Published in final edited form as:
*Psychol Rev*. 2004 January ; 111(1): 159–182.

# A Diffusion Model Account of the Lexical Decision Task

**Roger Ratcliff**,
*The Ohio State University*

**Pablo Gomez**, and
*De Paul University*

**Gail McKoon**
*The Ohio State University*

## Abstract

The diffusion model for 2-choice decisions (R. Ratcliff, 1978) was applied to data from lexical decision experiments in which word frequency, proportion of high- versus low-frequency words, and type of nonword were manipulated. The model gave a good account of all of the dependent variables —accuracy, correct and error response times, and their distributions—and provided a description of how the component processes involved in the lexical decision task were affected by experimental variables. All of the variables investigated affected the rate at which information was accumulated from the stimuli—called *drift rate* in the model. The different drift rates observed for the various classes of stimuli can all be explained by a 2-dimensional signal-detection representation of stimulus information. The authors discuss how this representation and the diffusion model's decision process might be integrated with current models of lexical access.

The lexical decision task is one of the most widely used paradigms in psychology. The goal of the research described in this article was to account for lexical decision performance with the diffusion model (Ratcliff, 1978), a model that allows components of cognitive processing to be examined in two-choice decision tasks. Nine lexical decision experiments, manipulating a number of factors known to affect lexical decision performance, are presented. The diffusion model gives good fits to the data from all of the experiments, including mean response times for correct and error responses, the relative speeds of correct and error responses, the distributions of response times, and accuracy rates.

In the diffusion model, the mechanism underlying two-choice decisions is the accumulation of noisy information from a stimulus over time. Information accumulates toward one or the other of two decision criteria until one of the criteria is reached; then the response associated with that criterion is initiated. In the lexical decision task, one of the criteria is associated with a word response, the other with a nonword response. The rate with which information is accumulated is called *drift rate*, and it depends on the quality of information from the stimulus. In lexical decision, some stimuli are more wordlike than others, and so their rate of accumulation of information toward the word criterion is faster; other stimuli, such as random letter strings, are so un-wordlike that information accumulates quickly toward the nonword criterion. For the nine experiments presented below, the drift rates can be summarized quite simply. First, the ordering of the drift rates from largest to smallest is as follows: high-frequency words, low-frequency words, very low-frequency words, pseudowords, and random letter

Correspondence concerning this article should be addressed to Roger Ratcliff, Department of Psychology, The Ohio State University, 1885 Neil Avenue, Columbus, OH 43210-1222..
Roger Ratcliff and Gail McKoon, Department of Psychology, The Ohio State University; Pablo Gomez, Department of Psychology, De Paul University.

strings. Second, the differences among the drift rates are larger when the nonwords in an experiment are pseudowords than when they are random letter strings.

For our framework, Figure 1 outlines the relationships among lexical decision data, the diffusion model, and word recognition (lexical) models, and shows how the data do not map directly to lexical processes but, instead, map to lexical processes only through the mediation of the diffusion model. Data enter the diffusion model, which produces the values of drift rates for the different classes of stimuli that give the best account of the data. In this framework, the role of a word recognition model is to produce values for stimuli for how wordlike they are. We call the measure of how wordlike a stimulus is its *wordness* value (a term intended to be neutral for the purposes of this article). Wordness values map onto the drift rates that drive the diffusion decision process to produce predictions about accuracy and response time.

In our framework, wordness values place fewer constraints on word recognition models for the lexical decision task than has been appreciated. All that is required is that a model produce the appropriate ordering of wordness values: from high-frequency words to low- and very low-frequency words to pseudowords and random letter strings, with larger differences among them when the nonwords in an experiment are pseudowords than when they are random letter strings. In other words, the disturbing and simple conclusion from the diffusion model's account of lexical decision is that, beyond what can be said from a bare ordering of wordness values, the lexical decision task may have nothing to say about lexical representations or about lexical processes such as lexical access. Lexical decision data do not provide the window into the lexicon that might have been supposed in earlier research.

