# A Bayesian Model for the Time Course of Lexical Processing

**Mark Steyvers**
**(msteyver@psych.stanford.edu)**
Department of Psychology, Stanford University
Stanford, CA 94305-2130

**Eric-Jan Wagenmakers**
**(pn_wagenmakers@macmail.psy.uva.nl)**
Department of Psychology, University of Amsterdam
Amsterdam, The Netherlands.

**Richard Shiffrin**
**(shiffrin@indiana.edu)**
Department of Psychology
Indiana University, Bloomington, IN 47405-7007

**René Zeelenberg**
**(pn_zeelenberg@macmail.psy.uva.nl)**

**Jeroen Raaijmakers**
**(raaijmakers@psy.uva.nl)**
Department of Psychology, University of Amsterdam
Amsterdam, The Netherlands.

## Abstract

A Bayesian-based model for lexical decision, REM-LD, is fit to data from a novel version of a signal-to-respond paradigm. REM-LD calculates the odds that a test item is a word, by accumulating likelihood ratios for each lexical entry in a small neighborhood of similar words. The new model predicts the time course of observed effects of nonword lexicality, word frequency and repetition priming.

## Introduction

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 and FUME) from nonwords (e.g., GREACH and ANSU). In tasks in which accuracy is near ceiling, three critical findings are seen in the response times: (1) The *word frequency effect*. Words that occur regularly 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). (2) The *repetition priming effect*. Prior exposure to a word leads to faster correct classifications for that word on a second presentation. This increase in performance is particularly pronounced for LF words (e.g., FUME benefits more from prior exposure than CHAIR). (3) The *nonword lexicality effect*. Nonwords that look like words (e.g. GREACH) take longer to be classified correctly than nonwords that are relatively dissimilar to words (e.g. ANSU). In this article[1], we use a new variant of a signal-to-respond procedure that produces findings in the accuracy domain that mimic those listed above for response times. We will fit a new Bayesian model, REM-LD, to the data. The advantage of the signal-to-respond technique is that it allows one to track the time course of processing, obtaining multiple data points for each stimulus category while reducing concerns about higher order task strategies and speed-accuracy trade-off's.

## Experimental Data

The signal-to-respond paradigm has occasionally been applied to lexical decision (Antos, 1979; Hintzman & Curran, 1997). We used our new version of the signal-to-respond paradigm to replicate and extend Experiment 2 from Hintzman and Curran (1997).

### Method

We used four types of stimuli: (1) 168 HF words, each occurring more than 30 times per million according to the CELEX lexical database (Burnage, 1998) (2) 168 LF words, each occurring 1 or 2 times per million (3) 168 pronounceable nonwords created by replacing one letter of an existing word (e.g., GREACH created from PREACH) (4) 168 pronounceable nonwords differing by at least two letters from any word (e.g., ANSU; this condition was absent in the Hintzman and Curran study). The first three stimulus categories were matched on neighborhood structure (i.e., a neighbor is a word differing from another word in one letter, so TIED is a neighbor of LIED); These categories had the same summed logarithmic word frequency of the neighbors. Stimuli were presented twice to study how prior exposure affects performance. To control for practice effects and shifts in response criteria, we presented stimuli in blocks of 48 trials, half of which were stimuli that were encountered in the previous block, half of which were new. Each block contained 24 words and 24 nonwords. Subjects were required to respond at six different lags: 350, 400, 450, 500, 550, and 600 ms. The appropriate lag was indicated to the subject by means of three tones (see Figure 1a). The tones were equidistant in time, and the onset of the third and last tone coincided with the onset of the stimulus. The subject had to respond at the fourth *imaginary* tone. We adopted this procedure in the hope that it would produce less interference than the presentation of a tone

[1] More details and related research can be found in Wagenmakers (2001).

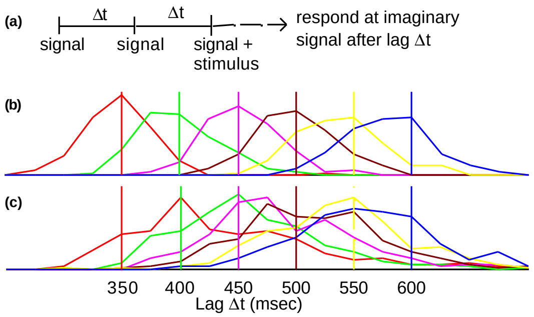

Figure 1. (a) the signal-to-respond procedure. (b) response time distributions for a participant with good timing and (c) for a participant with bad timing. Matching line-colors indicate correspondence between lag (vertical line) and response distribution.

during processing. After each trial, subjects received feedback concerning the accuracy and latency of their response relative to the desired latency.

