Psychological Review
2001, Vol. 108, No. 1, 257–272


0033-295X/01/$5.00 DOI: 10.1037//0033-295X.108.1.257

# THEORETICAL NOTES

# A Bayesian Model for Implicit Effects in Perceptual Identification

Lael J. Schooler
Pennsylvania State University

Richard M. Shiffrin
Indiana University at Bloomington

Jeroen G. W. Raaijmakers
University of Amsterdam

Retrieving effectively from memory (REM; R. M. Shiffrin & M. Steyvers, 1997), an episodic model of memory, is extended to implicit memory phenomena, namely the perceptual identification studies reported in R. Ratcliff and G. McKoon (1997). In those studies, the influence of prior study was greatest when words were presented most briefly and when forced-choice targets and foils were most similar. R. Ratcliff and G. McKoon use these data to argue against models in which prior study changes a word's representation. A model in which prior study changes a word's representation by adding context information is fit to their data; at test, the model uses a Bayesian decision process to compare the perceptual and context features associated with the test flash to stored traces. The effects of prior study are due to matching extra context information and are larger when alternatives share many features, thereby reducing noise that attenuates these effects.

In an implicit or indirect memory task, recent prior experience influences performance on a task that requires only general knowledge for completion, even when there does not appear to be awareness of the recent experience. For example, studying a word in a list improves identification of that word an hour later, when it is presented briefly and masked, in what the participant is led to believe is an unrelated task. A number of investigators have obtained evidence suggesting that at least the largest part of this effect is not due to a change in the quality of the sensory information extracted from the test presentation (e.g., Masson & MacLeod, 1996; Ratcliff & McKoon, 1997). Rather, Masson and MacLeod (1996) suggested that the degraded sensory information acts as a cue for the automatic retrieval of a prior study episode. It is the retrieval of this episode that increases the probability of responding with the target. Alternatively, Ratcliff and McKoon (1997) argued that prior study biases participants to respond with previously studied words. We return to their view of bias shortly.

In the task of Ratcliff, McKoon, and Verwoerd (1989), a target word (e.g., *lied*) was presented briefly on a computer screen (e.g., for 30 ms) and then masked (e.g., @@@@). Next, two words appeared and remained on the screen until the participant identified the flashed word by choosing between the two test words. Table 1 shows results from Ratcliff and McKoon's (1997) Experiment 1. When the alternatives were visually similar (e.g., *lied* vs. *died*), previous study of one member of the test pair led participants to choose it, whether or not it had been flashed. However, there was little effect of prior study when the alternatives were visually dissimilar (e.g., *lied* vs. *sofa*). The fact that the two bias conditions taken together do not exceed the no-prime baseline is one line of evidence suggesting that priming does not improve perception.

We have mentioned two critical effects: a bias that is symmetrical around baseline (usually interpreted as a failure to change perception), and the difference between similar and dissimilar choices (a finding that strongly constrains models). It is important to note bounds on these findings. First, in his versions of the same paradigm, Bowers (1999) has shown that dissimilar choice-words produce a bias virtually as strong as that for similar choice-words. Similar findings have been reported by Neaderhiser and Church (1998). This alternate pattern of results may be due to slight instructional differences (McKoon & Ratcliff, in press); it is conceivable, for example, that studies showing strong effects for dissimilar words might have induced participants to rely more heavily on access to episodic memory traces. Second, it is sometimes possible to demonstrate perceptual gains in the present paradigms, although the magnitude of the effect is smaller than

Lael J. Schooler, Department of Psychology, Pennsylvania State University; Richard M. Shiffrin, Department of Psychology, Indiana University at Bloomington; Jeroen G. W. Raaijmakers, Department of Psychonomics, University of Amsterdam, Amsterdam, the Netherlands.

This research was supported by National Service Research Award Fellowship 1F32HD/MHC7787-01A1; a National Science Foundation–North Atlantic Treaty Organization Fellowship to Lael J. Schooler; Office of Naval Research Grant N00014-90-J-1489; and National Institute of Mental Health Grant 12717. We thank Roger Ratcliff and Gail McKoon for making their data available to us. We also thank David Diller, Larry Jacoby, Colin MacLeod, Roger Ratcliff, and Roderick Smith for their comments on this article.

Correspondence concerning this article should be addressed to Lael J. Schooler, Department of Psychology, Moore Building, Pennsylvania State University, University Park, Pennsylvania 16802. Electronic mail may be sent to ljs24@psu.edu.

Table 1
*Forced-Choice Results From Ratcliff and McKoon (1997, Experiment 1)*

| Similarity Condition | Study condition: Target studied | Distractor studied | Neither studied |
|---|---|---|---|
| Similar (*died* vs. *lied*) | .85 | .66 | .75 |
| Dissimilar (*died* vs. *sofa*) | .83 | .88 | .87 |

*Note.* From "A Counter Model for Implicit Priming in Perceptual Word Identification," by R. Ratcliff and G. McKoon, 1997, *Psychological Review, 104,* p. 323. Copyright 1997 by the American Psychological Association. Adapted with permission of the author.

that found in Table 1 (Wagenmakers, Zeelenberg, & Raaijmakers, 2000). Our main goal in this article is to present a model for Ratcliff and McKoon's (1997) original pattern of results, although we will take up possible accounts of the other findings later.

Ratcliff and McKoon (1997) presented an elegant counter model to predict results involving perceptual identification. In their counter model, the perceptual system generates evidence based on the visual stimulus (e.g., the flashed word). Counters accrue the perceptual evidence associated with the alternatives and noise counts that are randomly assigned to the alternatives. As flash duration is reduced, the perceptual system generates less diagnostic evidence and more noise. When the counter for one of the choices exceeds the other by a criterion number of counts, a response is given. In this model, words inhabit a perceptual space in which visually similar words are near each other. A studied word tends to attract nondiagnostic counts that would otherwise have been accrued by counters for its near neighbors (Ratcliff & McKoon, 1997, do not present a model for the mechanism by which this occurs). The attractive force is weak, so counts bound for distant words are not captured.

Ratcliff and McKoon (1997) argue that their pattern of results poses a challenge for existing models of word identification, such as a logogen (Morton, 1969, 1970), or simple counter model. In a logogen model, prior study raises the resting level of the logogen for that word. Such an account does not explain why the advantage should be reserved for similar, but not for dissimilar, alternatives. More generally, Ratcliff and McKoon (1997) conclude,

> The main reason that other existing models cannot explain priming effects is the assumption that prior exposure to a word changes some property of the representation of the word itself. . . . When a property of the word itself changes, then processing of the word should always show facilitation relative to processing of other words. (p. 339)

In this article, we demonstrate that such claims are too broad. The bias results, and indeed the quantitative pattern of results obtained by Ratcliff and McKoon (1997), can be predicted by a model that assumes that priming does act to alter the word's lexical–semantic memory representation. We show that this is possible by embedding such an assumption within a model in which participants try to make the best possible decision, given available data and other processing constraints. Such models have been used successfully in other memory settings (e.g., Anderson, 1990; Anderson, Bothell, Lebiere, & Matessa, 1998; Anderson & Schooler, 1991; McClelland & Chappell, 1997; Schooler & Anderson, 1997; Shiffrin & Steyvers, 1997, 1998).

We cast this normative model within a Bayesian framework, a framework we have been using to develop a coherent account of major explicit and implicit memory phenomena. Previous work focused on explicit memory in the form of recognition and cued recall, and the model was termed retrieving effectively from memory (REM; Nobel, Diller, & Shiffrin, in press-a, b; Shiffrin & Steyvers, 1997, 1998). The present word-identification task is quite different from episodic memory tasks, but the use of a Bayesian decision process and the way in which study produces changes in memory representations are common to the two situations. In particular, we will demonstrate that implicit effects in masked word identification follow directly from the assumption that certain context and low-level features are added to the lexical–semantic memory trace when the prime is studied. We shall show that because the effects of the additional stored information are overwhelmed by the variability in the amount of available information when the alternatives are dissimilar, the resultant model predicts much smaller effects of prior study for dissimilar alternatives. We argue that a system acting in this way properly takes into account (in a Bayesian sense) the prior probabilities of encountering words.

## The REM Model in Brief

We begin with a brief description of the REM model that has been developed to predict the phenomena of explicit memory. In REM, memory images are represented as vectors of feature values. The first time a new event (e.g., ⟨4, 3, 2, 1, . . . , 8, 3⟩) is stored in memory, its image is incomplete and error prone (e.g., ⟨3, 0, 1, 0, . . . , 5, 0⟩; the number 0 indicates that no information is stored). When an event is repeated, a new image is usually stored for the new presentation; in addition, information is sometimes added to one or more previous images of that event if the previous images are similar enough. Thus, over developmental time, repetitions of inputs cause certain images to accumulate features and become increasingly complete, at which point the model terms them *lexical–semantic* images (e.g., ⟨4, 2, 2, 3, . . . , 8, 1⟩). As discussed below, it is critical for the model of implicit memory effects that in most cases a feature, once stored, does not subsequently change; this is desirable because it prevents continual storage of wrong information in general knowledge. Of course, some correction mechanism is needed for feature values initially stored incorrectly, so that over time lexical–semantic images tend to store valid information. REM therefore assumes that correction can occur in cases when attention is directed toward incorrectly stored features, a state that will usually occur when some sort of feedback provides pointers to the storage error. Over developmental time, this will happen often enough to correct errors in earlier storage for images that continue to attract additional storage. We therefore assume that lexical–semantic images contain mostly complete and correct information.

Images include both content and context features. Some of the content features describe the visual form of the word, whereas others encode semantic information. The context features define the setting or settings in which a memory image was constructed. Accumulation of contexts over developmental time causes lexical–semantic images to lose identification with any one context. The probability that a given feature takes on a value $j$ is termed $P(V = j)$. In the earlier implementations of REM, both context and con-

tent feature values were drawn randomly from a geometric distribution with parameter $g$:

$$P(V = j) = (1 - g)^{j-1}g, \quad j = 1, \ldots, \infty. \tag{1}$$

Choosing features according to this geometric distribution results in smaller feature values being more common than larger ones and captures the intuition that feature values should vary in base rates. Most predictions are not dependent on this assumption, however, so substituting binary or uniform distributions yields predictions that are just as good as those based on the geometric distribution. We will return to this point below.

