Running head: A Model for Evidence Accumulation

A Model for Evidence Accumulation in the Lexical Decision Task

Eric-Jan Wagenmakers[1], Mark Steyvers[2], Jeroen G.W. Raaijmakers[1], Richard M. Shiffrin[3], Hedderik van Rijn[1], and René Zeelenberg[1]

1 University of Amsterdam

2 Stanford University

3 Indiana University

Send correspondence to:

Eric-Jan M. Wagenmakers

Department of Psychology

Northwestern University

2029 Sheridan Road, 102

Swift Hall, Evanston, IL 60208

USA

e-mail: ej@northwestern.edu

phone: (847) 467 3141

Abstract

We present a new model for lexical decision, REM-LD, that is based on REM theory (e.g., Shiffrin & Steyvers, 1997). REM-LD uses a principled (i.e., Bayes' rule) decision process that simultaneously considers the diagnosticity of the evidence for the 'WORD' response and the 'NONWORD' response. The model calculates the odds ratio that the presented stimulus is a word or a nonword by accumulating likelihood ratios for each lexical entry in a small neighborhood of similar words. We report two experiments that used the signal-to-respond paradigm to obtain information about the time course of lexical processing. Experiment 1 verified the prediction of the model that the frequency of the word stimuli affects performance for nonword stimuli. Experiment 2 was done to study the effects of nonword lexicality, word frequency, and repetition priming and to demonstrate how REM-LD can account for the observed results. We discuss how REM-LD can be extended to account for effects of phonology such as the pseudohomophone effect, and how REM-LD can predict response times in the popular 'respond-when-ready' paradigm. Several other quantitative models of lexical decision are evaluated with respect to the findings reported here.

A Model for Evidence Accumulation in the Lexical Decision Task

In this paper, we propose a new model for lexical decision, REM-LD (standing for retrieving effectively from memory – lexical decision). The REM-LD model is a global familiarity model based on Bayesian principles similar to those used in the recently developed REM models for recognition memory (Diller, Nobel, & Shiffrin, 2001; Nobel & Shiffrin, 2001; Shiffrin & Steyvers, 1997; see also McClelland & Chappell, 1998), recall (e.g., Diller et al., 2001; Malmberg & Shiffrin, in press; Nobel & Shiffrin, 2001), long-term priming in perceptual identification (Schooler, Shiffrin, & Raaijmakers, 2001), and short-term priming in perceptual identification (Huber, Shiffrin, Lyle, & Ruys, 2001). The REM models constitute a general framework that describes how information is stored and retrieved from memory, and how an optimal decision can be made based on noisy information. The concept of optimal decision making provides a principled basis for modeling the functioning of human memory (cf. ACT-R, Anderson & Lebiere, 1998). We aim to show how the REM principles can be applied in a straightforward fashion to describe performance in a lexical decision task.

The outline of the article is as follows. First we will briefly describe the lexical decision task and the signal-to-respond paradigm that is used throughout this article. We then outline the general characteristics of the REM models. Next, we discuss the REM model as applied to lexical decision in more detail, presenting several simulations and a study that conforms a prediction of the model. A second study is used to demonstrate how REM-LD can account for the combined effects of processing time, word frequency, repetition priming, and nonword lexicality. Subsequently we will discuss how the REM-LD model can be extended to account for the pseudohomophone effect and how the model can generate response latencies when it is to be applied to the traditional 'respond-when-ready' paradigm. Finally, the REM-LD model is compared to several current models for lexical decision with

respect to the findings and simulations reported here. We argue that the most fashionable and most complete quantitative models of lexical decision to date cannot, in their current form, handle data from the signal-to-respond paradigm. Moreover, we believe the principled Bayesian decision mechanism inherent in the REM-LD model provides a parsimonious and attractive alternative to the often-used temporal deadline mechanism.

The Lexical Decision Task

It is generally assumed that the understanding of the skill of reading should be based in part on an understanding of the storage and retrieval of words. These processes are often studied through the use of the lexical decision task, requiring participants to distinguish words (e.g., CHAIR, FUME) from nonwords (e.g., GREACH, ANSU). Over the last decades, research in lexical decision has produced an enormous amount of data[1] and various empirical regularities have been established.

Among the myriad of findings available in the literature on lexical decision, we decided to select as targets for modeling by REM-LD three of the most robust and most general phenomena. In the traditional 'respond-when-ready' paradigm, when accuracy is usually near ceiling, these three important phenomena, as seen in the choice response latencies are: (1) the nonword lexicality effect (e.g., James, 1975; Joordens, Piercey, & Mohammad, 2000; Stone & Van Orden, 1989, 1993) -- nonwords that look like words (i.e., pseudowords such as GREACH) take longer to be classified correctly than nonwords such as EAGRCH that are relatively dissimilar to words; (2) the word frequency effect (e.g., Balota & Chumbley, 1984; Scarborough, Cortese, & Scarborough, 1977) -- words that occur relatively often in natural language (high frequency or HF words such as CHAIR) are classified correctly faster than words that occur relatively rarely (low frequency or LF words such as FUME); and (3) the repetition priming effect (e.g., Logan, 1988, 1990; Scarborough et al., 1977) -- the prior presentation of a word in an experiment leads to faster correct

classifications for the same word on its second presentation (this increase in performance is particularly pronounced for LF words; e.g., FUME benefits more from prior exposure than CHAIR -- see for instance Scarborough, Gerard, & Cortese, 1984).

Several models of visual word recognition have been proposed in order to give a theoretical account of the empirical effects revealed by the lexical decision task (for a review see Jacobs & Grainger, 1994). Most of the current models for lexical decision share a number of basic assumptions, and hence can be characterized in the following, very general way. The presented stimulus (i.e., a letter string) initially activates the word representations in memory that are orthographically and/or phonologically similar to the presented stimulus. In case the stimulus is a word, the positive evidence increases over time. Subsequently, a 'WORD' response is given when the positive evidence (e.g., the increase in activation due to the presentation of the stimulus) exceeds a criterion value. In many models for lexical decision, the 'NONWORD' response is a default response, because it is brought about by the absence or lack of positive information. In this simplified view, lexical decision is equivalent to lexical activation. We will discuss several current lexical decision models in some more detail later.

The activation-style models mentioned above have a decision mechanism that is very different from the one inherent to REM-LD. In REM-LD, a response is based on the balance between the positive evidence supporting a 'WORD' response and the negative evidence supporting a 'NONWORD' response. We aim to show that the REM-LD model provides a principled and unified account of lexical decision performance. One of the additional goals of the present approach toward modeling lexical decision is to provide an explicit account of how performance increases with processing time (i.e., the time-course of lexical processing), rather than to focus solely on the end result of the processes involved. To address this issue, we used a signal-to-respond procedure (Antos, 1979; Hintzman & Curran, 1997), forcing

participants to respond at specific times. The dependent measure of interest is the probability of correct classification at various times after stimulus onset. This procedure provides more data than the traditional lexical decision task in which instructions are given to 'respond as fast and accurately as possible' (i.e., the respond-when-ready paradigm). In addition, the increase of correct classification with processing time can constrain theories for the time course of lexical processing.

Defining Characteristics of the REM Models

The basic assumptions of REM can be conveniently classified with respect to the following three stages that jointly determine memory performance: (1) the storage of information in memory; (2) the retrieval of information from memory; and (3) the decision process.

With respect to storage and representation of information in memory, REM assumes that memory traces of higher-order units such as words consist of a number of lower-level elements or features (cf. Estes, 1950). Features can encode various types of information that are convenient to classify into two types: properties of the higher-order representation itself (i.e., content or item information including semantic, phonological, and orthographic information), or contextual information (i.e., properties that correspond to the "physical, spatial/temporal, environmental, physiological, and/or emotional states in which the item was experienced", Malmberg & Shiffrin, in press, p. 6). In the work presented here, the distinction between content and context information is not of central importance. In REM, memory traces are subdivided into episodic traces and lexical/semantic traces. Episodic traces contain incomplete and error-prone information about one specific encounter with the corresponding stimulus. In contrast, lexical/semantic traces reflect the accumulation of part of the information from each of the previous encounters with the corresponding stimulus, eventually producing a relatively complete and accurate trace (at least for the commonly

occurring features of the encoded stimulus). Therefore, the presentation of a known stimulus such as a word will have two effects: (1) the formation of a new episodic trace composed of relatively few features that encode error-prone information both about the item and about the context in which the item was presented; and (2) the addition and/or updating of information in the lexical/semantic memory trace that is not already stored. Since content or item-information is already stored almost perfectly, not much new item-information such as meaning will be added to the lexical/semantic trace. However, novel information such as the current context and any unique font for the current presentation can be added to the lexical/semantic trace.

For some memory tasks such as recall and recognition (e.g., Shiffrin & Steyvers, 1997), it is vital that the subject uses the experimental context to filter items recently presented on a study list (i.e., target items) from items that were not presented on a study list (i.e., foils). In these context-dependent tasks, performance will rely to a large extent on the quality and quantity of the stored episodic memory traces. For other memory tasks such as perceptual identification (Huber et al., 2001; Schooler et al., 2001) or lexical decision, performance does not usually depend on one specific past encounter with the presented stimulus. For the time being we will make the simplifying assumption that lexical decision involves only the lexical/semantic traces, and not the very weak and context-dependent episodic traces. Other possibilities will be taken up in the Discussion following Experiment 2.

With respect to the retrieval of information from memory, REM assumes that a memory probe (e.g., the stimulus combined with current context in lexical decision, or only context for the first retrieval attempt in a free recall task) is matched simultaneously to traces in memory. The matching process is based on a feature-by-feature comparison between the probe and each memory trace. Both the probe and the traces contain a complete set of

features, although not all of these become available instantly. This comparison process results in a number of matching features and a number of mismatching features for each separate probe-to-trace comparison. In Shiffrin and Steyvers (1997), feature values had different probabilities, corresponding to base rate differences, so that the value of a matching feature determines the likelihood of that match. For simplicity we assume in this article that feature values are equiprobable, so the only relevant information is whether features match or mismatch.

A simplified example of the feature-comparison process is given in Table 1 (for comparison see the episodic version given in Shiffrin & Steyvers, 1997, Figure 1). The probe is represented as a set (i.e., a vector) of features. Suppose a feature can take on any integer value from one to five, with equal probability. In the example given in Table 1, a probe is matched against two traces in memory. The four features representing the probe are compared to the corresponding features in the two memory traces. For Trace 1 in Table 1, only the third feature has the same value as the third feature from the probe. Hence, the feature comparison process results in one match, and three mismatches. As can be seen in Table 1, the probe is very similar to Trace 2, and the comparison process results in three matches and only one mismatch. One might think that for a trace that actually represents the probe, all feature comparisons would be matches, but that is too strict, and we allow for some discrepancies to arise even in such a case. The task for the system at any point in time is to make an optimal decision (i.e., 'WORD' or 'NONWORD') based on the observed number of matches and mismatches that result from the feature comparison process between the probe and each of the memory traces. The basic theme of the REM approach is the implementation of this idea of optimal or near-optimal decision making in the face of noisy information (an idea that also underlies the rational approach of ACT-R; e.g., Anderson & Lebiere, 1998). The idea can be illustrated by continuing our example from Table 1. Assume the system has

compared the probe to each memory trace, and obtained a count of matching and mismatching features. In order to make an optimal decision, the system needs to estimate two probabilities: (1) the probability that a probe feature will match a trace feature, given that the probe corresponds to the memory trace (i.e., P(match | same)), and (2) the probability that a probe feature will match a trace feature, given that the probe does not correspond to the memory trace (i.e., P(match | different)). These two probabilities determine the diagnosticity of a feature match, and the diagnosticity of a feature mismatch. In the example from Table 1, assume that the system estimates P(match | same) to be .8 (so P(mismatch | same) = .2), and P(match | different) to be .4 (so P(mismatch | different) = .6). Thus, the probability of a feature match is twice as likely (.8/.4), and of a feature mismatch is one third as likely (.2/.6), when the probe is compared to its corresponding memory representation than when it is not. An optimal solution multiplies these ratios of 2 (for matches) and 1/3 (for mismatches) for all features in a trace. In our example, Trace 1 has only one matching and three mismatching features giving a trace likelihood ratio of $2\times\left(\frac{1}{3}\right)^3=\frac{2}{27}$. Thus the likelihood ratio that the probe corresponds to Trace 1 is not very high. In contrast, the likelihood ratio of the probe corresponding to Trace 2 is much higher: $\frac{1}{3}\times 2^3=\frac{8}{3}$. As shown in Shiffrin and Steyvers (1997), the odds ratio of the probe corresponding to one of the memory traces is given by the average of the likelihood ratios. Thus, the odd ratio is $\frac{\frac{2}{27}+\frac{8}{3}}{2}=\frac{37}{27}\approx 1.37$. Since the optimal response criterion is set at an odds ratio of 1, a Bayesian system will assume that the probe indeed corresponds to one of the memory traces. If the calculated odds ratio were between 0 and 1, the system would have assumed the opposite. The next section will give a mathematical justification of these calculations.

The REM-LD Model

This section describes the assumptions of the REM model as applied to lexical decision in more detail, and subsequently gives a mathematical analysis of a Bayesian lexical decision process. First, with respect to storage of information in memory, both the probe and all the memory traces consist of $\underline{k} = 30$ features. Each of these features can take on equiprobable integer values, the range being immaterial for the current purposes. The assumptions regarding storage of information in memory are similar to those used in other applications of the REM model and remain the same throughout the work reported here.

Second, we assume that a probe is compared to the $\underline{n} = 10$ lexical/semantic traces in memory that are most similar to the probe orthographically. If the probe is a word, it will correspond to one of these ten lexical/semantic traces. If the probe is a nonword, it will correspond to none of these lexical/semantic traces. The limitation to ten traces was made for computational convenience and simplicity.

The similarity between the probe and a lexical/semantic trace is indexed by the probability $\underline{\beta}$ that a given feature value in the probe matches the corresponding feature value in the lexical/semantic trace. For the purposes of simulation, for a given probe vector, we construct the ten most similar traces as follows. Let the probe-to-trace similarity for corresponding (i.e., same) and non-corresponding (i.e., different) representations be indexed by β_1 and β_2, respectively. Then, $0 < \beta_2 < \beta_1 < 1$. The fact that $\beta_1 < 1$ means that there will always be a certain degree of dissimilarity between a presented word probe and the corresponding lexical/semantic trace: Not all features from the probe will match the features from the corresponding semantic/lexical trace even if the comparison process were faultless. This discrepancy can be due to various factors such as encoding variability, fallible perception or mismatching contextual information. The fact that $\beta_1 > \beta_2$ means that the probe-to-trace similarity is greater for corresponding representations than for non-

corresponding representations. Finally, the fact that $\beta_2 > 0$ means that even if the lexical/semantic trace does not correspond to the probe, they can still have several features in common. With probabilities $1-\beta_1$ and $1-\beta_2$ the trace features are dissimilar to the probe features. Thus for the case when the probe and trace encode the same item (denoted by s), the probability of a feature match is:

$$P(match \mid s) = \beta_1 . \quad (1)$$

For the case when probe and trace encode different items (denoted by d), we write $P(match \mid d) = \beta_2$.