## Results and Discussion

Forty-three students at Indiana University participated. we excluded 14 participants from the analyses because of extremely bad performance or bad timing. Figure 1b and 1c show the distribution of response times for a subject with good and bad timing, respectively. All response latencies were grouped into six bins for each subjects separately, the first bin containing the 16.7% slowest responses, and so forth. Next, the accuracy data from each bin were averaged over subjects. Other analyses such as binning by actual response latency or analyzing accuracy data by lag yielded similar results. The results can be seen in Figure 2. Performance for HF words is better than for LF words, and performance for nonwords that differ from any word in two letters (i.e., NW2) is better than for nonwords that differ from a word in one letter (i.e., NW1). Repeated stimuli (indicated by open symbols) are more likely to be classified as 'word' than new stimuli (indicated by the filled symbols), an effect larger for LF words than for HF words. As expected, performance increases dramatically with processing time, except perhaps for new LF words. This lack of increase could either be due to a very slow retrieval process for LF words, or to the possibility that some subjects might be uncertain concerning the lexicality of some LF words. One might argue that the gain in performance for repeated LF words reflects a retrieval of the feedback given on the earlier presentation ('I remember this stimulus is supposed to be a word'). However such a memory process would lead to improved performance for repeated nonwords ('I remember this stimulus is a nonword'), whereas the data show a *decrease* in performance for repeated nonwords. The hypothesis that repetition priming involves two distinct processes (i.e., familiarity and recollection) will be elaborated upon in the Discussion. Overall, the data consistently show effects of processing time, nonword lexicality, word frequency and repetition priming.

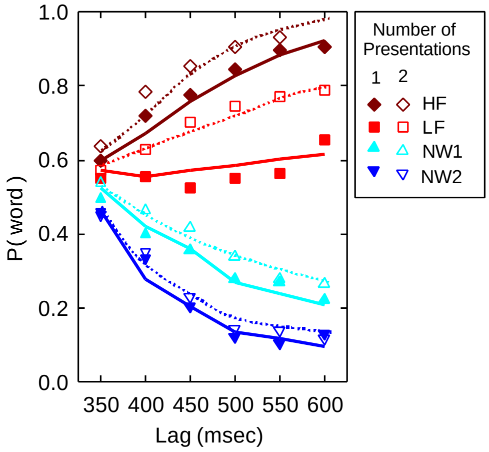

Figure 2. Results of the signal-to-respond lexical decision experiment. The observed data is indicated by the symbols while the REM-LD model fit is indicated by the solid lines (stimuli presented once) and dashed lines (stimuli presented twice).

## The REM-LD Model

The REM-LD model is similar to the REM model for episodic recognition (Shiffrin & Steyvers, 1997). In episodic recognition, participants have to distinguish 'old' words (i.e., words that were presented in a previous study list) from 'new' words. The REM-LD model is an application of the REM model to the lexical decision task. In the REM-LD model, we will make the following assumptions.

(1) Words and nonwords can be represented by vectors of feature values. We assume that these features arbitrarily represent attributes such as orthography and phonology. Here we represent each word by a collection of 30 features with values 1 to 10 randomly drawn from a uniform distribution.

(2) Words have lexical entries (i.e., representations) in memory whereas nonwords do not. The presentation of the probe (i.e., the stimulus) leads to activation of $n$ lexical entries that are orthographically *similar* to the probe (see Figure 3a). In a more complete model the value of $n$ would probably be smaller for tests of dissimilar nonwords (i.e., nonwords that differ in two letters from any word), but for simplicity we set $n$=10 for all test items and instead vary the feature similarity for dissimilar nonwords. In case the probe is a word, one of the activated lexical entries is the probe (denoted s-entry for 'same', e.g. BEG in Figure 3a). The other activated entries are similar but different from the probe (denoted d-entries for 'different'). Note that a nonword

can only activate lexical entries that are *similar* to it, since nonwords do not have lexical representations.