In REM, Bayesian inference procedures determine the probability that a memory probe refers to a particular memory image. A comparison between the probe vector and a memory vector yields the number of features and their values that match and mismatch between the two. These matches and mismatches are the inputs to the inference procedures. Shiffrin and Steyvers (1997) developed REM initially to predict the phenomena of episodic recognition memory. Their model was able to handle such phenomena as the list-strength effect (Ratcliff, Clark, & Shiffrin, 1990), the mirror effect (e.g., Glanzer & Adams, 1990), and the normal receiver-operating characteristic (normal ROC) slope effect (e.g., Ratcliff, McKoon, & Tindall, 1994), phenomena that had posed significant problems for existing models. Here we apply these principles in the development of a model of implicit memory.

## REMI: A Model of Implicit Memory

Theories within the Bayesian framework borrow from Anderson's (1990) rational analysis approach the strategy of developing normative models: These make the best decisions possible given certain processing limitations and given data that may include error resulting from imperfect storage and perception. Such an approach produces a mathematically sensible model, but the probability calculations involved do not strike most observers as plausible neural or mental behavior. One justification for such a model, therefore, is that the memory-perceptual system adapts to conform to optimal procedures over evolutionary and developmental time scales. Thus, the system acts as if it were carrying out a Bayesian analysis, whatever the actual neural or mental basis.

Presumably, the primary objective of the systems tapped by masked word identification is the identification of objects in the world based on limited perceptual data. Consider the normative solution to the problem of identifying an object from a set of possible alternatives, given limited perceptual data. Let $p_0(\mathbf{w}_i)$ be the prior probability that an object (in this case a word) $\mathbf{w}_i$ would be encountered. Let $p(\mathbf{d}|\mathbf{w}_i)$ be the probability that perceptual data, $\mathbf{d}$, would be observed given that $\mathbf{w}_i$ is in the environment. These probabilities can be combined using Bayes's rule to calculate $p(\mathbf{w}_i|\mathbf{d})$, the probability that $\mathbf{w}_i$ is in the environment, given the perceptual data (here $j$ indexes all words):

$$p(\mathbf{w}_i|\mathbf{d}) = \frac{p_0(\mathbf{w}_i)p(\mathbf{d}|\mathbf{w}_i)}{\sum_j p_0(\mathbf{w}_j)p(\mathbf{d}|\mathbf{w}_j)}. \tag{2}$$

This is essentially identical to the adaptive control of thought, rational's (ACT–R's) pattern recognition equation (Anderson & Lebiere, 1998; Anderson & Matessa, 1992). The similarity of the two recognition equations is not surprising, because a Bayesian solution cast at such a high level of abstraction would have to take this form. We see this as a strength of the normative approach outlined above because irrespective of who does the development, models developed following these prescriptions will tend to more or less converge onto the same solution.

What is critical for handling Ratcliff and McKoon's (1997) data is that the prior probability that the word would be observed enters into the choice. The question then is how to set this prior probability. It is important to distinguish how the system operates because of developmental learning from how it adapts to local constraints (such as the experiment-defined rules that the two choices are equally probable, or that the conditional probability that a prime will be the target is .50). At a more superficial level of operation, the participant imposes control processes in an attempt to be responsive to local constraints. A simple example would be setting a response criterion so that there is a tendency to choose a word remembered from the studied list, if such words are indeed targets at a higher than chance rate. Presumably, at a deeper level the system will be responsive to contingencies that are typical of lifelong environmental events (e.g., Anderson, 1990).

An example of the kind of contingency to which the system ought to exhibit sensitivity is recency, because recent words are more likely to appear again (Anderson & Schooler, 1991). If, for this reason, the system does exhibit a bias to choose a recently studied word, this would help explain one of the findings of Ratcliff and McKoon (1997). Of course, in a given model, it is necessary to provide a process by which the system can know how recently a choice has been encountered. The pattern of results obtained by Ratcliff and McKoon (1997), however, showed more than a sensitivity to recency: The bias for choosing a recently studied word was attenuated when the alternatives were dissimilar. This too is consistent with the just-described normative solution, because in the dissimilar case there is more diagnostic data to differentiate the two alternatives. A feature of a Bayesian analysis is that the influence of the prior decreases as the amount of data increases.

The above analysis illustrates how Bayesian principles could be used to predict some of the qualitative features of the results observed by Ratcliff and McKoon (1997). The next step is to produce a process model that will approximate the operation of these principles. We lay out such a model in the following sections, which we apply a model loosely based on REM to three perceptual identification tasks studied by Ratcliff and McKoon (1997). For convenience, we term the model retrieving effectively from memory, implicit (REMI). We demonstrate that the model can predict the Ratcliff and McKoon (1997) results, including the findings concerning the similarity of alternatives in a forced-choice task. Certain mathematical and computational details are provided in Appendixes A and B.

## REMI Applied to a Forced-Choice Task

Applying the normative model to a two-alternative forced-choice (2 AFC) task is straightforward. Let the vectors representing the target and foil be termed respectively $\mathbf{t}$ and $\mathbf{f}$. The Bayesian decision system then compares the perceived vector with the two choice-word vectors and calculates the probability that the choice word we have designated $\mathbf{t}$ is in fact the target $\mathbf{T}$, in the following version of Equation 2:

$$p(\mathbf{t} = \mathbf{T}|\mathbf{d}) = \frac{p_0(\mathbf{t} = \mathbf{T})p(\mathbf{d}|\mathbf{t} = \mathbf{T})}{p_0(\mathbf{t} = \mathbf{T})p(\mathbf{d}|\mathbf{t} = \mathbf{T}) + p_0(\mathbf{f} = \mathbf{T})p(\mathbf{d}|\mathbf{f} = \mathbf{T})}. \quad (3)$$

In this equation, the priors (indicated by the subscripts) refer both to those implied by the experimental procedure and to those taking recency into account (by storage of context features in lexical–semantic traces at study). The system should choose the target when $p(\mathbf{t} = \mathbf{T}|\mathbf{d})$ exceeds .5, and half the time when $p(\mathbf{t} = \mathbf{T}|\mathbf{d}) = .5$.

We next develop a process model that implements the normative solution. In a task presenting alternatives after a (masked) near-threshold presentation, we assume a first stage of sensory processing that produces some visual features appropriate for the stimulus and some noise features presumably due to internal noise and the actions of the mask; the probe of memory consists of these features and some context features. We take context to be a combination of internal and external cues. Examples of internal context cues might be mood or a memory from lunch. External cues might be the posters in a room, the hum of a fan, or a word's font.

We propose that the probe (the sensory features extracted from the presentation plus some current context) is compared in parallel with the traces in memory. In a forced-choice task, the decision will be restricted to the two alternatives. Thus, the probe vector (consisting of veridical, noise, and current context features) is compared with the lexical–semantic vectors for the two alternatives: The choice is based on the better of the two matches, *better* defined in the sense of Equation 3.

The effects of prior study are modeled in the following way: When a word is studied, the lexical–semantic trace for the word is contacted, and there is a tendency for features of the current presentation that are not already in the trace to be stored in that trace. Because the semantic features for a word are already known, these are not stored. What might be stored are context and low-level features (like font) that are unique to the current presentation. Because these features tend to be part of the probe at the time of test (as in a contextual drift model for context change; Estes, 1955; McGeoch, 1932; Mensink & Raaijmakers, 1988), they produce additional matching features for studied words. These extra matches are a form of bias because they add to the evidence for both target and foil, whichever had been studied, and do not affect the features extracted from the target flash. Thus, this source of bias toward choosing a word that had been studied is independent of any possible perceptual benefit that prior study might or might not have conferred and is therefore consistent with the assumption of the counter model that prior study does not increase the amount or quality of visual information extracted from the flash (Masson & MacLeod, 1996, also made this point).

For purposes of constructing a simulation model, these various assumptions may be simplified and made concrete as follows. First, we distinguish diagnostic and nondiagnostic feature values. Any feature that has the same value for the two choice-words cannot provide differential evidence, regardless of what is perceived, and hence is termed *nondiagnostic* and is ignored. Consider next the visual part of each vector, ignoring context. Assume there are $L$ visual features in all. Assume there are $L_d$ diagnostic visual features for dissimilar choices and $L_s$ diagnostic visual features for similar choices ($L$ will be larger than $L_d$, which will be larger than $L_s$). The nondiagnostic features are irrelevant for either similar or dissimilar conditions, so the visual vectors for **t** and **f**, consisting of the $L_d$ and $L_s$ diagnostic features in the dissimilar and similar conditions respectively, are filled with integer values drawn from a distribution $P(V = j)$ that is geometric with parameter $g$ (see Equation 1). A vector, **d**, representing the features perceived from the flashed target in the diagnostic positions, is constructed as follows: For each flash duration there is some probability that the $i$th visual feature, $\mathbf{t}_i$, of the target vector will be accurately copied to the $i$th position of the vector, **d**, that represents the flashed word. Features of the flashed word that are not copied are filled with random features drawn from a geometric distribution of Equation 1 with parameter $g$. The vector of perceived features, **d**, is matched against the lexical–semantic target vector, **t**, and the foil vector, **f**. For each of **t** and **f**, the result of visual matching is a set of matching and mismatching features, with particular values. These values are used to calculate the likelihoods for each alternative, termed $p(d|t)$ and $p(d|f)$ in Equation 3.

Consider next context features and context matching. Such features and their matching are assumed to determine the prior probabilities, $p_0(t)$ and $p_0(f)$ in Equation 3. Although there are a number of potentially complex ways in which one could characterize context matching (corresponding to the matching of visual features), for present purposes it is sufficient to simplify greatly and assume that a choice-word that has been studied has a probability $\alpha$ of contributing exactly one extra matching feature. If present, such a matching context feature is assumed to contribute some fixed amount of evidentiary value in favor of that alternative.