Throughout this paper, we generate predictions from the REM-LD model using the values of β_1 and β_2 that were in fact used to generate the traces (i.e., the true values). When one assumes the process by which β_1 and β_2 are estimated is noisy, the resulting variability around the true values will tend to decrease overall performance (an effect that may be offset by increasing the difference between the average values of β_1 and β_2). The issue of how the system estimates values of β_1 and β_2 is not the topic of interest here, but such estimation can be based on the information (i.e., the number of observed matches and mismatches) obtained on previous trials. Depending on the total amount of information, such an estimation process could operate quite accurately. Thus, for current purposes we use the true values of β_1 and β_2 throughout. To acknowledge the fact that the system (or, more accurately, the participant) has to estimate these values we will henceforth denote the values of β_1 and β_2 used in the decision process as $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

As mentioned above, one of our aims is to provide an explicit account of the time-course of lexical processing as revealed by the signal-to-respond paradigm. In order to model the increase in performance with processing time, we assume that it takes a variable amount of time to activate different probe features and compare them to trace features. For

simplicity, the time course of activation of probe features and comparison to trace features are combined into a single activation process: The probability of activation of a probe feature, α, increases monotonically over time according to

$$\alpha(t) = \begin{cases} 1 - \exp[-b(t - t_0)], & t \geq t_0 \\ 0 \quad , & t < t_0 \end{cases}, \tag{2}$$

where t equals processing time, b represents the rate of increase in α with t, and t_0 represents the starting point of the function, that is, the minimum processing time for correctly activating probe features. The specific form of Equation 2 is motivated primarily by simplicity; because this is not a focus of the present project we used fixed values of t_0 = 273 and b = .0025 for all simulations (these values were chosen after a cursory examination of the parameter space).

The probability that exactly r probe features will be active at time t since stimulus onset (out of a total of k = 30 features) is given by a binomial distribution:

$$P(R = r) = \binom{k}{r} \alpha(t)^r \left[1 - \alpha(t)\right]^{k-r}. \tag{3}$$

Equations 2 and 3 describe how, as processing time increases, more and more probe features are activated and become available to be compared to the traces. In other words, the amount of information that is available to the comparison process increases with processing time. These equations determine the distribution of the number of probe features involved in comparison at any given time, t. Matches and mismatches with any trace only occur for those features that are presently active.

Given r features are active at time t, the probability of observing exactly m matches and r-m mismatches in comparison with a trace depends on whether the trace encodes the same item as the probe. For the same encoding case, we have:

$$P(M = m \mid s) = \binom{r}{m} P(match \mid s)^m \left[1 - P(match \mid s)\right]^{r-m}. \tag{4}$$

The probability P(M=m | d) of observing m matches given that the probe does not correspond to the lexical/semantic trace can be obtained by replacing P(match | s) in Equation 4 by P(match | d). The likelihood ratio λ of the probe corresponding to a lexical/semantic trace, given that m matches were observed, is given by multiplying the ratios for each feature:

$$\lambda = \frac{P(D \mid s)}{P(D \mid d)} = \left[\frac{P(match \mid s)}{P(match \mid d)}\right]^{m} \left[\frac{1 - P(match \mid s)}{1 - P(match \mid d)}\right]^{r-m} \tag{5}$$

(Equation 5 is a special case of Equation 3 in Shiffrin and Steyvers, 1997). Putting Equation 5 and Equation 1 together, and adding a subscript j to refer to the j-th trace, we obtain:

$$\lambda_j = \left[\frac{\hat{\beta}_1}{\hat{\beta}_2}\right]^{m_j} \left[\frac{1-\hat{\beta}_1}{1-\hat{\beta}_2}\right]^{r-m_j} . \tag{6}$$

The number of matching and mismatching features (i.e. the exponents in this equation) have a distribution determined by Equations 2, 3 and 4.

Finally, we assume that the system makes an optimal decision given by Bayes' rule. According to Bayes' rule (e.g., Hoel, Port, & Stone, 1971), the posterior odds ratio Φ that the probe is a word can be obtained by multiplying the likelihood ratio and the prior odds ratio:

$$\Phi = \frac{P(W \mid D)}{P(NW \mid D)} = \frac{P(D \mid W)}{P(D \mid NW)} \frac{P(W)}{P(NW)}, \tag{7}$$

where P(W | D) and P(NW | D) indicate the probability that given the observed data (i.e., the number of matching and mismatching features), the probe is a word or a nonword, respectively. An unbiased Bayesian system will respond 'WORD' when $\Phi > 1$, respond 'NONWORD' when $\Phi < 1$, and guess when $\Phi = 1$. When the probe is equally likely to be a word or a nonword, as is usually the case in lexical decision experiments, the prior odds ratio is one and the posterior odds ratio is determined by the first ratio from the right side of Equation 7.

If the probe is a word, then exactly one of the activated traces corresponds to (i.e., matches) the probe. If the probe is a nonword, then none of the activated traces corresponds to the probe. Given the former case, the probability that a given trace matches is just 1/n (1/10 if we assume 10 traces in the comparison set); a simple derivation (Shiffrin & Steyvers, 1997, Appendix A) then shows that the posterior odds ratio Φ is the average of n likelihood ratios:

$$\Phi = \frac{P(D \mid W)}{P(D \mid NW)} = \frac{1}{n}\sum_{j=1}^{n}\frac{P(D_j \mid s_j)}{P(D_j \mid d_j)} = \frac{1}{n}\sum_{j=1}^{n}\lambda_j \,, \tag{8}$$

where $P(D_j \mid s_j)$ and $P(D_j \mid d_j)$ denote the probability of observing the data (i.e., the number of matching and mismatching features resulting from a comparison between the activated probe features and the features of the memory trace) given that the probe corresponds to the jth memory trace, and given that the probe does not correspond to the jth memory trace, respectively. In short, REM-LD bases its 'WORD' vs. 'NONWORD' decision on the posterior odds ratio that the probe corresponds to exactly one of the n lexical/semantic traces. This is equivalent to averaging the n separate likelihood ratios λ_j that the probe corresponds to lexical/semantic trace j.

Predictions and General Implications of the REM-LD Model

The most straightforward predictions of REM-LD follow from the fact that the system simultaneously evaluates the diagnosticity of the evidence supporting a 'WORD' response (i.e., P(W | D)) and the evidence supporting a 'NONWORD' response (i.e., P(NW | D)). A crucial aspect of REM-LD is that the 'NONWORD' response is not just a default response. Rather, the 'WORD' and 'NONWORD' responses are two sides of the same coin. This observation follows naturally from a Bayesian analysis of the lexical decision task, such as provided by the REM-LD model. We will illustrate this notion with two well-documented

phenomena in lexical decision: (1) the effect of nonword lexicality, and (2) the effect of word frequency.

Several researchers (e.g., James, 1975; Joordens, Piercey, & Mohammad, 2000; Stone & Van Orden, 1989, 1993) have shown that performance for nonwords that are very similar to words (i.e., pseudowords such as GREACH) is worse than performance for nonwords that are less similar to words (e.g., EAGRCH). Moreover, the similarity of the nonwords to words also affects performance for the word stimuli: Performance for word stimuli that have to be distinguished from word-like nonwords is worse than for words that have to be distinguished from less word-like nonwords. We will demonstrate by simulation that REM-LD predicts these results. For all simulations reported in this paper, each data point reflects the average of 10,000 binary decisions.

In REM-LD the similarity of the nonwords to the lexical/semantic traces in memory is quantified by the parameter β_2 (i.e., the probability of a matching feature given that the probe does not correspond to the lexical/semantic trace). In other words, the similarities of the ten most similar lexical images to the nonword test string will all tend to be lower the less word-like is the test string. Throughout this article, we make the simplifying assumption that the similarity between a word probe and a non-corresponding lexical/semantic trace is the same as the similarity between a word-like nonword probe and any of the lexical/semantic traces.

In this article we simulate the signal-to-respond procedure used in Experiments 1 and 2. The participant has to respond immediately after hearing a tone, and the dependent variable of interest is the probability of responding 'WORD' as a function of processing time (i.e., time after stimulus onset). In almost all of the simulations and the experiments reported here, the tone (i.e., the signal-to-respond) could be presented at one of six times after stimulus onset (i.e., deadlines): 75, 200, 250, 300, 350, and 1000 ms. In accordance with the

empirical results we let the model 'respond' after adding 200 ms to the deadlines. Figure 1a shows the behavior of the REM-LD model with the following parameter values: $\hat{\beta}_1$ = .82, $\hat{\beta}_2$ (word-like nonwords or pseudowords) = .46, and $\hat{\beta}_2$ (less word-like nonwords) = .37. It is assumed that two separate paradigms are modeled: One in which words are paired with nonwords (open symbols), and another in which words are paired with pseudowords (closed symbols). If it were assumed instead that the kinds of foils were mixed, then perhaps the model/system would choose an estimate of $\hat{\beta}_2$ somewhere between .46 and .37; in this case the curves for pseudowords and less word-like nonwords would still separate because of the differing number of matches, but the two word curves would not differ from each other.

The results show a number of effects that match those found in the literature: (1) performance is at chance accuracy at the shortest deadline, and asymptotes to near-perfect performance at the longest deadline (cf. Hintzman & Curran, 1997), (2) performance for word-like nonwords (i.e., pseudowords) is worse than for less word-like nonwords (Grainger & Jacobs, 1996), and (3) performance is worse for words that have to be distinguished from word-like nonwords than for words that have to be distinguished from less word-like nonwords (Grainger & Jacobs, 1996, Figure 25).

Another well-documented finding in lexical decision is the effect of word frequency: Performance for high-frequency or HF words is better than performance for low-frequency or LF words (e.g., Scarborough et al., 1977). In REM-LD we assume that the probability that a feature of a word probe matches the corresponding feature in its own lexical/semantic trace, $\hat{\beta}_1$, is higher for HF word probes than for LF word probes.[2] An increased matching probability for HF words over LF words may arise as a result of various mechanisms, for example: (1) HF word traces may match more readily with the experimental context. This could be due to the fact that HF words (e.g., CHAIR) generally occur in many different

contexts, whereas LF words (e.g., PYRAMID, PHARAOH) are often tied to relatively few contexts (e.g., Dennis & Humphreys, 2001; Landauer & Dumais, 1997).[3] (2) More accurate content-information (i.e., semantic, orthographic or phonological properties, such as spelling) might be stored in an HF trace than in an LF trace. To our knowledge, present empirical evidence does not allow a choice to be made from the various alternatives.

A second simulation was carried out to study whether REM-LD could produce the following effects: (1) the word frequency effect, and (2) the finding that the word frequency effect is attenuated when the nonwords are not very word-like. Again two paradigms are modeled, one in which the high and low frequency words are mixed with nonwords, and one in which the high and low frequency words are mixed with word-like nonwords (i.e., pseudowords). The results can be seen in Figure 1b. The parameter values are: $\hat{\beta}_1$ (HF words) = .865, $\hat{\beta}_1$ (LF words) = .775, $\hat{\beta}_2$ (pseudowords) = .505, and $\hat{\beta}_2$ (nonwords) = .415. Because words of different frequency are mixed in the simulated paradigm, the estimated value for the overall similarity of a word probe to its corresponding memory trace was set at the average of the β-values for HF and LF words. That is the actual values of β used to generate probe and trace vectors were .865 and .775, but the equations used to calculate likelihood ratios used a common value of (.865+.775)/2 for both kinds of words. The two important results illustrated in Figure 1b are: (1) performance for HF words (circle symbols) is better than performance for LF words (triangle symbols) (i.e., the word frequency effect), and (2) the word frequency effect is larger when the nonwords are very word-like (i.e., pseudowords, filled symbols) then when they are not (open symbols).

Up to this point we have illustrated the behavior of the model by showing how it accounts for the finding that nonword characteristics affect performance for the word stimuli. The mirror image of this result, namely that word characteristics affect performance for the nonword stimuli, has also been occasionally reported (e.g., Joordens et al., 2000; Stone &

Van Orden, 1993). More specifically, the aforementioned studies showed that the frequency of the word stimuli affects performance for the nonword stimuli: When all word stimuli are of high frequency, classification performance for nonword stimuli is facilitated relative to when all word stimuli are of low frequency. From a Bayesian perspective (cf. Equations 4 and 6) this result is to be expected, since lexical decision performance depends on the discriminability of the words and the nonwords. We begin by presenting an experiment carried out to test this prediction of the REM-LD model using the signal-to-respond paradigm.

Experiment 1

Method

Participants. Thirty-five students of the University of Amsterdam participated for course credit. All participants were native speakers of Dutch and reported normal or corrected-to-normal vision.

Stimulus Materials. We used three types of experimental stimuli: (1) 144 HF Dutch words, each occurring more than 25 times per million according to the CELEX lexical database (Baayen, Piepenbrock, & Van Rijn, 1993), (2) 144 LF Dutch words, each occurring one to five times per million, and (3) 288 pronounceable nonwords created by replacing one letter of an existing word (e.g., GREACH derived from PREACH). Specifically, the nonwords were created by replacing one letter from a word that was not used in the experiment. The letter position subject to replacement was determined randomly. A vowel was always replaced by a vowel, and a consonant was always replaced by a consonant. The replacement letters were sampled in proportion to the letter frequencies (e.g., the rare letter ‘z’ was unlikely to be used as a replacement, whereas the common letter ‘r’ was relatively likely to be used as a replacement). We verified that the letter string that resulted from the replacement operation was not another word.

The three stimulus categories were matched on neighborhood structure (a neighbor is a word differing from another word in one letter, so TIED is a neighbor of LIED): These categories had roughly the same summed logarithmic word frequency of the neighbors, defined as $\sum_i (\log N_i + 1)$, where N_i is the word frequency of the ith neighbor (cf. Massaro & Cohen, 1994). For each stimulus class (i.e., HF words, LF words and nonwords) one-third of the stimuli were four letters long, one-third were five letters long and one-third were six letters long. In addition to the experimental stimuli there were 192 lexical decision practice stimuli, consisting of 48 HF words, 48 LF words, and 96 nonwords. The lexical decision practice stimuli had the same general characteristics as the experimental stimuli. Finally, the stimuli ">" and "<" were used as stimuli to familiarize the subjects with the signal-to-respond procedure. The word and nonword stimuli (and their neighborhood characteristics) can be obtained from http://www.psych.nwu.edu/~ej/remldstimuli.xls.