The framework shown in Figure 1 is counter to much previous work that has assumed lexical decision data do map directly onto lexical processes. Often, lexical decision response time (RT) has been interpreted as a direct measure of the speed with which a word can be accessed in the lexicon. For example, some researchers have argued that the well-known effect of word frequency—shorter RTs for higher frequency words—demonstrates the greater accessibility of high-frequency words (e.g., their order in a serial search, Forster, 1976; the resting levels of activation in units representing the words in a parallel processing system, Morton, 1969). However, other researchers have argued, as we do here, against a direct mapping from RT to accessibility. For example, Balota and Chumbley (1984) suggested that the effect of word frequency might be a by-product of the nature of the task itself and not a manifestation of accessibility. In the research presented here, the diffusion model makes explicit how such a by-product might come about.

The sections below begin with a detailed description of the diffusion model; then nine experiments are presented, and the model is fit to the data from each one. The main result is that the differences in performance for various classes of stimuli are all captured by drift rate, not by any of the other components of processing that make up the diffusion model.

## The Diffusion Model

According to the diffusion model, the mechanism underlying binary decisions is the accumulation of noisy information over time toward one or the other of two decision criteria, or boundaries, as in Figure 2, where the boundaries are labeled *a* and *0* and the starting point is labeled *z*. The mean rate of approach to a boundary is called the drift rate, and the variation of sample paths around the mean values in the accumulation process is described by the diffusion coefficient $s^2$ (within-trial variability). This variation allows processes with the same drift rate to reach the same boundary at different times (the two jagged lines of Panel A in Figure 2). Variability also means that a process can reach the wrong boundary by mistake, yielding an error response. Drift rate is a function of the quality of the information produced from processing of the stimulus. For example, later in this article, we show that drift rate in

lexical decision is a function of word frequency and of the type of nonwords used as negative items. Speed–accuracy trade-offs occur when the boundaries are moved farther apart to produce slower and more accurate responses or closer together to produce faster and less accurate responses. For example, if the boundaries were moved very close to the starting point in Panel A, the slower of the two processes in the figure would terminate earlier at the bottom boundary.

Positively skewed RT distributions are automatically predicted by the geometry of the model. Increasing positive skew is the empirically observed behavior of RT distributions in many tasks when difficulty of the task is increased. If the drift rates of both the fastest and slowest processes (the solid lines in Panel B of Figure 2) are reduced by the same amount $x$, the fastest responses are slowed by less than the slowest responses (the dashed lines), which produces an increase in positive skew.

The diffusion model can also explain the relative speeds of correct and error responses by allowing variability across trials in drift rate and in starting point. With a fixed drift rate and symmetric boundaries, the model predicts error RTs to be the same as correct RTs. When drift rate varies across trials, the model predicts that error responses will be slower than correct responses. Across-trial variability in drift rates was introduced by Ratcliff (1978) in applications of the model to recognition memory because nominally equivalent items (e.g., the 10th item in a list of items to be remembered) are expected to vary in strength. Similarly, in the lexical decision task, nominally equivalent items (e.g., high-frequency words) vary in familiarity across trials. How across-trial variability in drift rate allows the model to predict slower error responses than correct responses is explained by Panel C of Figure 2. In the figure, two processes, one with drift rate $v_1$ and the other with drift rate $v_2$, are averaged to show how errors slower than correct responses come about as weighted averages of finishing times and response probabilities. Processes with drift $v_1$ have mean RT of 400 ms and accuracy .95, and processes with drift $v_2$ have mean RT of 600 ms and accuracy .80. The weighted mean RT for correct responses is 491 ms. The mean RT for errors is 560 ms because the contribution from processes with drift $v_2$ is four times larger than the contribution from processes with drift $v_1$ (error rates of .20 and .05, respectively). In real implementations of the model, there is assumed to be a normal distribution of drift rates, not just the two processes used here for illustration.

Error responses faster than correct responses are predicted when the position of the starting point varies across trials (e.g., Laming, 1968). Figure 2, Panel D, shows this with the averages of two processes that represent the highest and lowest starting points from a uniform distribution of starting point values. Processes starting near the error boundary hit it with shorter RT and greater probability than processes starting near the correct boundary. The weighted sum gives faster errors than correct responses.