The next step is to take the matching values and calculate the likelihood values. At this point, two variants of the model must be considered. The critical issue is whether the values of the matching features should be taken into account. In the REM models used for episodic memory, matching of rare feature values provided more evidence than matching of common feature values. This assumption fits well with the Bayesian justification for the present approach. However, calculating different probabilities for different feature values is asking quite a lot of the neural-mental machinery that implements the present system; it would clearly be much simpler for the system to count the number of matches irrespective of their values. We have implemented both versions, and both produce adequate predictions for present data sets. The version taking the values into account must be simulated. For the version simply counting matches, it is possible to derive analytical predictions, as shown in Appendix A; it is this version that we describe in this article. For this version, the likelihood term in Equation 3 is

$$p(\mathbf{d}|\mathbf{w}) = \prod_{i=1}^{M} p_m \prod_{i=1}^{L'-M} p_n. \quad (4)$$

where **w** refers to one of the choice words (**t** or **f**), **d** refers to the vector of perceived feature values, $p_m$ is the probability that there will be a match between a feature value for that choice word and the corresponding feature value in the perceived vector conditional on the choice-word actually being the target, $p_n$ is the probability that there will be a mismatch between a feature value for that choice word and the corresponding feature value in the perceived vector conditional on the choice word being the target, $M$ is the number of diagnostic feature values that match for that choice word, and $L'$ is the number of diagnostic features in **w**.

Consider next the prior probability, $p_0$, which we set equal to $p_{0E}p_{0L}$, where the first term refers to the factors in place in the test

setting, including differential probabilities of testing the two alternatives, payoffs, instructions, and so forth, and the second refers to differences found in the lexical representations of the two alternatives. The experimental design has been chosen to make the two alternatives equally likely, so $p_{OE}$ is set to .5 for both alternatives. We have assumed that with probability $\alpha$ the effect of prior study is to add one context-matching feature to the lexical representation of the studied word. It is convenient to assume that the evidentiary value of such an extra context feature is equal to that for one visual feature. Thus, an alternative whose lexical trace has an extra context feature that does not match that in the current test environment has $p_{OL} = p_n$, and an alternative with an extra context feature that does match has $p_{OL} = p_m$. Because $p_m$ is larger than $p_n$, it is quite obvious from Equations 3 and 4 that the alternative with the larger number of matching features (including the visual features and the extra context feature) will have the higher probability and hence be the one chosen in a maximum-likelihood decision. Thus, the maximum-likelihood decision rule is quite simple: Choose the alternative with the greater number of matching features; if the number is tied, guess with probability .5.

Although this simple rule combines the two kinds of evidence, to maintain contact with prior formulations such as the rational analysis of cognition (Anderson, 1990) it is useful to keep in mind that the features are actually of two kinds: $L'$ visual features that act as the likelihood, and one context feature that determines the prior.

Given this simple decision rule, the probability of a correct choice requires only the calculation of the probabilities that the target has more matches than the foil, termed $P_N(C)$, $P_T(C)$, and $P_F(C)$ for *neither studied, target studied,* and *foil studied,* respectively. The formulas for calculating these probabilities are given in Appendix A. To illustrate the model, the distribution of the difference between the number of matches for the target and the foil is graphed in Figure 1 for the neither-studied case and for the foil-studied cases. The distribution to the right of zero, plus one half of the zero value, is the probability correct in each case. Panel A shows the difference distributions for the similar and dissimilar cases, assuming neither choice had been studied. Panel B repeats the similar distribution from Panel A, and also shows the distribution when the foil had been studied. It can be seen that the study of a foil simply shifts the distribution a fraction to the left. If the target had been studied, the distribution would have shifted rightward instead. Panel C is similar to B but shows the situation for dissimilar words. The implications of the differences between similar and dissimilar choices are discussed next.

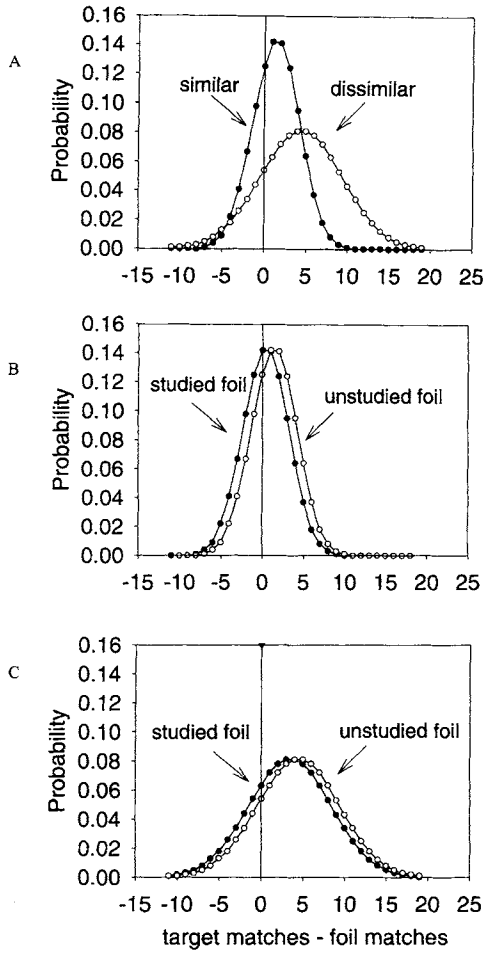


*Figure 1.* (A) Distributions of the number of perceived diagnostic features that match the target minus the number that match the foil. Filled circles illustrate the case when the alternatives are similar; open circles when the alternatives are dissimilar. (B) For the similar alternative case: Filled circles illustrate that the effect of studying the foil is to add (part of) a match, thereby shifting the difference distribution to the left and changing the probability of choosing the target because the area to the right of zero decreases. Studying the target would instead cause the distribution to shift right. (C) For the dissimilar case: The effect of study is like that for Panel B, but the change in choice probability is less because the dissimilar distribution has higher variance.

### *The Interaction Between Similarity and Prior Study*

Why prior study influences decisions between similar items more than dissimilar ones is fairly easy to understand with the help of Figure 1. The dissimilar condition has more diagnostic features. Diagnostic features are those that differ between **f** and **t**. For example, if we take letters to be features, when the alternatives are *lied* and *died,* only the *l* and *d* are diagnostic. When the alternatives are *lied* and *sofa,* all of the letters are diagnostic. (In a more realistic example, even the dissimilar condition will have some nondiagnostic features.) The diagnostic features are the only features we need to be concerned with in a forced-choice task, because a common feature, such as *d* in the *lied–died* example, will either match the perceived vector for both alternatives or will mismatch the perceived vector for both alternatives. In either case, the matching result provides no differential evidence favoring either alternative.

Because the similar condition has fewer diagnostic features, its difference distribution is narrower than that for the dissimilar condition (as depicted in Panel A). The critical point, however, is that the positive or negative shift of the distribution due to the possible presence of a matching context feature is of the same size regardless of the similarity between the choices (as depicted in Panels B and C). The figure makes it clear that such a constant shift will have a greater impact on $P(C)$ (i.e., the portion of the distribution at or above zero) for the similar case, when there are

fewer diagnostic features. In effect, the match of context is weighted more strongly when there is little other evidence to consider. Consider two extreme cases: If the choice depended on a single diagnostic visual feature, then the addition of an extra match because of context would markedly affect the decision; if the choice depended on 1,000 visual features, the addition of an extra match because of context would not noticeably affect the decision. In short, the extra matching feature behaves just like a weak prior in a Bayesian inference.

Panel A of Figure 2 shows parametric data from Ratcliff and McKoon (1997, Experiment 5). They varied flash time, whether the alternatives were similar or dissimilar, and which, if either, of the alternatives were studied. Note that for the similar conditions the effects of prior study are present at all flash times, even when performance is near chance. For the dissimilar alternatives, there is some bias at the shorter flash times, but next to none at the longest flash time.

We fit REMI to the forced-choice data from Ratcliff and McKoon's (1997) Experiment 5 with the program TORUS (Rabinowitz, 1995), which minimizes the total sum of squared error. Panel B of Figure 2 shows the best fit to the data. The model fit depends on seven parameters, whose best fitting values are given in Table 2. The first two parameters are the number of diagnostic features for the dissimilar ($L_d$) and similar ($L_s$) alternatives. The third, $\alpha$, is the probability of an additional matching context feature for a studied word. Parameters $s_{10}$, $s_{20}$, and $s_{40}$ are the probabilities of seeing a feature at the three flash times of 10, 20, and 40 ms, respectively. The last parameter, $g$, governs the geometric distribution of features for the target, foil, and perceptual noise, and therefore determines the probability that a feature value resulting from perceptual noise will match a visual feature value of a word by chance. The specific choice of the geometric distribution is critical only for those applications involving word frequency (because these depend on the frequency of various features). For the other applications, uniform feature distributions result in fits that are as good as those reported here.

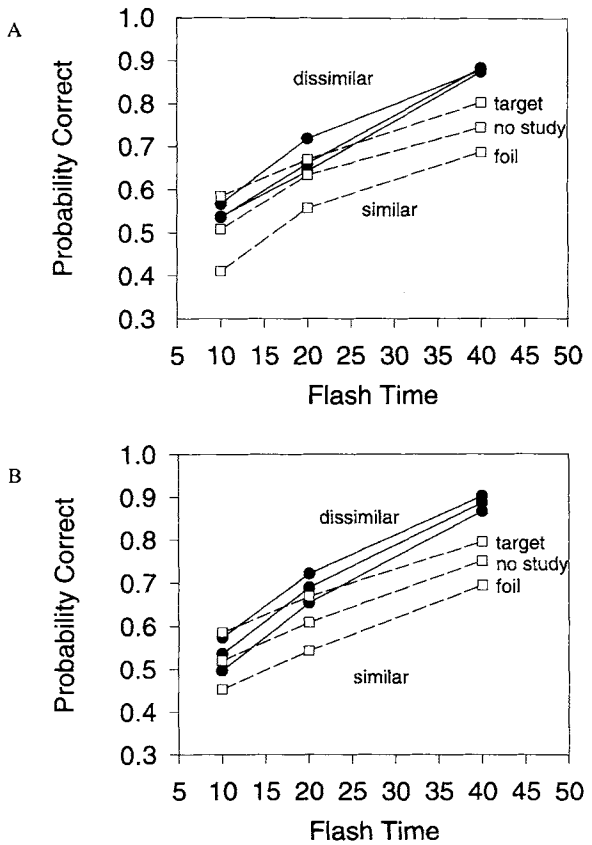


*Figure 2.* (A) Data from Ratcliff and McKoon's (1997) Experiment 5: Circles illustrate dissimilar choices, and squares similar choices. In each of these conditions, upper points represent target study, middle points no study, and lower points foil study. (B) The fit of REMI to the data shown in Panel A (parameters are given in the first column of Table 2).