Design. The experiment consisted of five blocks: (1) a general, non-lexical practice block during which subjects were familiarized with the signal-to-respond procedure. To this aim, we required subjects to classify arrows (">" and "<"). Throughout the experiment, subjects were required to respond immediately after hearing a tone. The tone could be presented at one of six times after the onset of the target stimulus (i.e., deadlines): 75, 200, 250, 300, 350, and 1000 ms. The general practice block consisted of 96 trials. (2) the first lexical decision practice block. In this block, subjects had to make 96 lexical decisions. For half of the subjects, the practice block contained 48 HF words and 48 nonwords, and for the other half of the subjects, the practice block contained 48 LF words and 48 nonwords. (3) the first experimental block. This block consisted of 288 trials. The frequency class of the 144 word stimuli was identical to that of the previous practice block. (4) the second lexical decision practice block, and (5) the second experimental block. Block four and five were identical to block two and three, respectively, except for the fact that new nonwords were

used and the frequency class of the word stimuli was reversed. Only responses to experimental stimuli were analyzed. The experimental stimuli were assigned to each of the six deadlines in a counterbalanced (Latin square) design. Also, two sets of 144 experimental nonword stimuli were assigned either to the block with only HF word stimuli or to the block with only LF word stimuli using a counterbalanced design. The order of the trials was randomly determined for each subject. All word and nonword stimuli occurred only once throughout the experiment. Participants were allowed a short break after completing the first experimental block (block three).

Procedure. Subjects received spoken and written instructions explaining the signal-to-respond lexical decision task. Subjects were instructed to respond immediately after hearing a tone (i.e., the signal-to-respond). In addition, subjects were informed about the frequency of the word stimuli before the start of each block (i.e., "the words in this block are not encountered very often" for LF words, versus "the words in this block are encountered often" for HF words). Each trial started with the 1000 ms presentation of a trial marker (##) at the center of the screen. Next, the trial marker was replaced by the stimulus. In order to further encourage timely responding, the stimulus was removed from the screen at the exact moment the signal-to-respond tone was presented. This 32 ms, 1000 Hz tone could be presented at one of six time intervals after stimulus onset. Subjects gave a 'NONWORD' response by pressing the 'z' key of the keyboard with the left index finger and a 'WORD' response by pressing the '?/' key with the right index finger. When no response was given after 500 ms since the presentation of the tone, the message 'TE LAAT' (Dutch for 'too late') was presented for 1500 ms. When the subject anticipated the tone (i.e., responding faster than 75 ms after presentation of the tone), the message 'TE VROEG' (Dutch for 'too early') was presented for 1500 ms. For all other responses, subjects received feedback on both accuracy

and timing, presented for 2000 ms during which the relevant stimulus was also presented on the screen.

Results

The results of Experiment 1 are presented in Figure 2a and Table 2. Figure 2a shows the accuracy data and Table 2 shows the response latencies. ANOVAs were performed on error percentages and on the mean latencies of correct responses. The data of three subjects were excluded from the analysis because of excess error rate and an apparent failure to obey instructions. Of the remaining 32 subjects, only data falling within a response time window extending from 100 ms to 350 ms after the onset of the tone were analyzed (cf. Hintzman & Curran, 1997). This resulted in the exclusion of 15.8% of the data. Other methods of analysis (e.g., binning the data or using different window-sizes) yielded similar results.

As can be seen in Figure 2a, HF words were responded to more accurately than LF words, $F(1, 31) = 61.7$, $MSE = 242$, $p < .001$. HF words were also classified correctly faster than LF words, $F(1, 31) = 5.7$, $MSE = 810$, $p < .05$. The crucial finding of this experiment is that nonwords presented in a block with only HF words were responded to more accurately than nonwords presented in a block with only LF words, $F(1, 31) = 42.6$, $MSE = 265$, $p < .001$. No effect of word frequency on performance for nonwords was apparent from the response latencies, $F < 1$. For all four stimulus categories, performance increased with processing time, all p's $< .001$.

Discussion

Experiment 1 demonstrated that the effect of word frequency on performance for nonwords (e.g., Joordens et al., 2000; Stone & Van Orden, 1993) is also consistently obtained in the signal-to-respond paradigm where accuracy rather than response time is the dependent variable. The finding that word frequency affects performance for nonwords is predicted by REM-LD. In a simulation study, we tested the prediction of REM-LD under the

conditions of Experiment 1 with the following three parameter values: $\hat{\beta}_1$ (HF words) = .865, $\hat{\beta}_1$ (LF words) = .73, and $\hat{\beta}_2$ = .415. Again simulations of two paradigms were carried out, one in which the words were all HF, and one in which the words were all LF. The results of this simulation are shown in Figure 2b. The model predicts that word frequency affects nonword performance because of the centering aspect of the Bayesian decision mechanism (cf. Equations 4 and 6): If classification accuracy for words is enhanced, for instance by using HF words instead of LF words, this will in turn make nonwords more discriminable and hence leads to an improvement in classification performance for nonwords. REM-LD predicts the effect of nonword lexicality on performance for words for the same reason: If classification accuracy for nonwords is enhanced (e.g., by using nonwords that are not very word-like), this will lead words to be more discriminable, and hence result in an increase in classification accuracy for words. As a final note, the predicted functions fail to capture the somewhat S-shaped form of the observed data. Although the form of the signal-to-respond functions is of secondary interest for this article, we note that variation in the time at which information accumulation begins would tend to produce such an S-shaped result.

Experiment 2

The objective of Experiment 2 was the study of lexical decision performance as a function of processing time, nonword lexicality, word frequency, and, particularly, repetition priming. Experiment 2 was inspired by the work of Hintzman and Curran (1997, Experiment 2). Hintzman and Curran used a signal-to-respond lexical decision task to track the time course of processing for four types of stimuli: (1) HF words, (2) LF words, (3) nonwords created by changing one letter from an HF word, and (4) nonwords created by changing one letter from an LF word. In addition, all stimuli were presented twice (see Hintzman & Curran, 1997, Figure 9, for their results). Because the two types of nonword stimuli did not differ significantly, we collapsed the data over the two types of nonwords to avoid clutter and

re-plotted the Hintzman and Curran data in Figure 3c. As can be seen, performance for HF words is better than LF words (i.e., the word frequency effect). Also, performance for repeated words is better than performance for words that are presented for the first time. This repetition priming effect is more pronounced for LF words than for HF words, thus reducing the word frequency effect (see also Scarborough et al., 1977; Scarborough et al., 1984). For nonwords, prior presentation led to a decrease in performance: repeated nonwords were more likely than novel nonwords to be classified as a word. The inhibitory repetition priming effect for nonwords is of considerable theoretical importance. Logan (1988, 1990) reported substantial facilitatory repetition priming effects for nonwords (i.e., performance for repeated nonwords is better than for novel nonwords), and argued that this finding constitutes evidence for a theory based on automatic retrieval of episodic information (i.e., instance theory). We will discuss the implications of both inhibitory and facilitatory effects of prior presentations for nonwords in more detail later.

One of the most important differences between the current experiment and that of Hintzman and Curran (1997, Experiment 2) is a more powerful manipulation of nonword lexicality. In our experiment, we used two types of nonwords: (1) nonwords such as GREACH created by changing one letter of an existing word, and (2) nonwords such as ANSU that differ in two letters from any existing word. We expected lexical decision performance to be better for the 'two letter replaced' nonwords than for the 'one letter replaced' nonwords. For modeling purposes, we also equated the HF words, LF words, and the 'one letter replaced' nonwords for certain orthographic neighborhood characteristics, as in Experiment 1, using a combined measure for both the number and the frequency of orthographically similar words (cf. Massaro & Cohen, 1994).

Method

Participants. Thirty-seven students of Indiana University participated for a small monetary reward. All participants were native speakers of English and reported normal or corrected-to-normal vision.

Stimulus Materials. We used four types of experimental stimuli: (1) 168 HF English words, each occurring more than 30 times per million according to the CELEX lexical database (Baayen et al., 1993), (2) 168 LF English words, each occurring one or two times per million, (3) 168 pronounceable nonwords created by replacing one letter of an existing word (e.g., GREACH derived from PREACH), (4) 168 pronounceable nonwords differing by at least two letters from any word (e.g., ANSU).[4] As in Experiment 1, the first three stimulus categories were matched on neighborhood structure, having roughly the same summed logarithmic word frequency of the neighbors. The nonword stimuli were constructed by applying the same rules as the ones used in Experiment 1. All stimuli were four, five, six, or seven letters long, occurring in the respective proportions 2:2:2:1. In addition to the experimental stimuli there were 72 fillers and 72 lexical decision practice stimuli, each group consisting of 18 HF words, 18 LF words, 18 'one-letter replaced' nonwords, and 18 'two-letters replaced' nonwords. Both fillers and lexical decision practice stimuli had the same general characteristics as the experimental stimuli. Finally, the stimuli ">" and "<" were used as stimuli to familiarize the subjects with the signal-to-respond procedure. The word and nonword stimuli can be obtained from http://www.psych.nwu.edu/~ej/remldstimuli.xls.

Design. The experiment consisted of three phases: (1) a general, non-lexical practice phase during which subjects were familiarized with the signal-to-respond procedure. As in Experiment 1, we required subjects to classify arrows (">" and "<"). Throughout the experiment, subjects were required to respond immediately after hearing a tone. The tone could be presented at one of six times after the onset of the target stimulus (i.e., the same deadlines as used in Experiment 1): 75, 200, 250, 300, 350, and 1000 ms. The general

practice phase consisted of 300 trials. (2) a lexical decision practice phase. In this phase, subjects had to make 96 lexical decisions to 72 different stimuli (i.e., one block of 48 new stimuli followed by a block of 24 new stimuli and 24 stimuli from the first block). (3) the experimental phase. This phase consisted of 30 blocks of 48 trials each, resulting in a total of 1440 trials. In each block except the first, half of the stimuli were new, and half of the stimuli had been presented in the previous block (i.e., a blocked design was used). In a blocked design (cf. Hintzman & Curran, 1997, Experiment 2; Logan, 1988, Experiment 3; Smith & Oscar-Berman, 1990), the presentation condition (i.e., 1^{st} or 2^{nd} presentation) of a stimulus and the total number of trials preceding the stimulus are not confounded. Therefore, any change in performance over the number presentations of a stimulus is due to a stimulus specific repetition effect and can not be ascribed to some general practice effect, skill learning, fatigue, or a criterion-shift due to improvement for a subset of stimuli (for a more detailed discussion see Wagenmakers, Zeelenberg, Steyvers, Shiffrin, & Raaijmakers, 2001). The transition from one block to another block was not marked in any way and from the point of view of the participants the experiment consisted of one long sequence of trials. The first block consisted of 48 filler stimuli. In the final block, the remaining 24 filler stimuli were added to 24 experimental stimuli that had been presented in the previous block. Each block consisted of an equal number of word and nonword stimuli, and each of the six deadlines occurred eight times in one block. Only responses to experimental stimuli were analyzed. The experimental stimuli were assigned to each of the six deadlines in a counterbalanced (Latin square) design. The order of the trials was randomly determined for each subject. Participants were allowed two short breaks, one after 480 trials in the experimental phase, and one after 960 trials in the experimental phase.

Procedure. The procedure was identical to the procedure of Experiment 1, with the exception that the feedback on response latency and accuracy was presented for 1500 ms

instead of 2000 ms, and the stimulus was not presented on the screen while this feedback was presented. In addition, of course, all messages (i.e., too late, too early) were translated from Dutch to English.

Results

The results of Experiment 2 are presented in Figure 3a and Table 3. Figure 3a shows the accuracy data and Table 3 shows the response latencies. ANOVAs were performed on the mean latencies of correct responses and on error percentages. The data of 14 subjects were excluded from the analysis, either because of evident failure to obey instructions, excess error rate, or poor response timing (i.e., over 30% of the responses outside the response window mentioned below).[5] Of the remaining 23 subjects, only data falling within a response time window extending from 100 ms to 350 ms after the onset of the tone were analyzed (cf. Experiment 1). This resulted in the exclusion of 18.8% of the data. Other methods of analysis (e.g., binning the data or using different window-sizes) yielded similar results.

As is apparent from Figure 3a and Table 3, both response latency and response accuracy increased with an increase in deadline, all p's < .001. HF words were responded to more accurately than LF words, $F(1, 22) = 224.8$, $MSE = 174$, $p < .001$. HF words were also classified correctly faster than LF words, $F(1, 22) = 73.5$, $MSE = 199$, $p < .001$. These word frequency effects for both response accuracy and response latency were attenuated by a prior presentation, $F(1, 22) = 49.5$, $MSE = 73$, $p < .001$, and $F(1, 22) = 5.3$, $MSE = 53$, $p < .05$, respectively. Nonwords that differed in two letters from a word were both classified more accurately and classified correctly faster than nonwords that differed in only one letter from a word, $F(1, 22) = 586.7$, $MSE = 51$, $p < .001$, and $F(1, 22) = 11.4$, $MSE = 134$, $p < .01$, respectively.

Facilitatory effects of repetition priming were observed for both HF stimuli and LF stimuli. More specifically, both HF words and LF words were responded to more accurately

on their second presentation than on their first presentation, $F(1, 22) = 17.7$, $MSE = 57$, $p < .001$, and $F(1, 22) = 209.6$, $MSE = 65$, $p < .001$, respectively. HF words and LF words were also classified correctly faster on their second presentation than on their first presentation, $F(1, 22) = 11.2$, $MSE = 84$, $p < .01$, and $F(1, 22) = 54.0$, $MSE = 55$, $p < .001$, respectively. Figure 4a also shows that for nonwords differing in only one letter from an existing word (i.e., 'one letter replaced' nonwords), inhibitory effects of repetition priming were observed with respect to response accuracy. More specifically, 'one letter replaced' nonwords were responded to less accurately on their second presentation than on their first presentation, $F(1, 22) = 7.4$, $MSE = 101$, $p < .05$. In addition, 'one letter replaced' nonwords were responded to faster on their second presentation than on their first presentation, $F(1, 22) = 12.0$, $MSE = 43$, $p < .01$. With respect to nonwords differing in two letters from any existing word (i.e., 'two-letters replaced nonwords'), the effects of repetition priming did not reach significance for either response accuracy or response latency, $F(1, 22) = 2.2$, $MSE = 44$, $p > .15$, and $F(1, 22) = 2.8$, $MSE = 42$, $p > .10$, respectively.