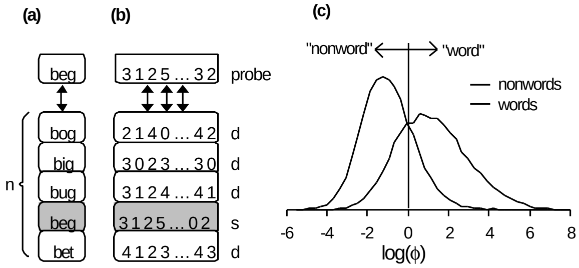

Figure 3. (a) a presented letter string activates *n* orthographically similar lexical entries. (b) letter strings and lexical traces are represented by vectors. (c) the distribution of log(ϕ) for two conditions.

The degree of similarity between the probe and the entries is determined by the probability that the same feature value is present in probe features as in features from the lexical entries. Stored s-entry features will match the probe features with probability $b_1$; stored d-entry features will match the probe features with probability $b_2$, where $b_2 < b_1$. With probability 1-*b*, feature values in the entries can differ from those in the probe. These feature values are obtained by sampling randomly from the uniform feature distribution, enabling matches to occur by chance. In the experiment, the similarity between the probe and the lexical neighbors was equated for NW1, LF and HF stimuli. However, the NW2 stimuli were more dissimilar from their lexical neighbors because they differed from any word in at least two letters. Therefore, we set $b_2(NW2) < b_2(NW1) = b_2(LF) = b_2(HF)$.

In REM-LD we assume that word frequency is represented by parameter $b_1$; the probability of a probe feature matching a trace feature is assumed to be higher for HF word probes than for LF word probes. The increased matching probability for HF words might involve various mechanisms such better matching context features because HF words occur in many different contexts. Alternatively, features from HF traces might be retrieved at a faster rate than features from LF traces. Therefore, we set $b_1(LF) < b_1(HF)$. In sum, the $b_1$ parameters determine the similarity or degree of overlap between a word probe and its corresponding lexical entry whereas the $b_2$ parameters determine the amount of overlap between the probe and lexical entries that are most similar to it.

(3) Mistakes in this task are made because the comparison process of the probe to the activated lexical entries is noisy. The function *a(t)* gives the probability that a feature from a lexical entry is retrieved correctly at time *t*, in order to be compared with a probe feature (if it is not retrieved correctly at time *t*, the value retrieved is obtained by sampling from the uniform feature distribution so that it can still match by chance). To account for the improvement in the lexical decision task as a result of more processing time, we assume that the function *a(t)* is monotonically increasing over time according to:

$$a(t) = 1 - e^{-b(t-t_0)}$$

where *b* and $t_0$ are parameters of rate and starting point.

(4) The decision to respond ‘word’ or ‘nonword’ is based on an assessment of the evidence that the activated set of lexical entries contains an s-entry. An optimal decision is based on the odds *f* that the probe is a word rather than a nonword, given the available data D: *f* = *P*(*w*|*D*) / *P*(*nw*|*D*) where the data consists of the number of matches and mismatches between the probe features and the features of all activated lexical entries. Figure 3c illustrates typical log *f* distributions generated by word and nonword probes, also illustrating the fact that log *f* = 0 is the optimal response criterion. By Bayes’ rule, *f* = [ *P*(*D*|*w*)*P*(*w*) ] / [ *P*(*D*|*nw*)*P*(*nw*) ]. Because in our experiment the prior probability of the probe being a word, *P*(*w*), equaled the probability of the probe being a nonword, *P*(*nw*), we have:

$$j = \frac{P(D \mid w)}{P(D \mid nw)}$$

When the probe is a word, there is an equal probability that any activated lexical entry is a s-entry. This can be used in a simple derivation (Shiffrin & Steyvers, 1997) that leads to:

$$\frac{P(D \mid w)}{P(D \mid nw)} = \frac{1}{n} \sum_{j=1}^{n} \frac{P(m_j \mid s_j)}{P(m_j \mid d_j)}$$

where $m_j$ is the number of matching features in the comparison of the lexical entry *j* to the probe. The terms $\mathrm{s_j}$ and $\mathrm{d_j}$ represent the assumptions that the lexical entry *j* is a s-entry and d-entry respectively. Therefore, the odds for *word* is an average of the likelihood ratios for each of the lexical entries in the activated set. We can calculate each likelihood ratio in the following way:

$$\frac{P(m_j \mid s_j)}{P(m_j \mid d_j)} = \left( \frac{a(t)\hat{b}_1 + \left(1 - a(t)\hat{b}_1\right) {}^{1}\!/_{v}}{a(t)\hat{b}_2 + \left(1 - a(t)\hat{b}_2\right) {}^{1}\!/_{v}} \right)^{m_j} \left( \frac{1 - a(t)\hat{b}_1}{1 - a(t)\hat{b}_2} \right)^{k - m_j}$$