The fit captures the qualitative patterns in the data: In particular, the effects of prior study are predicted to be larger when the alternatives are similar than when they are dissimilar. Also, as flash duration increases, bias decreases. Table 3 shows the predicted and observed performance. The influences of prior study for the studied conditions are reported in the Data bias and REMI bias columns. Bias is calculated by subtracting baseline performance (i.e., the no-study condition) from performance in the studied conditions. To get more stable estimates of this bias, Table 4 shows the average of the magnitude of the bias observed when the foil was studied and when the target was studied. The bottom row of Table 4 shows whether the model overestimates (positive values) or underestimates (negative values) the bias observed in the data. At a 10-ms flash time, the model exhibits too much bias when the alternatives are dissimilar and too little when they are similar. It should be noted, however, that at 20-ms flash duration the predictions slightly underestimate the bias for the dissimilar alternatives. Additional evidence suggests a small effect of prior study in the dissimilar case. Masson (personal communication, June 7, 1997) has run experiments like those of Ratcliff and McKoon (1997). He found that when the flashed word had been studied, the probability of correctly choosing that word over a dissimilar alternative was .78, and for unstudied words it was .76. Though not statistically significant, the amount of bias that Masson (personal communication) observed is consistent both with the small amount of predicted bias and with that found in Ratcliff and McKoon's (1997) data.

### *Extensions to Word Frequency Effects*

Ratcliff and McKoon (1997) carried out post hoc tests to examine the effect of word frequency. They found high–high (HH) frequency pairs produced almost equal performance to low–low (LL) frequency pairs, but in mixed pairs the high frequency alternative tended to be chosen. The present model comes close to predicting this pattern, based on the assumption that high frequency words have more common features, that is, feature values generated with a higher value of $g$ than that for low-frequency words (Shiffrin & Steyvers, 1997). More common feature values tend to be matched more often by the visual noise in the perceived vector. Thus in mixed pairs, extra (chance) matches favor the high frequency word. However, in same-frequency pairs (HH or LL) there is only a slight effect: The extra matches for high frequency pairs add a little more noise, leading to a very slight advantage for LL pairs. These data and the predictions with the assumptions that

Table 2
*Best Fitting Parameters of REMI for Forced Choice (FC), Yes–No, and Naming*

| Parameter | FC | FC feature frequency | FC revised | Yes–No | Naming |
|---|---|---|---|---|---|
| $L_d$ | 44 | 44 | | 44 | |
| $L_s$ | 14 | 14 | 14 | 14 | 0.14 |
| $g$ | 0.4 | 0.4 | | 0.4 | 0.4 |
| $g_L$ | | 0.36 | | | 0.36 |
| $g_H$ | | 0.43 | | | 0.43 |
| $u$ | | | 4 | | |
| $s_{av}$ | | 0.073 | | | |
| $s_{10}$ | 0.00913 | | | | |
| $s_{20}$ | 0.05161 | | 0.209 | | |
| $s_{28}$ | | | | | |
| $s_{30}$ | | | | .1173 | |
| $s_{35}$ | | | | | 0.23 |
| $s_{40}$ | 0.1306 | | | | |
| $\alpha$ | 0.4411 | | 0.874 | 0.4411 | 0.85 |
| $\alpha_H$ | | | 0.709 | | |
| L | | | 27 | 55 | 55 |
| $C_{FC}$ | | | 10 | | |
| $C_{YN}$ | | | | 15 | |
| $C_{RC}$ | | | | | 24 |
| $b$ | | | | | 0.75 |
| Neighborhood Size | | | | | 100 |

*Note.* A blank cell indicates that the parameter of that row does not affect the model of that column. REMI = retrieving effectively from memory, implicit.

the value of $g$ for high-frequency words is .43, that the value of $g$ for low-frequency words is .36, and that the value of $s$ is .073, are shown in Table 5.

The good fit to the data in Table 5, however, is not the whole story. The post hoc tests carried out by Ratcliff and McKoon (1997) had little power, and subsequent data collected by Bowers (1999), by Wagenmakers et al. (2000), and by McKoon and Ratcliff (in press) demonstrated a small but statistically reliable advantage for HH pairs over LL pairs (perhaps of 2–4%). This advantage cannot be fit by the present model. (In fact, if a different model were used, more like the REM model of Shiffrin and Steyvers (1997), in which the diagnosticity of feature matches is taken into account, the predicted advantage of LL pairs would be even larger). Thus, the advantage of HH pairs over LL pairs is outside the present model. After our discussion of the REMI models of perceptual identification in yes–no and naming, we will discuss a revised version of the forced-choice model that can predict the effects of word frequency (and other variables).

## The REMI Model Applied to a Yes–No Task

In a yes–no paradigm, a word is flashed and masked, a test word is presented clearly, and the participant says whether the flashed word and the test word were the same. In Ratcliff and McKoon's (1997) experiments, the test word was the flashed word, a word that was visually similar to the flashed word, or a word that was visually dissimilar to the flashed word; in addition, they varied whether the test word had been previously studied.

We adopt a decision model similar to that used for forced choice: The perceived vector is compared with the alternatives in memory, but the decision is based on the matches restricted to the single test word (in forced choice, the decision was restricted to the matches for the two choices). Thus the term $p_0(\mathbf{w})p(\mathbf{d}|\mathbf{w})$ is evaluated, where $\mathbf{w}$ represents the test word and $\mathbf{d}$ the perceived vector. Again we represent the effect of prior study in the term $p_0(\mathbf{w})$: There is one extra context feature and this matches the test item, if the test item had been studied, with probability $\alpha$. This context match is taken into account in this prior, and the effects of perception are represented in the likelihood term, $p(\mathbf{d}|\mathbf{w})$. If the product of the prior and likelihood exceeds a criterion, then the system decides that the test word was flashed. As in the forced-

Table 3
*Results From Ratcliff and McKoon (1997, Experiment 5) and REMI Predictions*

| Study | Flash | Similarity | Data | REMI | Data bias | REMI bias |
|---|---|---|---|---|---|---|
| Target | 10 | Similar | 0.585 | 0.5860 | 0.076 | 0.066 |
| Neither | 10 | Similar | 0.509 | 0.5200 | 0.0000 | 0 |
| Foil | 10 | Similar | 0.411 | 0.4527 | −0.098 | −0.0673 |
| Target | 10 | Dissimilar | 0.567 | 0.5740 | 0.029 | 0.0383 |
| Neither | 10 | Dissimilar | 0.538 | 0.5357 | 0 | 0 |
| Foil | 10 | Dissimilar | 0.536 | 0.4966 | −0.002 | −0.0391 |
| Target | 20 | Similar | 0.67 | 0.6689 | 0.036 | 0.0597 |
| Neither | 20 | Similar | 0.634 | 0.6092 | 0 | 0 |
| Foil | 20 | Similar | 0.558 | 0.5430 | −0.076 | −0.0662 |
| Target | 20 | Dissimilar | 0.719 | 0.7215 | 0.058 | 0.0318 |
| Neither | 20 | Dissimilar | 0.661 | 0.6897 | 0 | 0 |
| Foil | 20 | Dissimilar | 0.643 | 0.6543 | −0.018 | −0.0354 |
| Target | 40 | Similar | 0.804 | 0.7960 | 0.058 | 0.0444 |
| Neither | 40 | Similar | 0.746 | 0.7516 | 0 | 0 |
| Foil | 40 | Similar | 0.688 | 0.6948 | −0.058 | −0.0568 |
| Target | 40 | Dissimilar | 0.879 | 0.9027 | 0.005 | 0.0155 |
| Neither | 40 | Dissimilar | 0.874 | 0.8872 | 0 | 0 |
| Foil | 40 | Dissimilar | 0.884 | 0.8672 | 0.01 | −0.02 |

*Note.* REMI = retrieving effectively from memory, implicit.

Table 4
*Average Bias in Ratcliff and McKoon (1997, Experiment 5) and REMI Predictions*

| | 10 | | 20 | | 40 | |
|---|---|---|---|---|---|---|
| Source | Dissimilar | Similar | Dissimilar | Similar | Dissimilar | Similar |
| Data | 0.0155 | 0.0870 | 0.0380 | 0.0560 | 0.0075 | 0.0580 |
| REMI | 0.0387 | 0.0667 | 0.0336 | 0.0630 | 0.0178 | 0.0506 |
| Error | 0.0232 | −0.0204 | −0.0044 | 0.0070 | 0.0103 | −0.0074 |

*Note.* REMI = retrieving effectively from memory, implicit.

choice case, this decision rule comes down to counting matching features, including the extra context-matching feature, if there is one. There is a critical difference from the forced-choice situation, however: Because there is only a single test word, all $L$ visual features are diagnostic. Thus the decision rule counts all matches; if the total count exceeds a criterion, $C_{YN}$, a decision is made that the test word was flashed. Because only some features were diagnostic in forced choice, for both similar and dissimilar conditions, the parameter $L$ in yes–no that represents total vector length must be estimated (and is expected to be larger than $L_d$). The parameters $L_s$ and $L_d$ still play a role, because they define the similarity (i.e., feature overlap) between the flashed word and similar or dissimilar test words. The values of these two parameters were carried over from the fit to the forced-choice data. Finally, a new flash time parameter, $s_{30}$, corresponds to the 30-ms presentation time used in Ratcliff and McKoon's (1997) Experiment 8. The value of $s_{30}$ was constrained to be between the values of $s_{20}$ and $s_{40}$ that were obtained in the fit to the forced-choice results. The parameter values $g$ and $\alpha$ were carried over from the fit of the forced-choice model. The best fitting parameter values are given in Table 2. The mathematical and computational details are given in Appendix B.

Table 6 shows the results from Ratcliff and McKoon (1997, Experiment 8) along with the REMI fits to the data. The first column shows the relation of the test word to the flashed word. The Study column indicates whether the test word had been studied. The remaining columns show Ratcliff and McKoon's (1997) data and the fits of REMI. The fit looks reasonable, though the model slightly underestimates the effect of prior study when the flashed word and the test word are the same, and overestimates this effect when the test word is similar (*lied*) or dissimilar (*sofa*) to the flashed word (*died*).