Discussion

Experiment 2 showed substantial effects of stimulus type. Performance for HF stimuli was better than performance for LF stimuli (i.e., the word frequency effect) and performance for 'two letter replaced' nonwords was better than for 'one letter replaced' nonwords. In addition, prior presentation reduced the word frequency effect. Also, 'one letter replaced' nonwords showed inhibitory effects of nonword repetition. In a very similar experiment[6], Wagenmakers et al. (2001, Experiment 3) showed that inhibitory repetition priming for nonwords can be obtained for the 'two letters replaced' nonwords used in this study, albeit of a smaller magnitude than that observed for 'one letter replaced' nonwords. In general, then, the data from Experiment 2 are consistent with previous findings (i.e., Hintzman & Curran, 1997; Wagenmakers et al., 2001).

How can REM-LD account for the present results, and those of Hintzman and Curran (1997)? In the previous sections, we discussed how REM-LD models the word frequency effect (i.e., a higher value of β_1 for HF words than for LF words) and the nonword lexicality effect (i.e., an higher value of β_2 for word-like nonwords than for nonwords relatively dissimilar to words). To model the effect of repetitions for words we assume that study and test of a word adds information about the current presentation and context to the lexical/semantic trace of the tested word. In the REM framework generally, implicit memory effects are ascribed to such a mechanism (e.g., see Shiffrin & Steyvers, 1997; Schooler et al., 2001). Further, this assumption is consistent with the assumption in REM that such a mechanism is responsible for the development of lexical/semantic traces through repetitions of a word over developmental time (e.g., Shiffrin & Steyvers, 1997). Finally, we note that the approach in this respect is consistent with the approach to word frequency that is used in most models (e.g., McClelland & Rumelhart, 1981; Morton, 1969; Wagenmakers, Zeelenberg, & Raaijmakers, 2000; but see Ratcliff & McKoon, 1997). Thus in REM-LD, if the probe includes low level physical features like font, and current context features, these will produce better matches to traces that have been augmented by such features, namely those that represent traces of repeated words. Rather than implement this idea in detail, possibly by distinguishing types of features, we simply assumed that prior presentation increases the value of β_1. This simplification is quite sufficient for present purposes.

It is somewhat less straightforward to model the repetition priming effect for nonwords. It is assumed in the REM approach that presentation almost always produces storage of an incomplete and error prone episodic trace of the study event. Thus one approach would assume that this episodic trace is activated and produces the additional matching that is seen as an inhibitory effect in the data. However, this would introduce a different mechanism than that used for words. Thus, in an attempt to create a model for

lexical decision that is both conceptually and mathematically transparent, we adopt an approach based on that used for words: it is assumed that only lexical/semantic traces are matched to the presented stimulus. In particular, it is assumed that on the first presentation of a nonword (e.g., GREACH), participants will retrieve a number of words that are orthographically and/or phonologically similar to the test string. We further make the simplifying assumption that on a certain proportion of trials the subjects will retrieve one of the similar words (e.g., PREACH).[7] For instance, after the subject is presented with GREACH, he or she might think something like "this stimulus looks very similar to PREACH". In other words, the presentation of a nonword will sometimes lead to a trace-specific retrieval of an orthographically similar word representation. Although this example provides a description of a retrieval event that is 'aware and conscious', it is quite conceivable that such retrieval occurs implicitly, without lasting awareness. Whatever the degree of awareness, this retrieval event could produce storage of current context information in the trace of the retrieved word. When the nonword is tested again, the trace of this similar word will be part of the activated set of ten most similar traces, and will contribute more matching due to the additional context features stored. Consequently, the retest will lead to a relatively high estimate of familiarity (i.e., posterior odds ratio Φ), and bias the system to give a 'WORD' response. We implement this idea in the simplest way possible, by assuming that one of the lexical/semantic traces in the activated set has a slightly higher value of β_2 than on the first presentation.

Figure 3d shows how REM-LD handles the data from Hintzman and Curran (1997; Exp. 2, Figure 7; see our Figure 3c). Hintzman and Curran used seven deadlines instead of six. In their experiment, the signal-to-respond could be presented either at 75, 125, 200, 300, 400, 600, or 1000 ms after stimulus onset. Again, we let REM-LD 'respond' after adding 200 ms to these deadlines. The parameter values are: $\hat{\beta}_1$ (HF words) = .82, $\hat{\beta}_1$ (LF words) = .757,

the increase in $\hat{\beta}_1$ due to prior presentation for both HF and LF words = .063, $\hat{\beta}_2$ = .433, and the increase in $\hat{\beta}_2$ for one lexical/semantic trace due to prior presentation of a nonword = .072.

Figure 3b shows how REM-LD can account for the results of Experiment 2 (cf. Figure 3a). The parameter values are: $\hat{\beta}_1$ (HF words) = .82, $\hat{\beta}_1$ (LF words) = .73, the increase in $\hat{\beta}_1$ due to a prior presentation for both HF and LF words = .045, $\hat{\beta}_2$ (word-like nonwords or pseudowords) = .46, $\hat{\beta}_2$ (less word-like nonwords) = .433, the increase in $\hat{\beta}_2$ for one lexical/semantic trace due prior presentation of a word-like nonword = .072, and the increase in $\hat{\beta}_2$ for one lexical/semantic trace due to prior presentation of a less word-like nonword = .036.

In both experiments the materials are mixed across trials for each participant, so in the simulations (Figures 3b and 3d) the different values of β_1 and the different values of β_2 are used to generate the vector values and hence determine the number of matches and mismatches, but the calculations of the likelihood ratios are based on a single estimate of β_1, the arithmetic mean of the four β_1 values, and a single estimate of β_2, the arithmetic mean of the two (Fig. 3d) or four (Fig. 3b) β_2 values. Although the predictions are only qualitative, they are sufficient to illustrate that the model captures the observed pattern of results. We would like to stress that the performance of the model is not strongly dependent on specific parameter values. Most of the predictions of the REM-LD model are generated by the Bayesian decision mechanism that is inherent to the model. Consequently, the predicted results hold qualitatively across a range of parameter values and are quite general. In order to make this point clear, we attempted to reduce the number of free parameters to a minimum. In order to obtain a fit that is quantitatively closer, we could have let the criterion for

responding 'WORD' vary slightly from its optimal value of $\Phi = 1$. Also, we could have adjusted parameters $\underline{t}_0$ and $\underline{b}$ in the function that gives the probability of correctly retrieving a trace feature by time $\underline{t}$ (i.e., Equation 1). In addition, other functions (e.g., a sigmoid retrieval function) than the one given in Equation 1 could possibly have provided an even closer fit to the data. However, the precise shape of the retrieval function is not an inherent property of the REM-LD model, and hence we opted to illustrate the behavior of the model using the same retrieval function for all simulations reported here.

Note that in both simulations, the attenuation of the word frequency effect due to prior presentation follows from the differential effect that the same increase in $\hat{\beta}_1$ has on HF words and LF words. Turning to nonwords, recall that we propose that negative repetition priming for nonwords occurs because current context information is added to the trace of a similar word, a trace that is retrieved following presentation of the nonword. Assuming that such retrieval is harder and less likely for test strings that are less similar to words, the negative effect for such test strings will be smaller. This idea was implemented in the simulation by setting the increase in β_2 for one lexical/semantic trace due to prior presentation of a nonword to a lower value for nonwords that are dissimilar to words (i.e., .036) than for nonwords that are relatively similar to words (i.e., .072). In sum, Figures 3b and 3d show that REM-LD can, at a qualitative level, predict the observed effects on performance of processing time, word frequency, repetition priming, and nonword lexicality.

Logan (1988, 1990; see also Wagenmakers et al., 2001) reported substantial facilitatory effects due to prior presentation of a nonword. That is, in some experiments subjects classify nonwords more accurately on their second presentation than on their first presentation. In its present form, REM-LD predicts less accurate nonword performance (basically due to increased familiarity). It should be noted that under speed-stress such as imposed by the signal-to-respond paradigm, facilitatory nonword repetition priming is

usually not observed in lexical decision (Wagenmakers et al., 2001). In a study that provides some insight into these discrepant results, Wagenmakers et al. (2001; see also Smith & Oscar-Berman, 1990) presented empirical evidence that two opposing processes jointly determine performance for repeated nonwords: (1) a inhibitory familiarity process as for instance implemented by the REM-LD model, and (2) a facilitatory process that is perhaps based on automatic episodic retrieval of the interpretation associated with the nonword stimulus on its initial presentation (i.e., "I remember GREACH is a nonword", cf. Logan, 1990; Tenpenny, 1995; but see Bowers, 2000). That is, a particular form of episodic retrieval could in some studies dominate the familiarity factor that we propose affects lexical access. We will not carry this point further, because it goes beyond the scope of this article to extend the present model by adding an episodic retrieval component.

Extensions of the REM-LD Model

Up to this point we have shown how REM-LD provides a parsimonious explanation for the effects of word frequency, repetition priming, and nonword lexicality as observed in a signal-to-respond lexical decision task. We would like to stress that REM-LD correctly predicts the interactions of the above effects (e.g., the attenuation of the word frequency effect when the word-likeliness of the nonwords is reduced, the enhanced classification performance for nonwords when HF words are used instead of LF words), not by a careful exploration of the entire parameter-space, but by application of the likelihood-based statistical decision process that forms an integral part of the model. The above phenomena were selected for modeling based on their generality, robustness, and theoretical importance. However, our choice was up to a certain point arbitrary, and it is certainly possible to extend the REM-LD model to handle other phenomena than the ones considered so far. In this section we will tentatively explore how REM-LD can be applied to the pseudohomophone effect and the prediction of response latencies.

The Pseudohomophone Effect

For simplicity, we have so far assumed that the probe-to-trace comparison process involves only orthographic features (cf. Grainger & Jacobs, 1996). Therefore, REM-LD in its present simple form does not address the role of phonology in visual word recognition (or, more specifically, in lexical decision). Note that in the lexical decision task, activation of phonology is not required for successful performance as the distinction between a word and a nonword is purely based on orthography. Nonetheless, several findings have unambiguously demonstrated that phonological information does play an important role in lexical decision (e.g., Frost, 1998; Stone, Vanhoy, & Van Orden, 1997; Van Orden, 1987).

One of the most robust findings that attest to the role of phonology in lexical decision is the pseudohomophone effect (Coltheart, Davelaar, Jonasson, & Besner, 1977; Rubinstein, Lewis, & Rubinstein, 1971), that is, nonwords that are pronounced as words (e.g., BRANE) are more difficult to correctly reject than nonwords that are not pronounced as words (e.g., SLINT). It is relatively straightforward to extend REM-LD to account for the pseudohomophone effect. We assume that there are stages of processing that occur automatically en route to construction of the set of probe features, and that part of these stages involves production of phonological features. Such features are of course also part of the lexical/semantic representations in memory. Hence the matching process used to produce likelihood ratios includes both orthographic and phonological features. More specifically, we assume that a lexical trace contains $\underline{k_p}$ = 10 phonological features, in addition to $\underline{k_o}$ =15 orthographic features.[8] For both words and regular nonwords, $\hat{\beta}_1$ (orthography) = $\hat{\beta}_1$ (phonology) and $\hat{\beta}_2$ (orthography) = $\hat{\beta}_2$ (phonology), that is, the probability of matching a feature (i.e., $\hat{\beta}_1$ and $\hat{\beta}_2$ when the probe does and does not correspond to a trace, respectively) is the same for orthographic and phonological features. The difference between regular nonword probes and pseudohomophone probes such as BRANE is that the latter have a

particular lexical trace (e.g., BRAIN) for which the phonological information matches with probability β_1 instead of β_2. In other words, pseudohomophones have an average a higher odds ratio Φ than regular nonwords because the phonological information of the pseudohomophone probe (e.g., BRANE) will tend to match the phonological information of a similar sounding lexical trace (e.g., BRAIN), boosting the likelihood ratio that the probe matches the phonologically similar (but orthographic dissimilar) lexical trace.

Figure 4 shows two exploratory simulations of the pseudohomophone effect in a signal-to-respond setting. In both simulations we let the model respond after 275, 350, 450, 550, and 650 ms. Each trace consisted of 15 orthographic features and 10 phonological features, $\hat{\beta}_1$ = .8 and $\hat{\beta}_2$ = .35. The top left panel show the predictions of REM-LD when orthographic features and phonological features become available at the same rate. The bottom left panel of Figure 4 shows how the probability of activating/retrieving a feature increases over time (according to Equation 2). As can be seen from Figure 4, top left panel, classification performance for pseudohomophones (e.g., BRANE) is consistently lower than for regular nonwords (e.g., SLINT). The top right panel shows a second simulation of the pseudohomophone effect, this time using a different activation function for phonological features. Figure 4, bottom right panel shows that the activation function for phonological features first increases at the same rate as the activation function for orthographic features, but decreases after 450 ms. Specifically, the equation for the activation function of phonological features is identical to Equation 2 if $t \leq tp$; when $t > tp$, the activation function is given by $\exp[-b(t-tp)] - \exp[-b(t-t_0)]$. The form of this activation function reflects the hypothesis that in the first stages of processing phonological information is computed automatically, whereas it can be suppressed or discounted in later stages of processing. Such a process of discounting is plausible given that subjects should be able to correctly classify pseudohomophones as nonwords when given enough time.

The two simulations shown in Figure 4 serve to illustrate how REM-LD can be extended to handle the pseudohomophone effect. The simulations also show how the signal-to-respond paradigm can potentially be used to infer the relative time-course of activation of orthographic versus phonological information. Of course, the presented simulations are speculative in the sense that to our knowledge a signal-to-respond experiment with pseudohomophones has yet to be performed.

It is worth mentioning one recent result with respect to the role of phonology in lexical decision: Ziegler, Jacobs, and Klueppel (2001) replicated in German results from Van Orden (1991; Van Orden, Stone, Garlington, Markson, Pinnt, Simonfy, & Brichetto, 1992) showing that pseudohomophones derived from HF words are faster classified (i.e., correctly rejected) than pseudohomophones derived from LF words. Ziegler et al. (2001) noted that this result is at odds with predictions of the standard versions two of the most popular models for lexical decision (i.e., the Multiple Read-Out Model, Grainger & Jacobs, 1996, and the Dual Route Cascaded model, Coltheart et al., 2001). Such a result falls naturally out of REM models that incorporate differentiation (e.g., Shiffrin & Steyvers, 1997; see also Shiffrin, Ratcliff & Clark, 1990): the idea that traces stored better are better differentiated from (i.e., less similar to) traces of other items. In the REM-LD model, we could assume that for both HF and LF pseudohomophones, their corresponding word is in the activated set. However, differentiation would mean that HF similarity would be lower than LF similarity, reflected in the β values. To illustrate with an example, the well-stored information about BRAIN would produce relatively little confusion with BRANE, but the not-so-well stored information about FLOTSAM would produce relatively more confusion with FLOTSUM.