With the combination of across-trial variability in drift rate and across-trial variability in starting point, the diffusion model can produce the various patterns of correct versus error RTs that occur empirically: Error responses are sometimes faster than correct responses and sometimes slower, and in some experiments, there is a cross-over such that errors are slower than correct responses when accuracy is low and faster than correct responses when accuracy is high (Ratcliff & Rouder, 2000; Ratcliff, Van Zandt, & McKoon, 1999; Smith & Vickers, 1988). In the diffusion model as it is applied here (and in recent papers, e.g., Ratcliff & Tuerlinckx, 2002), drift rate is assumed to be normally distributed across trials, with standard deviation $\eta$, and starting point is assumed to be rectangularly distributed with range $s_z$ (a rectangular distribution has a maximum and minimum, and so with appropriate values of $z$, a, and $s_z$, no starting point can lie outside the response boundaries as it might if the distribution were normal).

Besides the decision process, any two-choice task includes other components of processing, such as encoding and response execution. These are summarized in the diffusion model into one parameter, the nondecision component of RT with mean $T_{er}$ (which is not shown in Figure 2). Recently, Ratcliff and Tuerlinckx (2002) investigated whether the diffusion model should include variability in the nondecision component of RT. In the course of fitting the model to a set of data using a chi-square method, Ratcliff and Smith (in press) found that variability among the shortest RTs (the .1 quantile RTs) led to poor fits of the model (with misses in the .9 quantile RT as large as 300 ms). Ratcliff and Tuerlinckx reasoned that the .1 quantile variability could be due to perceptual and encoding processes. Assigning variability to the nondecision component of processing led to good fits of the model.

Variability in the nondecision component of processing also plays an important role in fits of the model to lexical decision data. An important result obtained by Balota and Spieler (1999) was that there is a relatively large shift in the leading edge of the lexical decision RT distribution as a function of word frequency for correct responses. The leading edge of the distribution for high-frequency words is about 30 ms shorter than the leading edge for lower frequency words. This shift is larger than would be expected from the diffusion model if there were no variability in the nondecision component of processing. Without variability in the nondecision component of processing, the model can accommodate only 19-ms of the 30-ms leading edge shift (assuming parameter values similar to those for the fits of the diffusion model presented later in this article).

How variability in the nondecision component of processing contributes to leading edge shifts can be illustrated by considering just three values from a distribution of values of the nondecision component of processing: one value shorter than the mean value ($T_{er}$), another longer, and another at the mean. Suppose three sets of decision processes are averaged together, one set for each of the three values, to produce the cumulative RT distribution of the combination, and suppose this is done with all three sets of processes having the same high value of drift rate (and also holding all the other parameters of the diffusion model constant). Then, with the high value of drift rate, the cumulative RT function of the combination rises rapidly from zero, so rapidly that at its beginning, the RTs come entirely from decision processes for which the nondecision component of processing has its smallest value; there are no contributions from processes for which the nondecision component has its mean value or the long value. Only later in the function are there contributions from processes for which the nondecision component has the mean or the longer value. In contrast, the situation is different when the values of drift rates for the three averaged processes are lower. In this case, the cumulative RT function includes, from shortly after its beginning, processes for which the nondecision component has all three possible values: the smallest, the mean, and the longest. This function has none of the very short RTs that occur with the high drift rate function. In other words, the high drift rate function is shifted toward shorter times in its leading edge relative to the low drift rate function. As mentioned above, in fits of the diffusion model to data, generally about one half to two thirds of the shift in the .1 quantile RT as a function of word frequency (e.g., 19 ms in Balota & Spieler, 1999) can be accounted for with a change in drift rate without variability in the nondecision component of processing. With this source of variability, the change in drift rate accounts for an extra 11-ms shift in the .1 quantile (i.e., an extra third of the shift).

In fitting the diffusion model, variability in the nondecision component of processing is assumed, for simplicity, to have a rectangular distribution. Assuming a rectangular distribution instead of any other reasonable assumption has almost no effect on the shape of the predicted RT distributions relative to the case with no variability in the nondecision component of processing. This is because the shape of the combination of the rectangular distribution and the distribution generated from the diffusion decision process is largely determined by the

distribution with the largest standard deviation, which is the distribution from the diffusion process.