## An Alternative Interpretation of the REMI Model for Forced Choice and Yes–No

We have suggested that the perceived vector is compared with the set of lexical–semantic images, with a subsequent restriction to the two images presented in forced choice, or the one image presented in yes–no. Alternatively, the perceived vector could be held in some form of visual short-term memory until the alternatives are presented. The alternatives could then be read, their lexical–semantic images contacted, and some of the information found in those images, including context, retrieved. In this model, the perceived vector would be compared with the vectors constructed when the alternatives are read. For present purposes, these models make identical predictions, but the conceptual basis for the models is different. In the alternative conception prior study has its effect when the alternatives are presented, rather than in association with the flash, so future research could provide tests of the distinction.

## REMI Applied to a Naming Task

In a typical threshold-naming paradigm, a word is flashed and masked, and the participant attempts to name the word. The REMI model for naming is very similar to the model of forced choice. In naming, the vector of perceived and context features is compared against all the lexical–semantic images in memory, as before, but all of these are relevant instead of just the two (forced choice) or one (yes–no) presented alternatives. The probability that any particular word was flashed is given by Equation 2, where the sum in the denominator is taken over the words in the lexicon. The word most likely to have been presented is the one with the highest number of matches to the perceived vector. However, in a recall task there are good reasons not to emit the word with the highest

Table 5
*Illustrative Word Frequency Effects in the REMI Model*

| Target frequency | Foil frequency | Model | Data |
|---|---|---|---|
| Normal | Normal | .701 | |
| High | High | .697 | .701 |
| Low | Low | .705 | .696 |
| High | Low | .766 | .727 |
| Low | High | .627 | .613 |

*Note.* $g_h$ = .43 and $g_l$ = .36., $s$ = .073. Collapsed data from Ratcliff and McKoon (1997). REMI = retrieving effectively from memory, implicit.

Table 6
*Data From Ratcliff and McKoon (1997, Experiment 8), and REMI Predictions*

| Decision word | Study | Data | REMI |
|---|---|---|---|
| Same | No | .70 | 0.699 |
| Same | Yes | .77 | 0.741 |
| Similar | No | .51 | 0.516 |
| Similar | Yes | .55 | 0.568 |
| Dissimilar | No | .21 | 0.140 |
| Dissimilar | Yes | .23 | 0.176 |

number of matches unless the number of matches exceeds a criterion, $C_{RC}$. When the number of alternative responses is very high, as it is in a naming task, the probability of a correct identification is very low unless many features are perceived correctly. Thus, if the best alternative from the lexicon has only a few matches to the perceived vector, the response emitted has a very low probability of being right.

We believe that participants in naming tasks implicitly assume that emitted responses should have a reasonable chance of being correct. We therefore assume that a criterion number of matches must be exceeded before a response is emitted. This will ensure that the responses are correct on an acceptable proportion of the trials. We have described this process as a purely explicit one, but it may be that the system simply does not provide an alternative to be emitted unless a criterion is reached. By preventing the system from always making a response, the criterion functions like the recall criterion in the rational analysis of memory (Anderson, 1990), in which when the expected gain of retrieving an image fell below the expected costs of processing it, it was not worth retrieving.

Predicting naming performance requires specifying the similarity structure of the lexicon because our model depends on feature overlap. That is, the number of matches for various words in the lexicon will depend on the number of features shared between the flashed word and each entry in the lexicon. The question is how to produce such a specification without introducing an explosion of parameters. One approach is to assume that the similarities of these images to the target are distributed according to Zipf's law (Ijiri & Simon, 1977). A close approximation of Zipf's law is $f(j) = rj^{b}$, where $j$ is the rank (e.g., the most similar word, second most similar, etc.), $f(j)$ is the similarity of the word with rank $j$, $b$ is a constant, and $r$ is a scale factor. Such a system has only a few very similar words and many dissimilar ones. In Appendix C, we present an analysis of words like those used in Ratcliff and McKoon's (1997) study, showing that Zipf's law provides a good approximation.

Predictions for naming were obtained by simulation: To start, a target vector of length $L$ was randomly generated with values drawn from a geometric distribution with parameter $g$. Unless otherwise noted, the parameters were carried over directly from the previous simulations. Each target had $n$ near neighbors, and 100-$n$ distant neighbors. *Near neighbors* are words that are as similar to the target as the similar alternative in the forced-choice task. Ratcliff and McKoon (1997) referred to the group of a word's near neighbors as its "cohort." Because the parameter $r$ in Zipf's law equals the similarity of the most similar item, $r$ was set equal to $L - L_s$; thus, each near neighbor shared $L - L_s$ features with the target. The similarity of the remaining distant neighbors fell off as a function of rank, according to Zipf's law, with b set to .75. For example, if there were 3 near neighbors, each would share $L - L_s$ features with the target. The first distant neighbor would share $f(2)$ features, the second $f(3)$ features, and so on.[1]

Next, a vector of perceived features, **d**, was constructed as in the other models. The flash-time parameter, $s_{35}$, corresponds to the 35-ms presentation time used in Ratcliff and McKoon's (1997) Experiment 6. The value of $s_{35}$ was constrained to be between the values of $s_{30}$ and $s_{40}$ obtained from the fit of the yes–no and forced-choice REMI models, respectively. The perceived vector **d** is compared against the target, **t**, and against the 100 lexical–semantic images that represent the target's neighborhood. The response selected is the image with the largest number of matches if that number exceeded $C_{RC}$, or else no response is given. Some of the simulations modeled conditions in which the target or one of the near neighbors was studied. To the vector representing any studied item, an extra context feature that matches current context was added with probability $\alpha$; the value of $\alpha$ was initially set to that used in the previous simulations.

Ratcliff and McKoon (1997) list a series of effects that a model of naming should explain: (a) Studying a target should increase the probability of identifying it; (b) studying a similar word should not increase the probability of identifying the target; (c) words with larger numbers of near neighbors should be identified less well than those from smaller neighborhoods; (d) high-frequency words should be better identified than low-frequency words; and (e) low-frequency words should show greater priming effects than high-frequency words.

In coming to this picture of word naming, Ratcliff and McKoon (1997) collapsed responses from many different conditions and experiments. The naming simulations depend on critical assumptions about the structure of the lexicon (e.g., the parameter $g_h$ used to generate high-frequency words, the numbers of near and distant neighbors, etc.) and on the details of Ratcliff and McKoon's (1997) analyses (e.g., the exact proportion of high-frequency words included, the number of near neighbors, etc.). Given these complexities, our aim is not to give a precise fit to Ratcliff and McKoon's (1997) data, but rather to show that REMI's behavior follows, for the most part, Ratcliff and McKoon's (1997) broad prescriptions for how a model for naming should behave.

We carried out 27 naming simulations by fully crossing the variables of study (study target, study similar, or study neither), target frequency (low, normal, or high), and neighborhood size (2, 7, or 15 near neighbors). We ran the simulation 1,000 times for each combination of these variables. The responses were grouped into the four categories used by Ratcliff and McKoon (1997): correctly naming the flashed target, saying nothing at all, intruding with the similar alternative used in the forced-choice paradigm, or intruding with some other word. One of the near neighbors was randomly selected as the similar alternative, so if another of the near neighbors was chosen it was grouped in the other category. Table 7 shows data from their Experiment 6 broken into these categories. The model was fit to the data (by hand) with parameter values constrained by the results of the 2AFC, yes–no, and word frequency applications. There were two main problems with the predictions. First, the predicted effect of prior study was smaller than that observed (about 5% as compared with 14%), and there were too many intrusions of the similar alternative (about 5% as compared with 1%). The predictions can be brought into line with the data by adjusting the values of two parameters: raising $s_{35}$ from

[1] Appendix C gives empirical estimates of $r$ and $b$ (e.g., 2.5 and .253, respectively), whereas in the simulations, we used values of 41 and .75. For REMI, similarity is measured in terms of the number of shared features, whereas for the word analysis, similarity is measured in terms of a rough measure of letter shape. For REMI, similar items share 75% (41 of 55) of the features. For the word analysis, the maximal level of dissimilarity is 20, so differing by 2.5 in that metric means that similar words share 88% of letter shape.

Table 7
*Results of Ratcliff and McKoon's (1997) Experiment 6*

| Study | Response | | |
|---|---|---|---|
| | Target | Similar | Other |
| Target | .42 | .01 | .14 |
| Similar | .28 | .03 | .19 |
| Neither | .27 | .01 | .19 |

.13 to .23, and raising $\alpha$ from .406 to .85. These parameter values lead to the simulation results presented in Table 8, which are more in line with the data. One interpretation of these changes is that the forced-choice and yes–no tasks introduce additional noise that is not present in naming. Such noise could be perceptual because, in forced choice and yes–no, the presentation of the alternatives could cause additional interference, or could be due to forgetting because of the time delay until the alternatives are presented and read. Tables 9, 10, and 11 show the results of the REMI naming simulations.

Ratcliff and McKoon (1997) showed that the probability of naming the target falls off as the number of neighbors increases and the target's word frequency decreases (Tables 12 and 13). In the REMI naming simulations the target had 2, 7, or 15 near neighbors. Tables 14 and 15 show the naming simulation results collapsed across study conditions. The performance of the model does indeed fall as the number of neighbors increases, because of increased probability of responding with one of the neighbors. Averaging over the conditions in their Experiment 9, they found that the probability of a correct response was .333 for high-frequency words, and .237 for low-frequency words (Table 12). Collapsing across the corresponding simulation conditions from Table 14, the probabilities of a correct response for high- and low-frequency targets are .368 and .260, respectively. The reason REMI predicts an advantage for high-frequency words is that they tend to have more common features, which increases the probability of a chance match.

Ratcliff and McKoon (1997) also asserted that their counter model predicts greater priming effects for low-frequency words than for high-frequency words. Collapsing across their Experiments 6 and 9, they found that baseline performance for low- and high-frequency words was .20 and .30. When the target was studied, performance rose to .40 for low-frequency words and to .43 for high-frequency words. Another way to characterize these results is to say that a single prior study removes the effects of frequency. The present version of REMI cannot fit this pattern of results. Baseline performance for REMI for low- and high-frequency words is .227 and .335, respectively. Prior study improves performance by about .1 in both conditions to .330 and .435, respectively. In the section on forced choice, we noted that REMI requires additional mechanisms to predict a perception facilitation due to priming, and an advantage of high-frequency pairs over low-frequency pairs. It is conceivable that such a mechanism could also produce a larger gain for priming of low-

Table 8
*Fits of the REMI Simulation to Ratcliff and McKoon's (1997) Experiment 6*

| Study | Response | | |
|---|---|---|---|
| | Target | Similar | Other |
| Target | 0.387 | 0.012 | 0.082 |
| Similar | 0.280 | 0.037 | 0.118 |
| Neither | 0.283 | 0.017 | 0.114 |

*Note.* $s_{35} = .23$ and $\alpha = .85$.