Prediction of Response Times

Throughout this article, we have used a lexical decision signal-to-respond paradigm (Antos, 1979; Hintzman & Curran, 1997). In this paradigm, the variable of interest is

response accuracy, or more specifically the increase in classification accuracy with processing time. For the dominant paradigm in lexical decision, however, the variable of interest is response latency or response time (RT). In other words, in the majority of lexical decision experiments, subjects are typically instructed to 'respond as quickly as possible without making errors' or 'respond as quickly and accurately as possible'. These instructions (henceforth 'respond-when-ready') are meant to result in very few errors (e.g., about 5%), so that the difference in RT between various conditions is a valid indication of the differential processing demands associated with these conditions. It is worthwhile to consider how REM-LD can be extended to predict RTs in the respond-when-ready procedure, both because of the popularity of this procedure, and because it is desirable for any model to be able to account for RT as well as response accuracy.

In the REM-LD model, the odds ratio Φ equals 1 (i.e., no evidence to support either the 'WORD' response or the 'NONWORD' response) when no probe features have yet been compared to trace features. As probe and trace features become available for matching, information accumulates and the odds ratio starts to drift. Generally, the odds ratio will drift toward high values when the probe is a word, and will drift toward low values when the probe is a nonword. It is important, however, to realize that the drift of the odds ratio is noisy: sometimes the odds ratio will drift toward high or low values when the probe is a nonword or word, respectively.

In the signal-to-respond paradigm, the evidence (i.e., the odds ratio Φ) is evaluated at the time the system knows it has to respond (i.e., at some desired time $\underline{t}$ after stimulus onset): when $\Phi > 1$ the evidence favors the 'WORD' response, and when $\Phi < 1$ favors the 'NONWORD' response. In the respond-when-ready paradigm, in contrast, the system has to decide by itself when to respond. Intuition suggests that it is desirable for a system to respond, say, 'WORD' when there is reliable evidence in support of the 'WORD' response

over the alternative 'NONWORD' response. Responding before reliable evidence has accumulated will lead to many incorrect decisions; responding after reliable evidence has accumulated will lead to unnecessarily slow RTs.

The problem of when to halt processing of information and decide has been formally studied both in cognitive psychology and in other fields. For example, the use of an optimal stopping rule is important for quality control of industrial products (e.g., Sveshnikov, 1978, pp. 346-368). Consider the problem of assessing with some predetermined amount of confidence whether a batch of products is good or bad. When individual items are sampled from the batch one-by-one (and labeled after inspection to be 'defect' or 'not defect'), an optimal stopping rule provides the best criterion of when to stop sampling individual items and label the entire batch 'defect' or 'not defect'. The problem of optimal stopping rules was addressed by Wald (1947) and applied to decision-making in psychology by, among others, Edwards (1965), Stone (1960), and Laming (1968, 1973). The optimal stopping rule in the case of sequential sampling is given by the probability ratio test (see also Townsend & Ashby, 1983). That is, if the probability of stimulus A or stimulus B given the data sampled up to time $\underline{t}$ is denoted by $P_t(A \mid data)$ and $P_t(B \mid data)$, respectively, the probability ratio $\lambda = \frac{P_t(A \mid data)}{P_t(B \mid data)}$ gives the strength of evidence, based on the data, in favor of A over B. At time $\underline{t}$, if λ exceeds a preset upper bound, the response associated with stimulus A is executed. If λ exceeds a preset lower bound, the response associated with stimulus B is executed. In both cases, the decision time equals $\underline{t}$. When neither boundary has been reached, sampling continues. This procedure is termed the sequential probability ratio test (SPRT) and has been explored in some detail as a model for response time by Laming (1968, 1973).

Although the REM-LD model differs from the SPRT model in that the REM-LD model calculates the odds ratio based on the average of ten probe-to-trace likelihood ratios (cf. Equation 8), the underlying principles are in fact identical. Hence, the most principled

method to generate RTs from the REM-LD model is to monitor how the odds ratio Φ drifts over time, and respond when Φ reaches an upper or lower boundary. This REM-LD model for response times would inherit many of the desirable properties of the SPRT approach. To name one, the SPRT model accounts for the speed-accuracy trade-off in a straightforward way. The distance between the upper and lower boundary for the odds ratio Φ corresponds to the amount of evidence required to make a decision. Thus, when accuracy is stressed subjects can move the boundaries out, requiring greater certainty (i.e., more evidence or a more extreme odds ratio) before a choice is made. This greater certainty comes at the cost of having to sample more information, on average, before a response can be made.

Finally, note that the SPRT approach taken here is very similar to modeling RTs by means of a random walk model (or its continuous version, the diffusion process). The difference between a random walk/diffusion model (e.g., Ratcliff, Gomez, & McKoon, in press; Ratcliff, 1978) and the SPRT model is that in a random walk model the sampled units of information are evaluated with reference to a single criterion. Thus, in a random walk model the impact of a single unit of information on the decision process is either constant (i.e., +1 or –1, a unit step toward or away from the top boundary) or may vary based on the distance to the reference criterion. In the SPRT model, the information sampled from the stimulus is evaluated for its diagnosticity. Highly diagnostic information leads to a sizeable contribution to the decision process, whereas information that is not very diagnostic contributes little to the decision process. Thus, SPRT and the random walk model are closely related, all the more so since in most cases there will be a substantial correlation between diagnosticity and the distance to a reference criterion.

Model Evaluation in Lexical Decision

Several quantitative models for lexical decision have already been proposed, and many are able to account for an impressive amount of data (e.g., Coltheart et al., 2001;

Grainger & Jacobs, 1996). The existence of different models for the same task or process automatically leads to the question which of the candidate models is most likely to be the best. More generally, we would like to be able to rigorously evaluate the different candidate models, and perhaps select the model we believe to be the best as the most likely abstraction of what processes underlie performance in the lexical decision task.

The process of model evaluation (e.g., Myung, Forster, & Brown, 2000) is a challenging task. First, one needs to agree on a set of criteria against which each model can then be tested. Jacobs and Grainger (1994; see also Myung & Pitt, 1997) mention the following four: (1) generality (is the model applicable to other paradigms and experimental situations than the one it was originally applied to?); (2) explanatory adequacy (are the assumptions of the model plausible? Do the assumptions of the model follow from its structure or are they more ad hoc?); (3) complexity (how many free parameters does the model have? What is the functional form of these free parameters?); and (4) descriptive adequacy (how accurately does the model describe the data that is used to evaluate it?). To this list of model criteria we might add several other (related) criteria, such as, (5) falsifiability (how easily can the model be proven wrong?); (6) prediction adequacy (does the model make testable predictions? Are these predictions novel or perhaps even counterintuitive --Roberts & Pashler, 2000-- ?); (7) number of implicit or explicit assumptions (how many assumptions were needed to construct the model in the first place?); (8) conceptual and mathematical clarity (can the predictions of the model be easily understood?).

When all of the above criteria are taken into account the problem of model evaluation would become intractable: many of the criteria are subjective, or at least very difficult to quantify (e.g., the extent to which a model is judged to be conceptually clear or to have plausible assumptions). Moreover, even when all criteria would be objectively quantifiable, it

would still be problematic how to weigh the importance of the different criteria. For instance, should a model that has many parameters, is difficult to falsify, but can easily be generalized to other paradigms be preferred over a model that has only few parameters, is easily falsified but does not generalize well? The difficulty of model evaluation in lexical decision is exacerbated by the fact that many existing models focus on different phenomena and different data sets. In short, it is impossible to say with certainty which model for lexical decision is best (provided that nested models and falsified models are excluded from consideration). In the next section, we will therefore pursue the more realistic goal of comparing REM-LD to other quantitative models of lexical decision by noting differences and similarities with an emphasis on the data presented in this article.

Comparison to Other Quantitative Models of Lexical Decision

Quantitative models of lexical decision other than REM-LD have thus far not been applied to the signal-to-respond paradigm. A discussion of how these models can handle the results presented here is therefore to some degree speculative. In this section we will discuss the following models: (1) the Multiple Read-Out Model (MROM; Grainger & Jacobs, 1996); (2) the Dual Route Cascaded model of visual word recognition and reading aloud (DRC; Coltheart et al., 2001); (3) Parallel Distributed Processing models (PDP models; e.g., Seidenberg & McClelland, 1989; Plaut, McClelland, Seidenberg, & Patterson, 1996; Plaut, 1997); and (4) random walk models (e.g., Gordon, 1983; Ratcliff et al., in press; Stone & Van Orden, 1993). All of the above models are quantitative models that have recently been implemented or adjusted. In fact, we were unable to find any other published process models for lexical decision that have been implemented to generate quantitative predictions.[9]

The Multiple Read-Out Model

The Multiple Read-Out Model (MROM; Grainger & Jacobs, 1996) is in many ways similar to the Dual Route Cascaded model (DRC; Coltheart et al., 2001), and both models

use the same decisional mechanisms to account for lexical decision performance. Because our suggestions for adjusting MROM to account for performance in the signal-to-respond paradigm apply equally well to DRC we will discuss these suggestions after both MROM and DRC have been briefly introduced.

The representational assumptions of MROM are inherited from the Interactive Activation Model (IAM; McClelland and Rumelhart, 1981; Rumelhart & McClelland, 1982). In MROM, each word is represented by a separate local lexical unit or node. It is assumed that upon presentation of a printed word the incoming visual information from sub-lexical units such as letters and features gradually activates the associated word nodes. Each word node thus accumulates evidence that stems from lower levels of analysis. At the same time, activated word nodes inhibit each other by driving down the activation levels of their competitors (i.e., lexical inhibition). In MROM, word frequency is modeled as an increase in the resting level of the word nodes. This reflects the fact that high frequency words are more likely a priori, and hence it is adaptive to give such words a head start in the identification process (cf. Broadbent, 1967). Long-term repetition priming is not explicitly modeled by MROM, but one can assume a mechanism similar to that for word frequency.

MROM assumes that performance in lexical decision is based on three response criteria or thresholds. The first is a fixed criterion for the activation of a single lexical word node. When this single unit criterion is reached by any of the word nodes, the stimulus is identified as a specific word, and the corresponding lexical decision ‘WORD’ is made. Many researchers have, however, argued that under certain circumstances correct lexical decisions can be made without such lexical access to a specific word representation. For instance, when words have to be distinguished from ‘easy’, not very word-like nonwords (e.g., DJIPK), a superficial first-pass analysis of the stimulus might already provide sufficient evidence for

the correct response (e.g., Balota & Chumbley, 1984, p. 352; Balota & Spieler, 1999). Such a first-pass judgment is generally said to be based on familiarity.

The second criterion that MROM uses is based on the summed lexical activation over all word nodes. This criterion is specific to the lexical decision task, as it does not critically depend on selection of one particular word (such a selection is necessary for successful performance in other visual word recognition tasks such as perceptual identification). When the summed unit criterion is reached, the 'WORD' response is given. This criterion can be strategically set, depending on the list context and task instructions. For instance, when the word and nonword stimuli are orthographically dissimilar, and hence generate distinct overall values of familiarity (i.e., words activating the entire lexicon to a higher degree than nonwords), it is adaptive to lower the summed unit criterion for responding based on this discriminative information. Instructions stressing speed over accuracy are also assumed to lower the summed unit criterion.

Finally, MROM's third criterion provides a mechanism for generating a 'NONWORD' response. The nonword criterion takes the form of a temporal deadline criterion T, that is, the system defaults to the 'NONWORD' response when neither the single unit criterion nor the summed unit criterion have been reached by time T. If this deadline criterion were to be fixed, or vary stochastically around a fixed mean value, this would imply that nonword stimuli are always responded to at about the same speed. However, performance for nonword stimuli shows systematic effects of list context, effects of similarity to the word stimuli in the experiment, effects of similarity to words in general, and effects of task instructions such as stressing speed over accuracy. To account for these effects the temporal deadline in MROM cannot not fixed but needs to be variable (cf. Coltheart et al., 1977). As for the summed unit criterion, the setting of the temporal deadline criterion is assumed to be under strategic control. For instance, when the summed activation of all word nodes is high early in

processing, this constitutes evidence that the stimulus might be a word. Consequently, the temporal deadline is extended (and the summed unit criterion is lowered).

To summarize, when either the single or the summed unit criterion for lexical activation is reached before the temporal deadline, a 'WORD' response is made. When the temporal deadline is reached before either of the two activation criteria, this results in a 'NONWORD' response (for an illustration see Grainger & Jacobs, 1996, Figure 2). The flexible decision process enables MROM to handle a large number of phenomena in lexical decision. MROM has been applied to effects of neighborhood density and neighbor frequency (but see Davis, 1999, Chapter 7, and Paap, Johansen, Chun, & Vonnahme, 2000), and the model can also handle the frequency blocking effect (e.g., Glanzer & Ehrenreich, 1979; Gordon, 1983; Stone & Van Orden, 1993). Also, MROM can account for the effect of nonword lexicality (Grainger & Jacobs, 1996, p. 529), and for the increase in performance for nonwords when HF word stimuli are used instead of LF stimuli (Grainger & Jacobs, 1996, Figure 27). MROM has further been extended to account for phonological effects (i.e., MROM-p; Jacobs, Rey, Ziegler, & Grainger, 1998).

Both MROM and DRC, discussed next, account for a wide variety of phenomena in lexical decision and are arguably the most specific and the most complete quantitative models of lexical decision. A discussion of how MROM and DRC can be applied to the signal-to-respond paradigm is postponed until after introducing the DRC model.

The Dual Route Cascaded Model

The DRC model (Coltheart et al., 2001) has recently been developed to account for a wide range of empirical phenomena in both reading aloud (but see Seidenberg, Zevin, & Harm, 2002) and lexical decision. A detailed description of DRC is well beyond the scope of this article, and we focus instead on the more global properties of the model. The development of the DRC model was guided by earlier work of Morton (1969) and

McClelland and Rumelhart (1981). As in MROM, low level visual information increases the activation level of associated word nodes in an orthographic lexicon, and word frequency acts to increase activation levels regardless of the visual input (Coltheart et al., 2001, p. 216). Long-term repetition priming has not yet been modeled by DRC.

DRC has a modular architecture (see Figure 7 in Coltheart et al., 2001, for an illustration), the separate modules being 'visual feature units', 'letter units', 'orthographic input lexicon', 'semantic system', phonological output lexicon', 'grapheme-phoneme rule system', and a 'phoneme system'. DRC is a cascaded model because the modules continuously pass excitatory or inhibitory activation on to other modules.

