eds.), *Rethinking implicit memory*. Oxford, UK: Oxford University Press.

Wagenmakers, E. J., Zeelenberg, R., & Raaijmakers, J. G. W. (2000). Testing the counter model for perceptual identification: Effects of repetition priming and word frequency. *Psychonomic Bulletin & Review, 7*, 662–667.

Weidemann, C. T., Huber, D. E., & Shiffrin, R. M. (submitted). Spatiotemporal confusion and compensation in visual word perception. *Journal of Experimental Psychology: Human Perception and Performance*.