Table 9
*REMI Simulation Results for Combinations of Near Neighbors, Study, and Word Frequency*

| Near neighbors | Word frequency | Response | | |
|---|---|---|---|---|
| | | Target | Similar | Other |
| | | Study nothing | | |
| 2 | Low | 0.254 | 0.019 | 0.023 |
| 2 | Normal | 0.342 | 0.019 | 0.021 |
| 2 | High | 0.383 | 0.016 | 0.030 |
| 7 | Low | 0.232 | 0.014 | 0.106 |
| 7 | Normal | 0.283 | 0.023 | 0.099 |
| 7 | High | 0.347 | 0.022 | 0.113 |
| 15 | Low | 0.196 | 0.014 | 0.216 |
| 15 | Normal | 0.239 | 0.015 | 0.196 |
| 15 | High | 0.276 | 0.016 | 0.221 |
| | | Study target | | |
| 2 | Low | 0.357 | 0.013 | 0.008 |
| 2 | Normal | 0.428 | 0.010 | 0.016 |
| 2 | High | 0.470 | 0.015 | 0.015 |
| 7 | Low | 0.333 | 0.014 | 0.081 |
| 7 | Normal | 0.417 | 0.011 | 0.072 |
| 7 | High | 0.441 | 0.012 | 0.078 |
| 15 | Low | 0.300 | 0.013 | 0.143 |
| 15 | Normal | 0.343 | 0.014 | 0.159 |
| 15 | High | 0.394 | 0.009 | 0.171 |
| | | Study similar | | |
| 2 | Low | 0.277 | 0.048 | 0.017 |
| 2 | Normal | 0.333 | 0.038 | 0.017 |
| 2 | High | 0.403 | 0.052 | 0.021 |
| 7 | Low | 0.219 | 0.034 | 0.085 |
| 7 | Normal | 0.304 | 0.036 | 0.104 |
| 7 | High | 0.328 | 0.031 | 0.104 |
| 15 | Low | 0.173 | 0.034 | 0.214 |
| 15 | Normal | 0.209 | 0.025 | 0.256 |
| 15 | High | 0.270 | 0.035 | 0.248 |

Table 10
*Probabilities of Responding With the Target in Naming Averaged Over Study Conditions in Ratcliff and McKoon's (1997) Experiment 9*

| Target frequency | Number of neighbors | | | |
|---|---|---|---|---|
| | 1–4 | 5–8 | 9 and over | Marginal |
| Low | 0.34 | 0.21 | 0.16 | 0.237 |
| High | 0.37 | 0.33 | 0.30 | 0.333 |
| Marginal | 0.355 | 0.27 | 0.23 | — |

Table 11
*Probabilities of Responding With Nontargets in Naming Averaged Over Study Conditions in Ratcliff and McKoon's (1997) Experiment 9*

| Response | Target frequency | Number of neighbors 1–4 | 5–8 | 9 and over |
|---|---|---|---|---|
| Similar | Low | 0.06 | 0.02 | 0.02 |
| Similar | High | 0.04 | 0.01 | 0.01 |
| Other | Low | 0.077 | 0.071 | 0.146 |
| Other | High | — | — | — |

frequency than high-frequency words. We return to this issue shortly.

The counter model handles word frequency effects by raising the resting levels of the counters, such that high-frequency words have a head start over their lower frequency counterparts. The effects of prior study result from previously studied words having an increased likelihood of stealing noise counts. It is not clear to us why the counter model in its original form predicts that an increase in the probability of stealing a noise feature should mitigate the advantage that high-frequency words hold over low-frequency words (however, McKoon & Ratcliff, in press, presented a revised version of the counter model that probably would do so).

## A Revised REMI Model of Forced Choice

A direct way to implement frequency effects in REMI is through the prior. That is, the prior should reflect the observation that high-frequency words are more likely to appear in the environment than low-frequency words. In the mixed case, this would bias the response toward the high-frequency alternative. However, in the current version of the model, whenever the alternatives were of the same frequency, the priors would cancel, so performance would be the same for low-frequency and high-frequency pairs. A similar argument holds for why our model does not predict performance gains when both alternatives have been studied.

Although these results would be consistent with the analyses of Ratcliff and McKoon (1997), they differ from more recent findings. Masson and MacLeod (1996), in seven forced-choice perceptual identification experiments, found that when both alternatives had been previously studied, performance was on average 1.8% better than when neither word had been studied, though in only one of these experiments was this difference statistically reliable. Raaijmakers, Schooler, and Shiffrin (1997), using a forced-choice perceptual identification task, varied the number of times in the experiment that test pairs were presented. Participants performed 3.9% better when tested with pairs that had five prior presentations than when tested with novel pairs. This difference was statistically significant. As discussed earlier, others have found performance advantages for high-frequency pairs compared with low-frequency pairs. In particular, Wagenmakers et al. (2000) provided excellent data for testing purposes. They systematically varied the frequency of the target and foil, and whether the target, foil, neither, or both had been studied. The results are presented in Table 14. The striking result is that accuracy is better when both alternatives are high-frequency words, and that prior study of both words leads to improved performance for low-frequency words.

We investigated a number of ways to augment the REMI model to predict these findings, and discovered a number of different models that appear capable of doing so, albeit at the cost of additional assumptions, processes, and parameters. By way of example, we provide one such model in this section but emphasize that it would be premature to commit to any one version at this time.

The approach we present incorporates a decision criterion into the REMI forced-choice model (bringing it more in line with models of word naming and yes–no performance). The revised forced-choice model has two plausible assumptions. The first is that there is a higher prior probability of encountering high-frequency words than low-frequency words. This is implemented in terms of high-frequency words sometimes receiving an extra context match (with probability $\alpha_H$). The second assumption is that

Table 12
*Probabilities of REMI Simulation Responding With the Target in Naming Averaged over Study Conditions*

| Target frequency | Number of neighbors 1–4 | 5–8 | 9 and over | Marginal |
|---|---|---|---|---|
| Low | 0.296 | 0.261 | 0.223 | 0.260 |
| High | 0.419 | 0.372 | 0.313 | 0.368 |
| Marginal | 0.358 | 0.317 | 0.268 | |

Table 13
*Probabilities of REMI Simulation Responding With Nontarget in Naming Averaged Over Study Conditions*

| Response | Target frequency | Number of neighbors 2 | 7 | 15 |
|---|---|---|---|---|
| Similar | Low | 0.026 | 0.020 | 0.020 |
| Similar | High | 0.028 | 0.022 | 0.020 |
| Other | Low | 0.016 | 0.091 | 0.191 |
| Other | High | 0.022 | 0.098 | 0.213 |

Table 14
*Proportion of Correctly Identified Targets as a Function of Word Frequency of Target and Foil and Study Condition from Wagenmakers, Zeelenberg, and Raaijmakers (2000)*

| Target | Foil | Study Target | Foil | Neither | Both |
|---|---|---|---|---|---|
| HF | LF | 0.87 | 0.773 | 0.868 | 0.842 |
| LF | HF | 0.842 | 0.683 | 0.757 | 0.793 |
| HF | HF | 0.862 | 0.753 | 0.816 | 0.822 |
| LF | LF | 0.874 | 0.72 | 0.765 | 0.821 |

*Note.* HF = high frequency word; LF = low frequency word. Adapted from Table 1 in "Testing the Counter Model for Perceptual Identification: Effects of Repetition Priming and Word Frequency," by E. M. Wagenmakers, R. Zeelenberg, and J. G. W. Raaijmakers, 2000, *Psychonomic Bulletin & Review, 7,* p. 664. Copyright 2000 by the Psychonomic Society.

Table 15
*REMI Fit of Wagenmakers, Zeelenberg, and Raaijmakers (2000)*

| | | Study | | | |
|---|---|---|---|---|---|
| Target | Foil | Target | Foil | Neither | Both |
| HF | LF | 0.903 | 0.797 | 0.845 | 0.867 |
| LF | HF | 0.821 | 0.663 | 0.740 | 0.758 |
| HF | HF | 0.874 | 0.740 | 0.806 | 0.822 |
| LF | LF | 0.858 | 0.729 | 0.786 | 0.813 |

*Note.* HF = high frequency word; LF = low frequency word.

response is chosen when one of the alternatives matches more features than the other, and it has at least a criterion number of matches, $C_{FC}$; otherwise, the model guesses. The addition of the response criterion means that the nondiagnostic features also play a role. Previously, responses were based only on the number of diagnostic matches.

These changes allow the model to capture the qualitative pattern of results in the Wagenmakers et al. (in press) data: More than a criterion number of matches is needed to avoid guessing. Nonstudied high-frequency words will exceed this criterion more than nonstudied low-frequency words because of the extra matching feature for high-frequency words. Prior study of both might tend to push both high and low frequency words above criterion, reducing the frequency advantage.

We applied this forced-choice model with criterion to the Wagenmakers et al. (2000) data. Because these fits are for illustrative purposes only, they were not constrained by the previous fits. For simplicity we assumed a uniform distribution of four features, a probability $\alpha$ that a previously studied word will get an extra match and a probability $\alpha_H$ that a high-frequency word will get an extra match. The parameter values are in Table 2 under the FC revised column, and the resulting fits are presented in Table 15. The revised model fails to predict the magnitude of all of the effects, but does predict the qualitative pattern correctly. In particular, the model shows an improvement in performance with practice and word frequency. It should be noted that the model predicts improvement with study for both high-frequency (1.6%) and low-frequency (2.7%) words, but in the data the corresponding gains are 0.6% and 5.6%. This prediction could be put more in line with the data if the reasonable assumption were made that for high-frequency words the extra context features due to study and frequency were correlated. That is, the extra context feature that matches current context for a high-frequency item is presumably due to that word's recent occurrence in the environment, possibly causing storage of the same context feature that is stored in conjunction with a recent study episode in the experiment.