Reading aloud (i.e., generating the pronunciation of a word from the visual input) can be accomplished in DRC via three pathways. The first pathway for translating print to speech is called the lexical nonsemantic route, and is characterized by activation passing through the following sequence of modules: 'visual letter features' → 'letter units' → 'orthographic input lexicon' → 'phonological output lexicon' → 'phoneme system'. The second route is the lexical semantic route, which differs from the above sequence by insertion of the 'semantic system' module between the 'orthographic input lexicon' and the 'phonological output lexicon'. The third route from print to speech is the grapheme-phoneme conversion route, which bypasses the orthographic lexicon entirely and proceeds as follows: 'visual letter features' → 'letter units' → 'grapheme-phoneme rule system' → 'phoneme system'. The interested reader is referred to Coltheart et al. (2001) for an overview of empirical results in reading aloud that are consistent with this architecture, as well as a historical overview of how the architecture of DRC was developed.

Lexical decision in the DRC model is solely based on activation from the 'orthographic input lexicon'. The decisional mechanisms also used by MROM (i.e., the three criteria described earlier) are then brought to bear on the activation of the word nodes in this

orthographic lexicon. DRC can explain several important effects in lexical decision, among which certain neighborhood effects (e.g., Andrews, 1997), and the effect of pseudohomophony. It should be pointed out that the DRC model explains the effect of pseudohomophony by utilizing feedback connections from the 'phonological output lexicon' to the 'orthographic input lexicon'. DRC has not yet been applied to semantic effects such as semantic priming, or concreteness. However, we would like to note that DRC can potentially handle such results, since the 'semantic system' module also has feedback connections to the 'orthographic input lexicon'.

Discussion of how to apply MROM and DRC to the signal-to-respond paradigm.

It is important to note that the decisional mechanisms that DRC uses to account for lexical decision are derived from MROM, and the following discussion on how to model the signal-to-respond paradigm therefore applies equally well to DRC as it does to MROM. As mentioned in the Introduction, the 'NONWORD' response is a default response in MROM/DRC. A 'NONWORD' response is given when neither of the two activation criteria (i.e., the single unit criterion and the summed unit criterion) have been reached before the temporal deadline T. It is unclear to us what MROM/DRC predicts when the system is forced to respond before any of the three criteria has been reached. This situation will presumably arise when subjects are forced to respond at specific short deadlines after stimulus onset, such as those imposed by a signal-to-respond procedure.

In the standard MROM/DRC application the temporal deadline T is set by the subject. It is not entirely clear where to set the deadline criterion T in the signal-to-respond procedure, but one might let T be determined by the imposed deadline for responding, so that a 'NONWORD' response would be given when neither of the two activation criteria has been reached by the imposed deadline. However, this proposal would lead the system to display a very large bias toward the 'NONWORD' response at the early stages of processing (i.e.,

when it is unlikely that either of the activation criteria have been reached), and this is clearly not what is observed in the data.

An approach that might overcome the large bias to respond 'NONWORD' is one in which the system adjusts the summed unit criterion as a function of processing time: the summed unit criterion is set low when subjects are forced to respond relatively fast and the criterion gradually increases as the signal-to-respond is presented later. This criterion drift reflects the expectation of the system. Even if the stimulus is a word, it is unlikely that it would generate high levels of summed activity immediately after stimulus onset. Although such a solution might possibly fit the data, it should be pointed out that allowing the summed unit criterion to drift over time adds substantial flexibility and freedom to the model. As described earlier, the summed unit criterion is also adjustable with respect to stimulus variables. The task of adjusting the summed unit criterion as a function of time and, simultaneously, as a function of stimulus variables would present a formidable and delicate challenge.

In sum, we believe that MROM or DRC might be adjusted to account for data from the signal-to-respond paradigm, but it appears to us that doing so would involve adding additional and fairly complex processes.

Parallel Distributed Processing Models

In both MROM and DRC, as well as in REM-LD, each word is represented by an explicit structure such as a node or a feature vector. In contrast, Parallel Distributed Processing (PDP) models of lexical processing do not posit such 'local' word structures. Rather, PDP models represent words by means of the entire pattern of activation over groups of simple, neuron-like units. The main attraction of PDP models is arguably their ability to learn a lexicon of words from scratch. MROM and DRC do not address the issue of learning, as they come equipped with a lexicon that is already fully developed.

One of the first PDP models for lexical processing was the Seidenberg and McClelland (1989) model (SM89). The implemented part of the model consists of three layers of neuron-like units, a grapheme input layer, a phoneme output layer, and a layer of hidden units. The hidden units transform activation arriving from the grapheme layer and pass this activation back to the orthographic layer and on to the phoneme layer (Seidenberg & McClelland, 1989, Figure 2).

The entire SM89 network is first trained on a set of words using a back-propagation algorithm (e.g., Rumelhart, Hinton, & Williams, 1986). Because during learning high-frequency (HF) words are presented to the network more often than low-frequency (LF) words, HF words have a relatively high influence on how the layer-to-layer connection weights are set. At test, a specific input activation pattern is instantiated in the grapheme input layer, and the network oscillates until it settles in to stable pattern of activity (note that for the implemented SM89 model, however, the stable pattern was deterministic and calculated in one processing cycle).

Lexical decisions in the SM89 model are based on the amount of mismatch between the orthographic input pattern and the feedback pattern from the hidden layer computed by the grapheme layer, called the orthographic error score: “Because the orthographic input is in fact presented to the subject, it seems reasonable to assume that subjects can compare this input to the internally generated feedback from the hidden units and use the result of this comparison process as the basis for judgments of familiarity” (Seidenberg & McClelland, 1989, p. 529). A criterion is then set on the orthographic error score dimension, patterns associated with an error score that is lower than the criterion being judged as familiar and classified as a word (cf. Balota & Chumbley, 1984). It is further assumed that in the event that orthographic familiarity is not sufficiently diagnostic to achieve reasonably good classification performance, phonological error scores can serve to fine-tune and adjust the

decision process. Note that in this SM89 model, the locus of the word frequency effect is also the locus of the long-term repetition priming effect, as it is in REM-LD. Also, the SM89 model predicts that priming for nonwords will drive the decision process towards the 'NONWORD' response, as observed in Experiment 2 – this occurs because the orthographic error score reflects familiarity, and previous training on a nonword will therefore obscure the differences between word stimuli and nonword stimuli.

The SM89 model was applied to frequency blocking (e.g., Glanzer & Ehrenreich, 1979), the effect of pseudohomophony, and orthographic and phonological short-term priming. However, several problems with the SM89 model for lexical decision were identified (e.g., Besner, Twilley, McCann, & Seergobin, 1990; Fera & Besner, 1992).[10] Specifically, Besner et al. (1990) argued that in circumstances that would not call for the involvement of phonological error score involvement (i.e., only orthographically regular words are used), performance based on the orthographic error score in terms of percentage correct classifications was much lower for the model than it was for participants. Also, Fera and Besner (1992) demonstrated that the orthographic error score variable as calculated by the SM89 model was not correlated with lexical decision performance of human participants.

Several PDP models were developed to address these and other criticisms (e.g., Plaut et al., 1996; Plaut, 1997). In particular, Plaut (1997) demonstrated how an adjusted version of the SM89 model could produce accurate performance that is not directly based on the distributions of orthographic error scores. The model proposed by Plaut (P97; Plaut, 1997) differed from SM89 in its representational assumptions, that is, in the manner in which the words are coded (for details see Plaut et al., 1996). In addition, and more important for the present discussion, the P97 model implements a semantic layer that receives its activation, via a layer of hidden units, from both the grapheme layer and the phoneme layer (cf. Plaut, 1997, Figure 6). In the P97 model, lexical decisions are based on the pattern of activation in

the semantic layer. A measure of information uncertainty or randomness over the units in the semantic layer is then used to guide lexical decision. That is, words tend to drive semantic units to particular values more strongly than do nonwords, and this information can therefore be used to distinguish words from nonwords. Plaut (1997) demonstrated that the P97 model was able to accomplish almost perfect classification performance.

It should be stressed here that both the P97 model and the SM89 model do not produce response latencies, but rather produce a measure of semantic uncertainty or an orthographic error score, respectively. Performance of the models can roughly be assessed by considering the overlap of the distributions of error scores for words and nonwords. One method to make the P97 and SM89 models –or at least the fully recurrent versions of those models– generate response latencies is to let the models cycle until a stable state is reached, and relating the number of cycles to response latency (cf. Grainger & Jacobs, 1996). Another possibility is to use the obtained measure of semantic uncertainty (P97) or orthographic error (SM89) to drive a continuous random walk or diffusion process (Ratcliff et al., 2001). With respect to modeling the data from the signal-to-respond paradigm, this again leads to two modeling options. First, adding noise to the SM89 model will make the orthographic error distribution generated by the presentation of a word gradually separate from the distribution generated by the presentation of a nonword as the number of cycles increases (cf. Seidenberg & McClelland, 1989, p. 527). Thus, classification performance should increase from chance performance (i.e., completely overlapping distributions) to asymptotic performance, as is usually the case in the signal-to-respond paradigm.

A second option is to scale the orthographic error (SM89) or the measure of semantic uncertainty (P97) and use it as input for a powerful model for the generation of response times such as the continuous random walk or diffusion model (e.g., Ratcliff, 1978). In the next section we discuss several ways in such a random walk model can be applied to the

signal-to-respond paradigm. Future work with PDP models, possibly along the lines suggested above, will have to show whether PDP models can provide a valid alternative to MROM or DRC with respect to lexical decision in general and the lexical decision signal-to-respond paradigm in particular.

Random Walk Models

The close conceptual relation between REM-LD and random walk models has already been mentioned in the section on how to extend REM-LD to generate response latencies in the respond-when-ready paradigm by means of the SPRT model. More specifically, REM-LD has much in common with Gordon's resonance model (Gordon, 1983) and Stone and Van Orden's canonical random walk model (Stone & Van Orden, 1993; other random walk models for lexical decision were recently proposed by Joordens & Becker, 1997, and Joordens et al., 2000). Recently, Ratcliff et al. (in press) provided the first quantitative fits for this type of model for lexical decision.

In the canonical random walk model, that is "almost identical"(Stone & Van Orden, 1993, p. 765) to Gordon's resonance model, information accumulates over time. Incoming information can either support the 'WORD' response or support the 'NONWORD' response, and a decision is made when the difference in the amount of supportive evidence for the two response options reaches some criterion value.

Random walk models can be applied to the signal-to-respond paradigm in various ways. For instance, a decision can be based on the position of the random walk at the time the signal-to-respond is detected (cf. Ratcliff, 1988). The system can then go with the favored response either in a discrete all-or-none fashion (i.e., when the position of the walk is closer to the word boundary or closer to the nonword boundary respond 'WORD' or 'NONWORD', respectively) or in a continuous fashion (i.e., the distance of the position from the neutral point corresponds to a continuous response probability).

An additional issue is what happens to a random walk that has reached one of the response boundaries before the signal-to-respond is detected. In such a case one can either assume that when a boundary is reached the information-accumulating process halts completely, or one can assume that information-accumulation continues until the signal-to-respond is detected.

A quite different approach is to assume that the system has no access to partial information that precedes a decision (e.g., Anderson & Lebiere, 1998; De Jong, 1991; but see Meyer, Irwin, Osman, & Kounios, 1988). This implies that when the signal-to-respond is detected when no boundary has yet been reached the system would make a random guess (e.g., Wagenmakers, Zeelenberg, Schooler, & Raaijmakers, 2000).

Thus, as in REM-LD, random walk models base their decision on an evaluation of both positive lexical information (i.e., supporting the 'WORD' response) and negative lexical information (i.e., supporting the 'NONWORD' response). With respect to underlying representational assumptions we believe the REM-LD model to be potentially more informative than random walk models – random walk or diffusion models often make no representational assumptions at all. Future work involving the SPRT model would hope to obtain the descriptive power of random walk models without sacrificing the representational assumptions inherent in the REM framework.

Conclusions

We have shown that the global memory model REM, previously applied to recognition memory (Diller et al., 2001; Nobel & Shiffrin, 2001; Shiffrin & Steyvers, 1997), recall (Diller et al., 2001; Malmberg & Shiffrin, in press; Nobel & Shiffrin, 2001), long-term priming in perceptual identification (Schooler, Shiffrin, & Raaijmakers, 2001), and short-term priming in perceptual identification (Huber, Shiffrin, Lyle, & Ruys, 2001), can be extended in a straightforward fashion to account for several key phenomena in a signal-to-

respond lexical decision paradigm. Our simulations show how the new model, REM-LD, qualitatively accounts for the time-course of effects for word frequency, nonword lexicality, repetition priming, the interaction of word frequency with both repetition priming and nonword lexicality, the effect of word frequency on nonword classification (cf. Experiment 1), and the decrease in classification performance for repeated nonwords (cf. Experiment 2).

The optimality-constraint as incorporated in the REM models has been shown to provide a very useful theoretically motivated perspective on performance in a number of different memory tasks (for a 'biologically plausible' interpretation of this optimality-constraint see Gold & Shadlen, 2001). Our ultimate goal is to construct a principled model that is able to explain various phenomena in different memory/perceptual tasks (for an overview see Shiffrin, in press). We believe that the recent developments of the REM model, particularly including the present application to lexical decision, constitute a promising step toward a fairly comprehensive understanding of human memory.

References

Anderson, J. R., & Lebiere, C (1998). The atomic components of thought. Lawrence-Erlbaum Associates.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. Psychonomic Bulletin & Review, 4, 439-461.

Antos, S. J. (1979). Processing facilitation in a lexical decision task. Journal of Experimental Psychology: Human perception and performance, 5, 527-545.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). The CELEX lexical database (CD-ROM), Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. Journal of Experimental Psychology: Human Perception and Performance, 10, 340-357.

Balota, D. A., & Spieler, D. H. (1998). The utility of item-level analyses in model evaluation: A reply to Seidenberg and Plaut. Psychological Science, 9, 238-240.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. Journal of Experimental Psychology: General, 128, 32-55.

Besner, D., Twilley, L., McCann, R. S., & Seergobin, K. (1990). On the connection between connectionism and data: Are a few words necessary? Psychological Review, 97, 432-446.

Bowers, J. S. (2000). In defense of abstractionist theories of repetition priming and word identification. Psychonomic Bulletin & Review, 7, 83-99.

Broadbent, D. E. (1967). Word-frequency effect and response bias. Psychological Review, 74, 1-15.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), Attention and performance VI (pp. 535-555). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. Psychological Review, 108, 204-256.

Davis, C. (1999). The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition. Unpublished doctoral dissertation, University of New South Wales, Australia.

De Jong, R. (1991). Partial information or facilitation? Different interpretations of results from speed-accuracy decomposition. Perception & Psychophysics, 50, 333-350.

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

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

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

Estes, W. K. (1950). Toward a statistical theory of learning. Psychological Review, 57, 94-107.

Fera, P., & Besner, D. (1992). The process of lexical decision: More words about a parallel distributed processing model. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 749-764.