To summarize, the parameters of the diffusion model are as follows: the starting point, $z$; across-trial variability in starting point, range $s_z$; the boundary separation parameter, $a$; within-trial variability in drift, $s$ (a scaling parameter which is set to 0.1 in all fits); across-trial variability in drift, standard deviation $\eta$; a different value of mean drift rate ($v$) for each condition of an experiment (e.g., for high- versus low-frequency words); a mean residual time for nondecision parts of RT, $T_{er}$; and across-trial variability in the nondecision component of processing, range $s_t$ (see Table 1).

The diffusion model quantitatively fits the data from a number of binary decision tasks, including recognition memory, numerosity judgments, brightness discrimination, color discrimination, auditory discrimination, same–different judgments, letter discrimination, and visual search (Ratcliff, 1978, 1981, 1988; Ratcliff & Rouder, 1998, 2000; Ratcliff et al., 1999, Strayer & Kramer, 1994). The model can fit all aspects of the data—the probabilities of and the RTs for both correct and error responses and the shapes of RT distributions (and their hazard functions). It accurately accounts for the patterns of data that result from manipulations of speed versus accuracy and from manipulations that curtail decision processes, such as deadline, response signal, and speed–accuracy decomposition procedures (e.g., Meyer, Irwin, Osman, & Kounios, 1988). Diffusion models have also been applied in the domains of simple RT (Smith, 1995) and decision making (Busemeyer & Townsend, 1992, 1993; Diederich, 1997; Roe, Busemeyer, & Townsend, 2001), and diffusion models are close cousins of random walk models (Laming, 1968; Link, 1975; Link, 1992; Link & Heath, 1975; Smith, 1990; M. Stone, 1960).

In applying the diffusion model to data from the lexical decision task, we expected the degree of wordness to be higher for high-frequency words than for low-frequency words. For nonwords, we expected the degree of wordness to be lower for random letter strings than for pronounceable pseudowords. Degree of wordness determines drift rate in the diffusion model. A criterion is placed in the distribution of wordness values such that word stimuli generally have positive drift rates and nonwords generally have negative drift rates (see the drift rate criterion or relatedness criterion in Ratcliff, 1978, 1985; Ratcliff et al., 1999).

## Overview of the Experiments

The lexical decision experiments presented in this article were designed to provide data for evaluating the diffusion model as a model of the decision process in lexical decision. The experiments included manipulations of word frequency, the type of nonwords (nonword lexicality, e.g., Davelaar, Coltheart, Besner, & Jonasson, 1978; James, 1975; Shulman & Davison, 1977), the proportion of high- versus low-frequency words, and repetition of words and nonwords. The effects of these manipulations have been targets for modeling lexical access, so they were chosen to allow examination of how their effects can be interpreted through the decision process in the diffusion model.

## Experiments 1–6

The experiments varied word frequency and whether nonwords were pronounceable pseudowords or unpronounceable random strings of letters. The aim was to examine accuracy and the shapes of RT distributions for correct and error responses as a function of the two variables. The words were high-, low-, and very lowfrequency words (with mean frequency values of 325, 4.4, and .37 per million, respectively; Kučera & Francis, 1967). Experiments 1 and 2 included all three levels of frequency, and Experiments 3 and 4 included only the high- and low-frequency words. In Experiments 1, 3, 5, and 6, the nonwords were pseudowords, and

in Experiments 2 and 4, they were random letter strings. Experiments 5 and 6 examined the hypothesis (e.g., Glanzer & Ehrenreich, 1979) that word frequency effects are a product of strategies used by subjects, such that the choice of strategy depends on the proportions of high- versus low-frequency words in the experiment. In Experiment 5, 80% of the words were high-frequency words, and in Experiment 6, only 13% were high-frequency words.

## Method

**Subjects—**Northwestern undergraduates participated in the experiments for credit in an introductory psychology class. Sixteen students participated in Experiment 1, 14 in Experiment 2, 15 in Experiment 3, 17 in Experiment 4, 15 in Experiment 5, and 9 in Experiment 6.