In this revised model, it is difficult to know whether to attribute REMI's predicted performance gains for high-frequency pairs and studied pairs to bias or improved perception. The answer might hinge on research that establishes what factors cause or enable adjustment of the criterion, $C_{FC}$. If this criterion can be adjusted by instructions and other factors, then it would be tempting to conclude that perception is not being affected. This interpretation would be consistent with the fact that the parameter $s$ has been held constant for high-frequency, low-frequency, studied, and unstudied words, and with the fact that in the revised model what varies between high-frequency and low-frequency and studied and unstudied words is $\alpha$ and $\alpha_H$, which in the previous applications was responsible for bias due to prior study. However, if the criterion is a relatively fixed part of the perceptual system, and cannot be altered, then the interaction of such a threshold with experimental manipulations like study could be viewed as a form of perceptual change. More generally, we agree with the following assessment:

> Hintzman (1990) has pointed out that these kinds of labels and distinctions become irrelevant once a model has been spelled out; then "the explanatory burden is carried by the nature of the proposed mechanisms and their interactions, not by what they are called" (p. 121). (Ratcliff & McKoon, 1997, p. 341)

## Comparison With ACT-R

Our Bayesian justification for the present model is similar to the rationale underlying ACT–R (e.g., Anderson et al., 1998; Anderson & Lebiere, 1998). ACT–R models procedural knowledge with sets of production rules (i.e., if–then rules) whose conditions (the *if* part) are matched against the contents of declarative memory. The fundamental declarative representation in ACT–R is the *chunk*, which is something like a proposition. Central to ACT–R is the idea that chunks take on activations. A chunk's activation is a combination of the underlying strength of the chunk, like $p(\mathbf{w})$ in Equation 1, and the activation it receives from associated active chunks. In ACT–R, activation quite explicitly represents the posterior probability (log-odds) that a particular chunk will fulfill a processing goal of the system. Its strength represents the prior probability that the chunk will meet a processing goal, before taking into account associations that the chunk might have with the current context.

By combining the ACT–R models of word-fragment completion and the word-superiority effect (e.g., McClelland & Rumelhart, 1981; Reicher, 1969; Wheeler, 1970), it seems likely that an ACT–R model of primed word identification could be constructed that could handle Ratcliff and McKoon's (1997) data. Implicit memory effects in stem completion are observed when prior study of a target word (e.g., *class*) increases the probability that a word-stem (e.g., *CL*_____) will be completed with the studied word as compared with a baseline condition in which the target had not been studied. In ACT–R, reading the word *class* strengthens the class chunk, so that when the participant is later presented with *CL*_____, residual activation increases the chance of responding with "class." The ACT–R model of stem completion essentially implements similar proposals put forward by Bower (1996) and by Reder and Gordon (1996).

ACT–R also has a well-worked-out model of visual perception (Anderson & Lebierre, 1998; Anderson & Matessa, 1992) that can handle the word superiority effect. The ACT–R model of visual perception allows for chance matches that result from "seeing" features that are absent in the stimulus. Such chance matches are critical to our accounts of Ratcliff and McKoon's (1997) data, and the ACT–R account of the word superiority effect. Potential interactions between residual chunk strength and ACT–R's model of visual perception means that it should be able to handle implicit memory effects in perceptual identification. It remains to be seen whether these pieces can be put together to handle Ratcliff and McKoon's (1997) results in detail.

This leads to the general question concerning what classes of models can handle the effects under discussion in this article. A successful model generally requires both a valid conceptual basis and an appropriate quantitative implementation. Fulfilling the second of these requirements under any circumstances is often far more difficult than external observers appreciate, and the rich set of data collected by Ratcliff and McKoon (1997) is likely to make attainment more difficult. At a more conceptual level, we have demonstrated with the REMI model that two basic and rather simple assumptions can enable the model to succeed (at least when implemented within an appropriate quantitative system): (a) Study makes the representation used for perception slightly more available, and (b) the system includes a rudimentary model of perception that includes some noise. It might be possible to incorporate these assumptions into other types of models and thereby adapt them to handle the present data. For example, TODAM2 (Murdock, 1993), and MINERVA 2 (Hintzman, 1986) are feature-based models that have been applied to explicit memory tasks. During retrieval, a noisy vector of features is retrieved, and for recall this vector must be de-blurred. Although these models are not entirely explicit concerning the process of de-blurring, one method matches the noisy vector to a lexicon of stored entries, searching for a good or best match. Such a process might be adaptable to threshold perception, in which a noisy set of input features must be matched to a lexicon. If study makes the entries in the lexicon slightly more accessible, the inherent noise in these models would seem to make them meet the minimal requirements for handling priming; however, making such models quantitatively adequate is another matter entirely. The details of a given implementation would determine whether it could predict the observed quantitative relations among performance for studied and unstudied words, similar and dissimilar words, yes–no and forced-choice tasks, and so on. Whether the two basic assumptions are required for any model applied successfully to the Ratcliff and McKoon (1997) data is an open question. At first glance it appears that Ratcliff and McKoon (1997) succeeded without invoking the assumption that study makes the lexical trace more available. However, they assumed that study changes the assignment of noise counts in favor of a studied alternative but did not propose a mechanism by which this occurs; if a mechanism were invoked, it might well be compatible with the assumption that study makes the lexical trace more available.

## General Discussion

Within the REM framework, explicit and implicit memory effects depend on similar storage and retrieval processes. Successful retrieval relies on matching the contextual and content features of the probe against the images required to make a response. In explicit tasks, retrieval is defined contextually, so the context cues play a large role. In implicit tasks, especially those in which the participant believes access to general knowledge is all that is needed to accomplish the stated goals, context cues play a lesser and indirect role: We propose that the memory probe always contains at least some current context because such context is omnipresent in the participant's external and internal environment. We propose also that lexical–semantic images gain some current context features when any item is studied, for similar reasons. These two factors combine to produce many implicit memory effects.

REM's account of retrieval has roots in Bayesian statistics: Shiffrin and Steyvers (1997) developed their recognition model by assuming a near-optimal probabilistic decision acting in an environment with a variety of constraints imposed by imperfect storage and noise; the present account also assumes that decisions are near-optimal given the constraints imposed by memory storage and perception. Given this approach, one can think of the effects of context in implicit memory as a reasonable attempt to incorporate prior odds into perception, retrieval, and decision making. Because context changes over time, the addition of contextual information to lexical–semantic images, and the use of context information in retrieval probes, can be seen as an appropriate adaptation to the environment. This is not to say that participants are oblivious to experiment-defined constraints; these also should be reflected in the priors. However, the two sorts of contributions to the priors ought in most cases to be distinguishable. For example, one response to an experimenter-imposed constraint might be an adjustment *late* in processing of the response criterion.

In our basic model, in the initial version of the counter model (Ratcliff & McKoon, 1997), and in the conclusions of Masson and MacLeod (1996), the quality of perceptual processing is unaffected by prior study. More recent data (e.g., Wagenmakers et al., 2000) has shown that priming both alternatives in a forced-choice task can produce a small performance increase relative to priming neither alternative. This result is most simply explained as being due to a perceptual gain in processing the flash. However, the revised model we discussed, which contains a response threshold, is able to explain such a result and could easily be interpreted to imply no change in quality of perceptual processing. This issue notwithstanding, any claim that perceptual processing is unaffected by prior study must be task dependent. When a word is presented once, one more study instance is added to a lifetime of such events, perhaps totaling 10,000. Even allowing for the adaptive role of recency (and its implementation in context storage and retrieval), it is easy to see why changes in the lexical–semantic representation due to study, and changes in subsequent retrieval due to that change, are small in magnitude. However, the study of a novel event, such as a nonword, may add a representation to a memory that contains no previous exemplars. In this case, it would be much less surprising to find clear demonstrations of perceptual processing gains, perhaps gains due to top-down facilitation of lower level processing. Just one of many examples would be the word-superiority effect (e.g., McClelland & Rumelhart, 1981; Reicher, 1969; Wheeler, 1970; see also Feustel, Shiffrin, & Salasoo, 1983; and Salasoo, Shiffrin, & Feustel, 1985).

Finally, a few words are in order concerning the relation between our model and the counter model of Ratcliff and McKoon (1997). The counter model predicts no effect for sufficiently dissimilar alternatives, whereas our model predicts a small effect for dissimilar alternatives. Whether real data would produce sufficiently powerful results to distinguish the models on this basis is an open question. Applied to the present tasks, both models assume no effect of prior study on the quality of initial perceptual processing. Both assume an important role for noise in producing the interaction of prior study and similarity: As a function of similarity, the counter model differentially reallocates the assignment of noise features, whereas our model includes a differential number of noise features in the decision process. It is not imme-

diately obvious to us how to carry out an empirical test based on this distinction, but the models are conceptually quite different, so continued development of both models may eventually lead to empirical tests.

The results of our modeling efforts support the view that in REMI, context acts as a prior. The influence of a prior weakens as more diagnostic evidence is gathered. Likewise, for REMI the influence of context (and prior study) weakens as more veridical diagnostic features are made available by either increasing the flash duration or increasing the dissimilarity of the alternatives. These are characteristics that, Ratcliff and McKoon (1997) argued, a model of implicit memory must display. Thus, a model of perceptual identification built along Bayesian principles exhibits the properties we sought and demonstrates that Ratcliff and McKoon's (1997) data can be handled by a model in which "prior exposure to a word changes some property of the representation of the word" (Ratcliff & McKoon, 1997, p. 339). In building this model, we have maintained the viability of REMI and other models (Anderson & Lebierre, 1998; Bower, 1996; Reder & Gordon, 1996) that depend on strengthening memories to account for implicit memory effects.

## References


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

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38,* 341–380.

Anderson, J. R., & Lebiere, C. (Eds.). (1998). *Atomic components of thought.* Mahwah, NJ: Erlbaum.

Anderson, J. R., & Matessa, M. (1992). Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning, 9,* 275–308.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408.

Bower, G. H. (1996). Reactivating a reactivation theory of implicit memory. *Consciousness and Cognition, 5,* 27–72.

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

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62,* 145–154.

Feustel, T. C., Shiffrin, R. M., & Salasoo, A. (1983). Episodic and lexical contributions to the repetition effect in word identification. *Journal of Experimental Psychology: General, 112,* 309–347.