Frost, R. (1998). Toward a strong phonological theory of visual word recognition: True issues and false trails. Psychological Review, 123, 71-99.

Glanzer, M., & Ehrenreich, S. L. (1979). Structure and search of the internal lexicon. Journal of Verbal Learning and Verbal Behavior, 18, 381-398.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. Trends in Cognitive Sciences, 5, 10-16

Gordon, B. (1983). Lexical access and lexical decision: Mechanisms of frequency sensitivity. Journal of Verbal Learning and Verbal Behavior, 22, 24-44.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. Psychological Review, 103, 518-565.

Hintzman, D. L., & Curran, T. (1997). Comparing retrieval dynamics in recognition memory and lexical decision. Journal of Experimental Psychology: General, 126, 228-247.

Hoel, P. G., Port, S. C., & Stone, C. J. (1971). Introduction to probability theory. Boston: Houghton Mifflin.

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

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition - sampling the state of the art. Journal of Experimental Psychology: Human Perception and Performance, 20, 1311-1334.

Jacobs, A. M., Rey, A., Ziegler, J. C., & Grainger, J. (1998). MROM-p: An interactive activation, multiple readout model of orthographic and phonological processes in visual word recognition. In J. Grainger and A. M. Jacobs (Eds.), Localist connectionist approaches to human cognition (pp. 147-188). Mahwah, NJ: Erlbaum.

James, C. T. (1975). The role of semantic information in lexical decisions. Journal of Experimental Psychology: Human Perception and Performance, 1, 130-136.

Joordens, S., & Becker, S. (1997). The long and short of semantic priming effects in lexical decision. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 1083-1105.

Joordens, S., Piercey, C. D., & Mohammad, C. (2000). The referent model of lexical decision: A random walk on the harmonious side. Manuscript submitted for publication.

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

Laming, D. R. J. (1973). Mathematical psychology. New York: Academic Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Logan, G. D. (1988). Toward an instance theory of automatization. Psychological Review, 95, 492-527.

Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms? Cognitive Psychology, 22, 1-35.

Malmberg, K. J., & Shiffrin, R. M. (in press). The "one-shot" context hypothesis: Effects of study time in explicit and implicit memory. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Massaro, D. W., & Cohen, M. M. (1994). Visual, orthographic, phonological, and lexical influences in reading. Journal of Experimental Psychology: Human Perception and Performance, 20, 1107-1128.

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

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review, 88, 375-407.

Meyer, D. E., Irwin, D. E., Osman, A. M., & Kounios, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. Psychological Review, 95, 183-237

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

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

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. Psychonomic Bulletin & Review, 4, 79-95.

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

Paap, K. R., Johansen, L. S., Chun, E., & Vonnahme, P. (2000). Neighborhood frequency does affect performance in the Reicher task: Encoding or decision? Journal of Experimental Psychology: Human Perception and Performance, 26, 1691-1720.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. Language and Cognitive Processes, 12, 1-19.

Plaut, D.C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading : Computational principles in quasi-regular domains. Psychological Review, 103, 56-115.

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85, 59-108.

Ratcliff, R. (1988). Continuous versus discrete information processing: Modeling accumulation of partial information. Psychological Review, 95, 238-255.

Ratcliff, R., Gomez, P., & McKoon, G. (in press). A diffusion model account of the lexical decision task. Psychological Review.

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, 107, 358-367.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. (1971). Evidence for phonemic recoding in visual word recognition. Journal of Verbal Learning and Verbal Behavior, 10, 645-657.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part II. The contextual enhancement effect and some tests and extensions of the model. Psychological Review, 89, 60-94.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. Journal of Experimental Psychology: Human Perception and Performance, 3, 1-17.

Scarborough, D. L., Gerard, L., & Cortese, C. (1984). Independence of lexical access in bilingual word recognition. Journal of Verbal Learning and Verbal Behavior, 23, 84-99.

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

Seidenberg, M. S., Zevin, J. D., & Harm, M. W. (2002). DRC doesn't read correctly. Paper presented at the 43rd annual meeting of the Psychonomic Society, Kansas City, Missouri.

Shiffrin, R. M. (in press). Bayesian modeling of perception and memory. Cognitive Science.

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

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

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. Psychological Review, 96, 523-568.

Seidenberg, M. S., & Plaut, D. C. (1998). Evaluating word-reading models at the item level: Matching the grain of theory and data. Psychological Science, 9, 234-237.

Smith, M. E., & Oscar-Berman, M. (1990). Repetition priming of words and nonwords in divided attention and in amnesia. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 1033-1042.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. Psychological Science, 8, 411-416.

Stone, G. O., & Van Orden, G. C. (1989). Are words represented by nodes? Memory & Cognition, 17, 511-524.

Stone, G. O., & Van Orden, G. C. (1993). Strategic control of processing in word recognition. Journal of Experimental Psychology: Human Perception and Performance, 19, 744-774.

Stone, G. O., Vanhoy, M. D., & Van Orden, G. C. (1997). Perception is a two-way street: Feedforward and feedback phonology. Journal of Memory and Language, 36, 337-359.

Stone, M. (1960). Models for choice-reaction time. Psychometrika, 25, 251-260.

Sveshnikov, A. A. (1978). Problems in probability theory, mathematical statistics and theory of random functions. New York: Dover.

Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. Psychonomic Bulletin & Review, 2, 339-363.

Townsend, J. T., & Ashby, F. G. (1983). The stochastic modeling of elementary psychological processes. Cambridge: Cambridge University Press.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. Memory & Cognition, 15, 181-198.

Van Orden, G. C. (1991). Phonological mediation is fundamental to reading. In D. Besner & G. Humphreys (Eds.), Basic processes in reading: Visual word recognition (pp. 77-103). Hillsdale, NJ: Erlbaum.

Van Orden, G. C., Stone, G. O., Garlington, K. L., Markson, L. R., Pinnt, G. S., Simonfy, C. M., & Brichetto, T. (1992). “Assembled” phonology and reading: A case study in how theoretical perspective shapes empirical investigation. In R. Frost & L. Katz (Eds.), Orthography, phonology, morphology, and meaning (pp. 249-292). Amsterdam: North Holland.

van Rijn, H. (2001). An ACT-R model for lexical decision. Presentation for the EPOS workshop 'Computational Models of Memory', Amsterdam, The Netherlands, September 2001.

van Rijn, H., & Wagenmakers, E. J. (2001). Predicting time course effects: An ACT-R model of lexical decision. Proceedings of the XII ESCOP and XVIII BPS Cognitive Section Conference, University of Edinburgh. Edinburgh: Ecosse.

Wagenmakers, E. J., Zeelenberg, R., Steyvers, M., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). Nonword repetition in lexical decision: Evidence for two opposing processes. Manuscript submitted for publication.

Wagenmakers, E. J., Zeelenberg, R., Schooler, L. J., & Raaijmakers, J. G. W. (2000). A criterion-shift model for enhanced discriminability in perceptual identification: A note on the counter model. Psychonomic Bulletin & Review, 7, 718-726.

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

Wald, A. (1947). Sequential analysis. New York: Wiley.

Ziegler, J. C., Jacobs, A. M., & Klueppel, D. (2001). Pseudohomophone effects in lexical decision: Still a challenge for current word recognition models. Journal of Experimental Psychology: Human Perception and Performance, 27, 547-559.

Author Note

Portions of this paper were presented at the 33rd Annual Meeting of the Society for Mathematical Psychology, August 2000, Ontario, and at the 41st Annual Meeting of the Psychonomic Society, November 2000, New Orleans. René Zeelenberg was supported by a grant from the Foundation for Behavioral and Social Sciences of the Netherlands Organization for Scientific Research.

We thank Douglas Hintzman for sending us the data from Hintzman & Curran (1997, Experiment 2). We also thank Amy Criss, Beau Stephens, and Ken Malmberg for helpful comments and discussions, and we thank Charissa de Ruijter for her help in running subjects. We thank reviewers Max Coltheart, Andrew Heathcote, and Jay Holden for their constructive comments. The stimuli used in Experiment 1 and 2, and an analyses of their neighborhood characteristics can be obtained in Excel format from the internet website http://www.psych.nwu.edu/~ej/remldstimuli.xls. Correspondence concerning this article can be addressed to Eric-Jan Wagenmakers, Department of Psychology, Northwestern University, 2029 Sheridan Road 102 Swift Hall, Evanston, IL 60208, USA. E-mail may be sent to ej@northwestern.edu.

Footnotes

1 The scientific database PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) lists over 600 published papers since 1985 that have the words “lexical decision” in their abstract.

2 We believe it is difficult to pinpoint one specific mechanism that is related to word frequency. Word frequency is correlated with many variables such as concreteness, age of acquisition, feature frequency, context frequency, neighborhood density, neighbor frequency, etc. Rather than introduce a different parameter for every variable that we know is related to word frequency, we decided to use a more general approach, consistent with extant models in which word frequency manifests itself in better ‘resonance’(e.g., Gordon, 1983), or a higher ‘resting level of activation’ (e.g., McClelland & Rumelhart, 1981).

3 It should be noted that this contextual specificity account of word frequency correctly predicts an advantage for LF words over HF words in episodic recognition. In episodic recognition, subjects have to decide whether or not a probe word occurred in the context of the experiment. Since LF words generally occur in fewer different contexts than HF words, the LF words are more discriminable with respect to the experimental context than are HF words. In the REM model for episodic recognition (Shiffrin & Steyvers, 1997) subjects match the word probe to a set of episodic memory traces. Since it is assumed that episodic memory traces of LF words consist of more specific (i.e., less common or more general) features, matches for LF words tend to provide more diagnostic information (i.e., higher likelihood ratios).

4 Due to a programming error, some nonwords that were created by changing two letters from a 'parent' word only differed by one letter from yet another word. Despite this inaccuracy, the data showed substantial differences between the two types of nonwords.

5 The difficulty of the signal-to-respond procedure is also witnessed by the fact that Hintzman and Curran (1997, Experiment 2) had to exclude 6 out of their initial 25 participants, either because of low accuracy or because of bad timing.

6 Experiment 3 from Wagenmakers et al. (2001) used the same stimulus materials, but adopted a slightly different 'signal-to-respond' procedure (i.e., subjects were required to respond at an imaginary tone, the 'occurrence' of which was indicated by a rhythmic sequence of three prior tones). Also, Wagenmakers et al. (2001) used different deadlines than those used in the present study.

7 Previous REM-LD simulations were done using the assumption that all of the similar lexical/semantic traces were slightly more accessible after the first presentation of a nonword. These simulations yielded similar results to those reported here.

8 Explicitly modeling the process by which phonology is computed from orthography is a complicated task (e.g., for details see Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). For the exemplary simulations presented here, we take this process as a given as its details are beyond the aim of this illustration.

9 Quantitative models for lexical decision that have not yet been published include the SOLAR model (Davis, 1999), and an ACT-R model (van Rijn, 2001; van Rijn & Wagenmakers, 2001).

10 For the naming task, Spieler and Balota (1997) showed that the correspondence between performance of the SM89 model and the empirical data was high at the factor level but broke down at the level of individual items (see also Balota & Spieler, 1998; Seidenberg & Plaut, 1998)

Table 1

An Example of the Feature-Comparison Process and the Bayesian Decision Process. See Text for Details.

Stage	Probe	Trace 1	Trace 2
Representation	[1 3 1 4]	[2 4 1 5]	[1 3 4 4]
Retrieval		1 match	3 matches
Decision			
Likelihood		2/27 ≈ .074	8/3 ≈ 2.67
Odds Ratio	37/27 ≈ 1.37		

Table 2

Mean Response Times (in Milliseconds) in Experiment 1 as a Function of Target Word Status and Deadline.

	Deadline					
Target	75	200	250	300	350	1000
HF	370	453	490	532	572	1207
LF	377	464	501	538	581	1205
NW (HF)	373	466	502	542	583	1208
NW (LF)	372	462	507	550	592	1172

Note. Response times are from stimulus onset, independent of response accuracy. HF: high frequency words, LF: low frequency words, NW (HF): nonwords presented in one block with only HF words, NW (LF): nonwords presented in one block with only LF words.

Table 3

Mean Response Times (in Milliseconds) for First Presentations and Second Presentations (After the Comma) in Experiment 2 as a Function of Target Word Status and Deadline.

	Deadline (ms)					
Target	75	200	250	300	350	1000
HF	361, 360	433, 422	465, 460	504, 501	553, 551	1196, 1196
LF	362, 362	442, 431	483, 476	529, 517	568, 561	1199, 1197
NW1	363, 361	447, 440	487, 484	529, 527	574, 571	1202, 1203
NW2	361, 360	443, 446	482, 488	521, 524	562, 563	1200, 1197

Note. Response times are from stimulus onset, independent of response accuracy.

HF: high frequency words, LF: low frequency words, NW1: ‘one letter replaced’ nonwords, NW2: ‘two letters replaced’ nonwords.