In this equation, *v* is the number of distinct values from the uniform feature distribution (always 10 here) and *k* is the number of features (always 30 here). The calculations of the system are partly based on the estimates $\hat{b}_1$ and $\hat{b}_2$. These system estimates are based on an arithmetic average of the different values that $b_1$ and $b_2$ can take on in the different experimental conditions. The calculations of the system also depend on an estimate of $a(t)$, the time course of retrieving features from the entries. In other words, the system weighs the diagnosticity of the evidence with processing time. We are currently exploring alternative models that incorporate different assumptions.

(5) Prior exposure to a word primes the corresponding lexical entry. Therefore, the features of a repeated word probe will better match the features in the corresponding lexical entry. We model this by assuming that $b_1$ is increased by a small amount *Db* for repeated word probes. Similarly, prior exposure to a nonword primes the lexical entry of the word that is most similar to it. Therefore, the second occurrence of the nonword string will lead to more matching features in the comparison of the repeated nonword probe and the most similar lexical neighbor. We model this by increasing $b_2$ for one lexical entry by the amount *Db* for repeated nonword probes.

### Simulation results

Figure 2 shows the results of a quantitative model fit of the REM-LD model to the observed data involving seven free parameters. The mean squared error (MSE) of the fit is 7.74e-004. The values of the seven parameters values found to produce the predictions were: $b_1(LF)$=.674, $b_1(HF)$=.832, $b_2(NW1)$=.398, $b_2(NW2)$=.359, *Db*=.079, $t_0$=330, *b*=0.0051. The qualitative predictions were found to be relatively robust against variations in these parameter values. Because accuracy for HF words is higher than for LF words, $b_1(HF)$ was set higher than $b_1(LF)$ so that lexical probes would match their lexical entries better for HF words than LF words. Because NW1 nonwords are more often mistakenly judged to be words than NW2 nonwords, $b_2(NW1)$ was set higher than $b_2(NW2)$ so that a NW1 probe would activate its similar lexical neighbors to a greater extend than a NW2 probe. For both word and nonword conditions, probes that are repeated are classified as 'word' more often then probes encountered for the first time. The model predicts this because a repeated word probe primes the corresponding lexical entry while a repeated nonword probe leads primes the lexical entry of the word that it is most similar to. The model also predicts that the repetition priming effect is more pronounced for LF words than for HF words. This is because the average value of log $f$ is closer to zero for the LF words than for the HF words. Hence, an identical increase in log $f$ due to activation of the episodic trace will have a greater impact on performance for LF words than for HF words.

## Discussion

We have shown that a Bayesian-based model, REM-LD, can predict lexical decision effects such as word frequency, repetition priming, and nonword lexicality. This model takes into account the similarity of nonwords to words, thereby keeping the system 'centered' around the optimal criterion of log $f$ of zero. REM-LD can also handle the observed improvement in performance with processing time. In contrast to most extant models and empirical work in lexical decision, we focused on changes in accuracy over time, as seen in a variant of a signal-to-respond procedure. A Bayesian model is particularly suited toward explaining data from the signal-to-respond paradigm, since the system bases it decisions on the diagnosticity of the evidence, simultaneously considering the evidence for and against the 'word' response. When, early in processing, the evidence is noisy and supports neither the 'word' response nor the 'nonword' response, performance is at chance. Empirically, the most interesting finding is the decrease in performance for repeated nonwords. The current model assumes prior exposure of a nonword primes the most similar activated lexical entry, predicting the observed decrement in performance. However, with subject-paced responding, an *increase* in performance for repeated nonwords is sometimes observed (Logan, 1988). Therefore, it is possible that the net result of repetition priming for nonwords is the sum of two opposing effects: (1) An *implicit priming* component such as modeled by REM-LD, leading subjects to give the *erroneous* 'word' response, and (2) A *recollection* component that leads subjects to remember the *correct* 'nonword' response. This recollection process might be operative when subjects are under less pressure to give speeded responses, such as in experiments in which responding is subject-paced (e.g., Wagenmakers, 2001).

## References

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

Burnage, D. (1998). *CELEX: a guide for users*. Nijmegen: Centre for Lexical Information.

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

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

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

Wagenmakers, E.J. (2001). *Priming in Visual Word Recognition: Empirical Studies and Computational*

*Models*. Unpublished doctoral dissertation.