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.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Hintzman, D. L. (1990). Human learning and memory: Connections and dissociations. *Annual Review of Psychology, 41,* 109–139.

Ijiri, Y., & Simon, H. A. (1977). *Skew distributions and the sizes of business firms.* Amsterdam: North-Holland.

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

Masson, M. E. J., & MacLeod, C. M. (1996). Contributions of processing fluency to repetition effects in masked word identification. *Canadian Journal of Experimental Psychology, 50,* 9–21.

McClelland, J. L., & Chappell, M. (1997). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 734–760.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88,* 375–407.

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review, 39,* 352–370.

McKoon, G., & Ratcliff, R. (in press). The counter model for word identification: Reply to Bowers (1999). *Psychological Review.*

Mensink, G. J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review, 95,* 434–455.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165–178.

Morton, J. (1970). A functional model for memory. In D. A. Norman (Ed.), *Models of human memory* (pp. 202–254). New York: Academic Press.

Murdock, B. B., Jr. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100,* 183–203.

Neaderhiser, B. J., & Church, B. A. (1998, November). *Can perceptual bias account for priming in forced choice tasks?* Paper presented at the 39th Annual Meeting of the Psychonomic Society, Dallas, TX.

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

Nobel, P. A., Diller, D. E., & Shiffrin, R. M. (in press-b). A model for accuracy and response time in cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Raaijmakers, J. G. W., Schooler, L. J., & Shiffrin, R. M. (1997, November). *Effects of repeated study on word identification in a 2AFC task.* Paper presented at the 38th Annual Meeting of the Psychonomic Society, Philadelphia.

Rabinowitz, F. M. (1995). A stochastic algorithm for global optimization with constraints. *ACM Transactions on Mathematical Software, 21,* 194–213.

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

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785.

Ratcliff, R., McKoon, G., & Verwoerd, M. (1989). A bias interpretation of facilitation in perceptual identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 378–387.

Reder, L. M., & Gordon, J. S. (1996). Subliminal perception: Nothing special, cognitively speaking. In J. Cohen & J. Schooler (Eds.), *Cognitive and neuropsychological approaches to the study of consciousness* (pp. 125–134). Mahwah, NJ: Erlbaum.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81,* 274–280.

Salasoo, A., Shiffrin, R. M., & Feustel, T. C. (1985). Building permanent memory codes: Codification and repetition effects in word identification. *Journal of Experimental Psychology: General, 114,* 50–77.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology, 32,* 219–250.

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

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

Wagenmakers, E. M., 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.

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology, 1,* 59–85.

# Appendix A

## Forced Choice

Let the $g$ appropriate for the target be denoted $g_t$ and the $g$ appropriate for the foil be denoted $g_f$. Let the $g$ used to fill in the values of the perceived vector be denoted $g$; let subscripts $i$, $j$, and $k$ refer to the feature values that each of these might attain. Let **t** refer to the target choice vector, **f** to the foil choice vector, and **d** to the perceived vector. For any given diagnostic feature in these vectors (i.e., for which $i$ and $j$ are not equal), the following equation gives the distribution of the feature values that might occur:

$$P_{ijk} = P(\mathbf{t} = i, \mathbf{f} = j, \mathbf{d} = k | i \neq j)$$

$$= \begin{cases} g_{ti}g_{fj}[s + (1 - s)g_i]/P(i \neq j) & \text{if } i = k \\ g_{ti}g_{fj}(1 - s)g_k/P(i \neq j) & \text{if } i \neq k. \end{cases} \quad \text{(A1)}$$

In Equation A1, the term $P(i \neq j)$ is calculated by summing the terms in the numerators of the expressions for which $i$ and $j$ are not equal. Let $P_t$, $P_f$, and $P_n$ refer to the probabilities that a given diagnostic feature might match the target, the foil, or neither, respectively:

$$P_t = \sum_{\substack{j=1 \\ j \neq i}}^{\infty} \sum_{i=1}^{\infty} P_{iji}; \quad P_f = \sum_{\substack{j=1 \\ j \neq i}}^{\infty} \sum_{i=1}^{\infty} P_{ijj}; \quad P_n = 1 - P_t - P_f.$$

Let $N$ be the sum of $n1$, $n2$, and $n3$, where these refer respectively to the number of diagnostic features that match target, foil, and neither. Then:

$$P(n1, n2, n3) = \binom{N}{n1, n2, n3} P_t^{n1} P_f^{n2} P_n^{n3};$$

$$P(n1 = n2 + 1) = \sum_{n1 = n2 + 1} P(n1, n2, n3);$$

$$P(n1 = n2 - 1) = \sum_{n1 = n2 - 1} P(n1, n2, n3);$$

$$P_N(C) = P(n1 > n2) + .5P(n1 = n2), \quad \text{if neither studied;}$$

$$P_T(C) = P_N(C) + .5\alpha[P(n1 = n2) + P(n1 = n2 - 1)],$$

if target studied; and

$$P_F(C) = P_N(C) - .5\alpha[P(n1 = n2) + P(n1 = n2 + 1)],$$

if foil studied.

# Appendix B

## Yes–No

Let $P_{M=}$ equal the probability of a match between $d$ and the test word when the target and test words match in that feature position. Let $P_{M\neq}$ equal the probability of a match in a feature position when the target and test words mismatch in that feature position. Let $g_{ti}$ and $g_{ei}$ be the respective probabilities that the target and the test word have a feature with value $i$ in a given position (depending on the experimental design, either of these could be the basis for the feature values obtained when the target and foil test word match in a given position; usually similar foils are matched to targets, so $g_{ti}$ is used in the expressions below). Let $g_i$ be the probability that the value in the perceived vector has value $i$ if stored there by noise. Then:

$$P_{M=} = s + (1 - s) \sum_{i=1}^{\infty} g_{ti}g_i; \quad P_{M\neq} = (1 - s) \sum_{i=1}^{\infty} g_{ti}g_i.$$

Let $P_{ij}$ = probability that $i$ matches occur for the $M$ feature positions in which the target and choice word match, and $j$ matches occur for the $M^*$ feature positions in which the target word and choice word mismatch.

$$P_{ij} = \binom{M}{i} P_{M=}^{i}(1 - P_{M=})^{M-i} \binom{M^*}{j} P_{M\neq}^{j}(1 - P_{M\neq})^{M^*-j}.$$

Let $C$ be the criterion number of matches for responding yes:

$$P(\text{yes}) = \sum_{i+j>C} P_{ij}. \quad \text{(B1)}$$

The numbers $M$ and $M^*$ ($M + M^* = L$) are parameters determined by the experimental condition. For target tests, $M = L$. For similar foils, $M^* = L_s$. For dissimilar foils, $M^* = L_d$. The features common to all words ($L - L_d$ in number) could in principle have a $g$ value different from the other $g$ values, but we have for convenience set it to the common default value for noise (denoted $g$ above). If the test item had been studied, then the probability of responding yes is

$$P(\text{"yes"}) = (1 - \alpha) \sum_{i+j>C} P_{ij} + \alpha \sum_{i+j>C-1} P_{ij}, \quad \text{(B2)}$$

where $\alpha$ is the probability of an extra context match due to prior study. These equations were verified by simulations.

*(Appendixes continue)*

# Appendix C

## Estimating Orthographic Similarity

To assess the validity of using Zipf's law, we examined the orthographic similarity of printed words to each other. Restricting our analysis to four-letter words with frequencies of at least one in a million in Kučera and Francis (1967), because Ratcliff and McKoon (1997) culled their words from this corpus, we calculated a rough measure of visual similarity for each pair of words. Consistent with Ratcliff and McKoon's (1997) description of orthographic similarity, letters were coded according to whether they have ascenders (e.g., *l*, *d*, *h*), descenders (e.g., *j*, *y*, *q*), or neither (e.g., *i*, *n*, *s*). A measure of word dissimilarity was constructed by matching corresponding letters. Differently shaped, same shaped, and identical letters added respectively 5, 2, and 0 units to the measure. For example, the distance between *lied* and *died* on this scale would be 2 and the distance between *lied* and *sofa* would be 17. For each of the 1,500 words that met the inclusion restrictions, the similarity distances of the other words were calculated and were ranked from most similar to least similar. The average similarity distance was calculated for the most similar word (i.e., the nearest neighbor), the second most similar (the second nearest neighbor), and so on. Figure C1 (Panel A) plots these similarity distances as a function of rank, and Figure C1 (Panel B) plots the results in log–log coordinates. The straight line observed in Figure C1 (Panel B) ($r^2$ = .99) supports the assumption that Zipf's law approximates the distribution of similarities between words, with the values of $r$ and $b$ estimated to be 2.5 and .253, respectively.

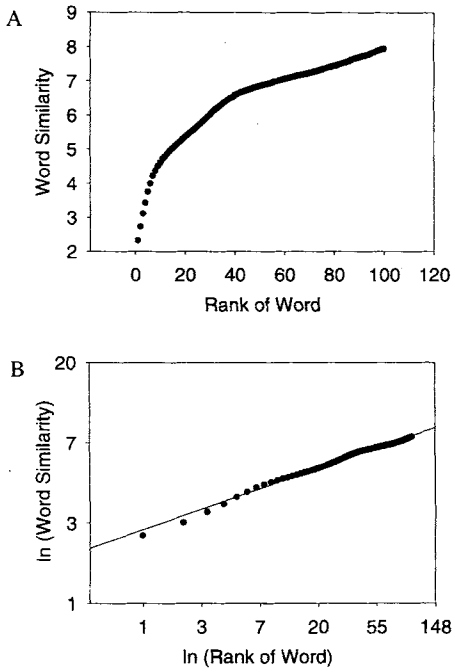


*Figure C1.* A measure of visual similarity was constructed for four-letter words (lower values indicating greater similarity). All the neighbors of a word are ranked from 1 (*most similar*) to 20 (*least similar*) on this measure. Panel A gives the average across all words of the visual similarity of a word's first-ranked neighbor, second-ranked neighbor, etc. Panel B shows the data from Panel A in log–log coordinates. The straight line in this scale indicates that Zipf's law approximates the distribution of a word's neighbors in terms of their similarity.

Received August 12, 1997
Revision received June 5, 2000
Accepted June 14, 2000 ■