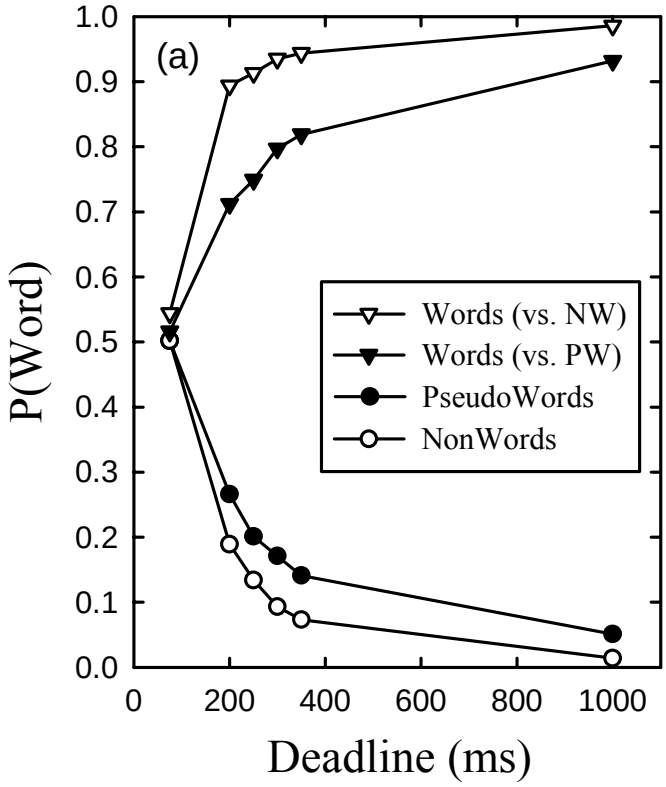
Figure Captions

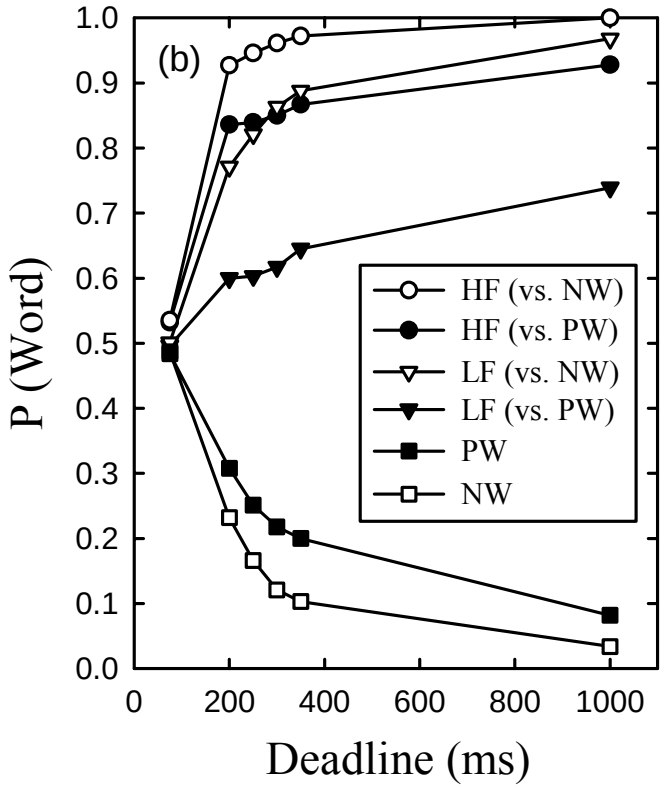
Figure 1. (a) The predicted effect of nonword lexicality in the REM-LD model. Performance for words is worse when the nonword stimuli are word-like then when the nonword stimuli are not word-like. P(Word): probability of responding 'WORD', PW: pseudowords (i.e., word-like nonwords), NW: nonwords (i.e., less word-like nonwords). (b) The predicted effect of nonword lexicality with respect to the word frequency effect in the REM-LD model. The word frequency effect is larger when the nonword stimuli are word-like than when they are not. HF: high frequency words, LF: low frequency words. See text for details.

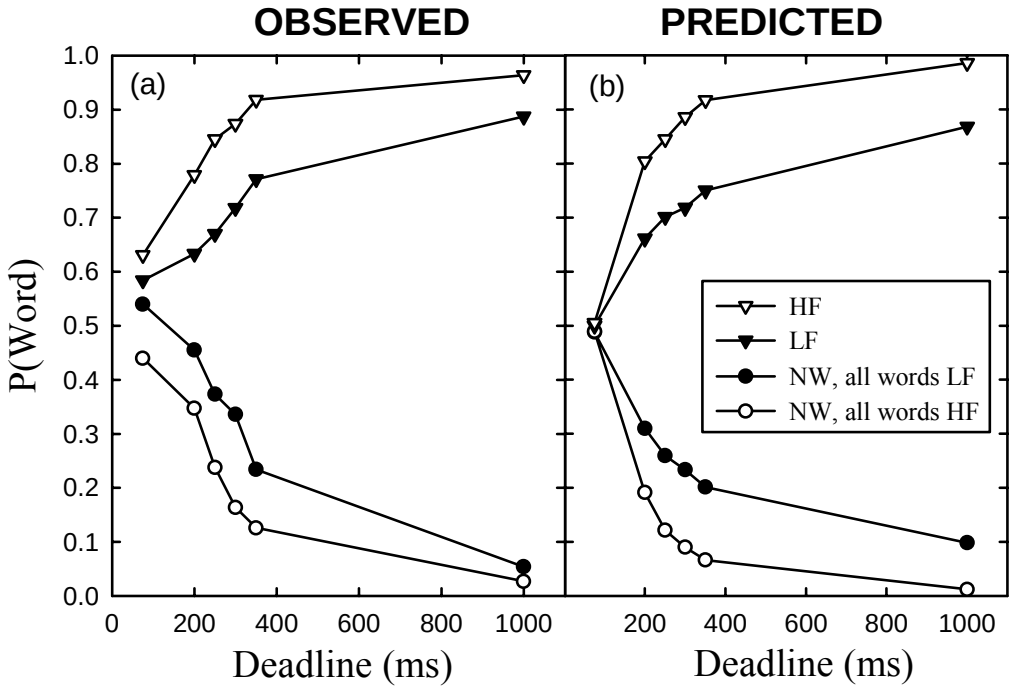
Figure 2. (a) Results from Experiment 1. Nonwords are responded to more accurately when presented in one block with only high-frequency words than with only low-frequency words. P(Word): probability of responding 'WORD', HF: high frequency words, LF: low frequency words, NW: nonwords. (b) The predicted effect of word frequency on performance for nonwords in the REM-LD model. Performance for nonwords is better when they have to be distinguished from HF words than when they have to be distinguished from LF words.

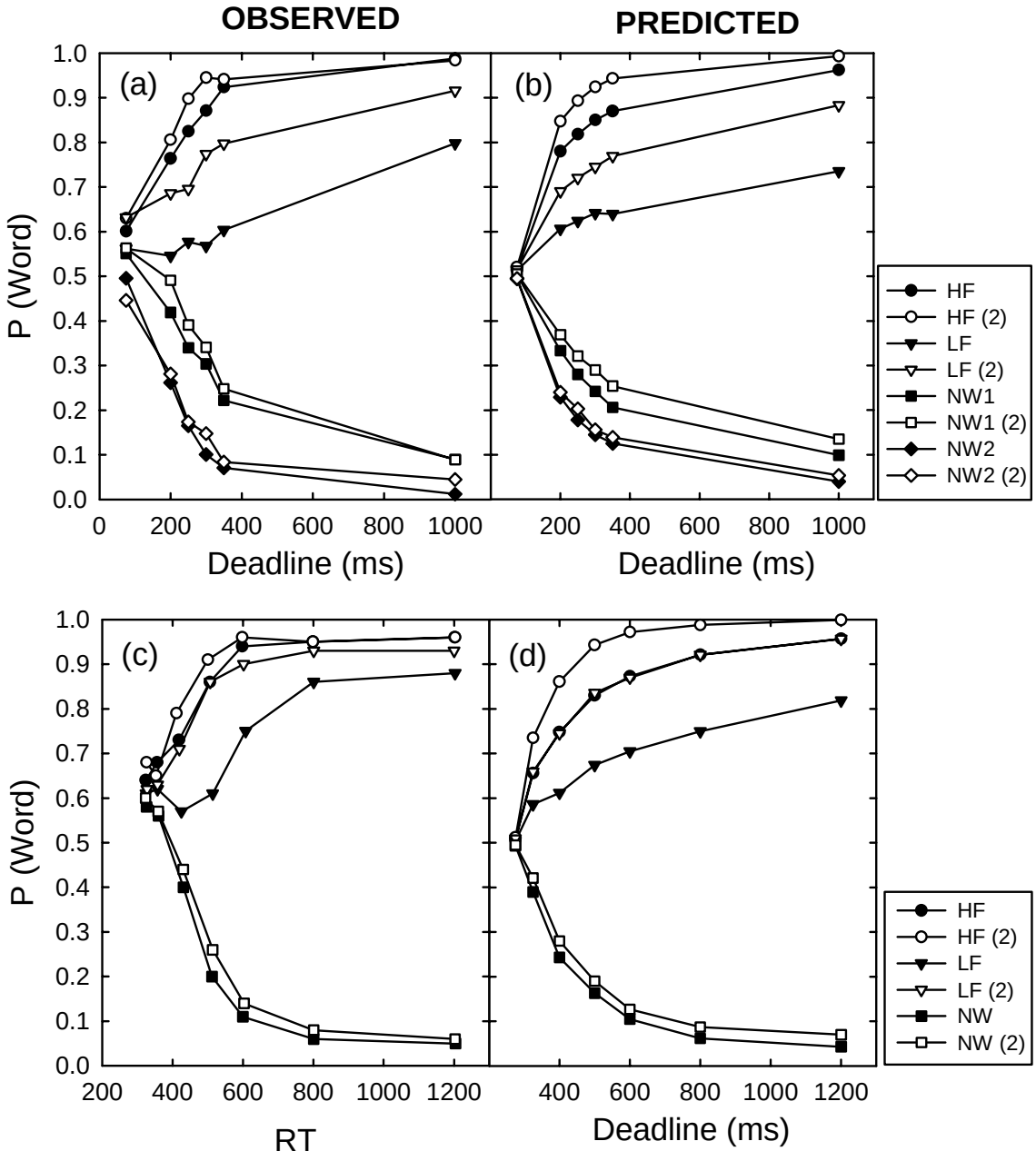
Figure 3. (a) Results from Experiment 2. Repeated stimuli are more likely than novel stimuli to be classified as a word. P(Word): probability of responding 'WORD', HF: high frequency words, LF: low frequency words, NW1: 'one letter replaced' nonwords, NW2: 'two-letters replaced' nonwords. The digit 2 in brackets indicates the second presentation. (b) Predictions of the REM-LD model for the conditions from Experiment 2. (c) Re-plotted data from Hintzman & Curran (1997, Experiment 2, Figure 9). As is apparent from the figure, prior exposure increases the probability of classifying a stimulus as a word, for all stimulus categories. (d) Predictions of the REM-LD model for the conditions from Hintzman and Curran (1997; Experiment 2).

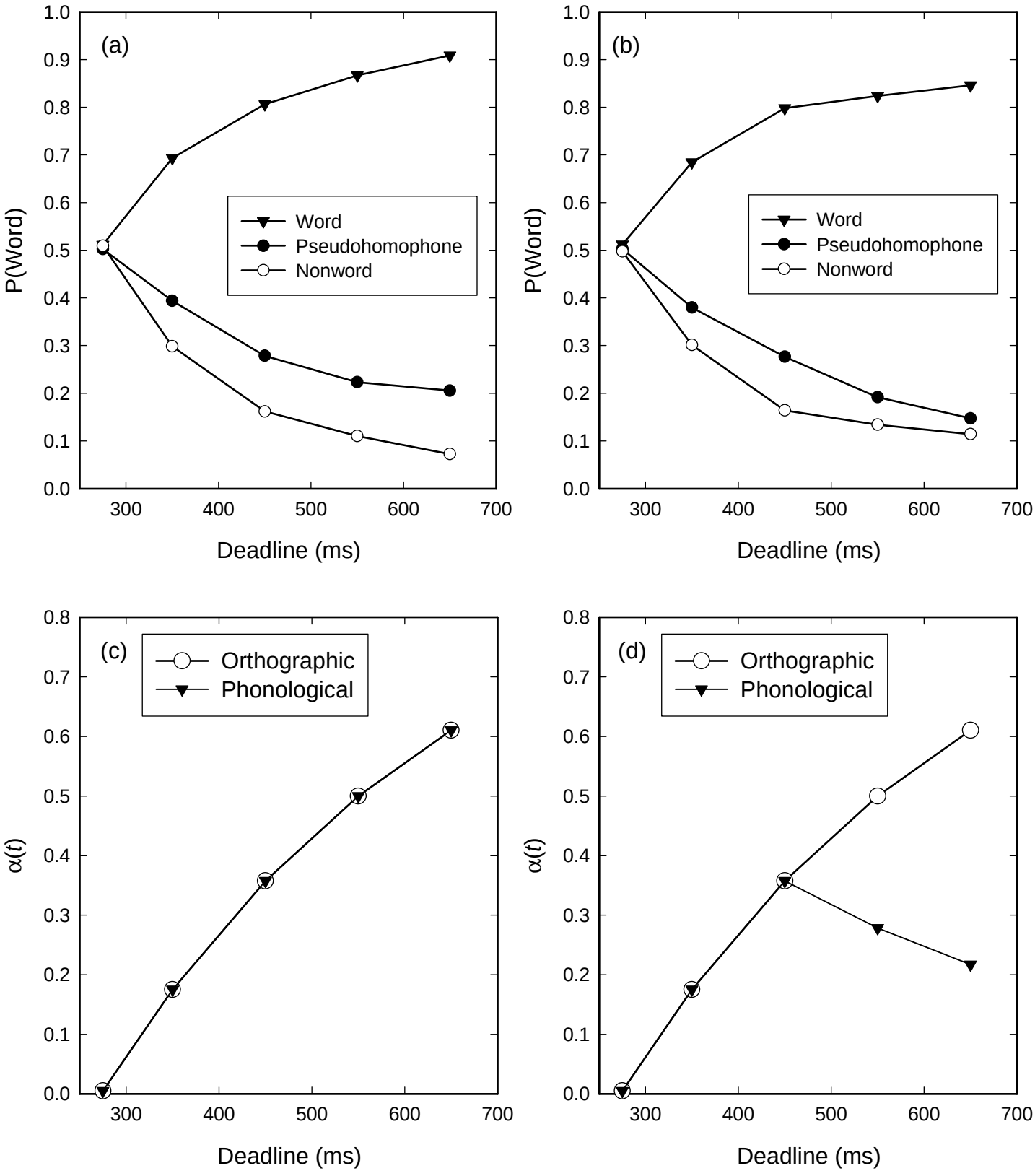
Figure 4. The pseudohomophone effect simulated by REM-LD. P(Word): probability of responding 'WORD'. (a) Predicted increase in accuracy with processing time for words, regular nonwords, and pseudohomophones with the same time course for activating orthographic and phonological features (see panel c). (b) Predicted increase in accuracy with processing time for words, regular nonwords, and pseudohomophones with a different time course for activating orthographic and phonological features (see panel d). (c) An identical time course for activating orthographic and phonological features in REM-LD (for results see panel a). (d) A different time course for activating orthographic and phonological features in REM-LD (for results see panel b).


(a)
P(Word)
Deadline (ms)
Words (vs. NW)
Words (vs. PW)
PseudoWords
NonWords
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
200
400
600
800
1000


(b)
P (Word)
Deadline (ms)
HF (vs. NW)
HF (vs. PW)
LF (vs. NW)
LF (vs. PW)
PW
NW
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
200
400
600
800
1000


OBSERVED
PREDICTED
(a)
(b)
P(Word)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
200
400
600
800
1000
Deadline (ms)
HF
LF
NW, all words LF
NW, all words HF


OBSERVED
PREDICTED
(a)
(b)
(c)
(d)
P (Word)
Deadline (ms)
RT
HF
HF (2)
LF
LF (2)
NW1
NW1 (2)
NW2
NW2 (2)
NW
NW (2)


(a)
P(Word)
Deadline (ms)
Word
Pseudohomophone
Nonword
(b)
P(Word)
Deadline (ms)
Word
Pseudohomophone
Nonword
(c)
α(t)
Deadline (ms)
Orthographic
Phonological
(d)
α(t)
Deadline (ms)
Orthographic
Phonological