**Materials—**There were 800 high-frequency words, with frequencies from 78 to 10,600 per million ($M$ = 325, $SD$ = 645; Kučera & Francis, 1967); 800 low-frequency words, with frequencies of 4 and 5 per million$M$ = 4.41, $SD$ = 0.19); and 741 very low-frequency words, with frequencies of 1 per million or no occurrence in Kuc era and Francis's corpus ($M$ = .365, $SD$ = .48). All the very low-frequency words did occur in the *Merriam-Webster's Ninth Collegiate Dictionary* 1990, and they were screened by three Northwestern undergraduate students; any words that any one of the three students did not know were eliminated.

From each word, a pseudoword was generated by randomly replacing all the vowels with other vowels (except for *u* after *q*), giving a pool of 2,341 nonwords. There was also a pool of 2,400 random letter strings, created by randomly sampling letters from the alphabet and then removing those strings that were pronounceable. The distributions of the numbers of letters per word for each type of word are shown in Table 2. The random letter strings had the same proportions for each length as the word strings for the three frequency groups combined, and these are also shown in Table 2.

**Procedure—**Stimuli were presented on a personal computer screen, with responses collected from the keyboard. Stimulus presentation and response recording were controlled by a real-time computer system.

Subjects were presented with strings of letters and instructed to decide if each string of letters was or was not an English word, pressing the/key for a word response and the *z* key for a nonword response. If a response was incorrect, the word "ERROR" was presented on the screen for 750 ms. The intertrial interval was 150 ms. Trials were grouped in blocks of 30; after each block, subjects had a self-paced break. The first block was used for practice and was not included in the data analysis.

In Experiment 1, 5 high-frequency, 5 low-frequency, 5 very low-frequency words, and 15 pseudowords were randomly selected without replacement for each of 50 blocks. No participant was ever presented with both a word and the pseudoword derived from it, and pseudowords were selected from the three pools in proportion to the words used in the experiment in this and subsequent experiments. Each subject was tested on 250 words of each type and on 750 pseudowords. The design of Experiment 2 was the same, except that the nonwords were random letter strings.

In Experiments 3 and 4, we did not use the very low-frequency words to test whether their absence would change the results from Experiments 1 and 2. There were 50 test blocks, each composed of 8 high-frequency words, 7 low-frequency words, and 15 nonwords (pseudowords in Experiment 3 and random letter strings in Experiment 4).

In Experiment 5, in each of 50 blocks of trials, there were 12 high-frequency words, 2 low-frequency words, 1 very low-frequency word, and 15 pseudowords. In Experiment 6, there

were also 50 blocks, each with 2 high-frequency words, 13 very low-frequency words, and 15 pseudowords.

## Results From Experiments 1–6

Responses longer than 2,000 ms and shorter than 350 ms (around 0.6% of the responses across all the experiments) were eliminated from the analyses. Data from three subjects who stopped participation early were discarded. Table 3 shows error mean RTs and correct mean RTs as well as standard errors for those RTs. It also provides predictions (discussed later) of the diffusion model. Observed and predicted .1 quantile RTs are also shown in Table 3.

The three pools of words—high-, low-, and very low-frequency—had different numbers of words for each word length. Consequently, any observed effects of word frequency could be due to the differing distributions of word lengths. For each experiment, we analyzed the data using only four- and five-letter strings (which allowed us to almost equate word length) and found that the patterns of results were in each case similar to the ones found with all the stimuli. Thus, all of the analyses that we present are based on all the stimuli.

## Correct Responses for Words: Accuracy and Mean RT

As shown in Table 3, the data replicated previous research: RTs increased and accuracy decreased as word frequency decreased, and responses were slower and less accurate when pseudowords were used in the experiment than when random letter strings were used. The differences in accuracy rates and RTs among the frequency conditions were larger with pseudowords than with random letter strings.

In all six experiments, the effects of word frequency were significant. In Experiment 1, with pseudowords, the difference in mean correct RTs between high- and low-frequency words was 68 ms, and the difference between low- and very low-frequency words was 40 ms, $F(2, 30) = 188.45$, $MSE = 252$ ( $p < .05$ throughout this article). The difference in probability correct from high- to very low-frequency words was .167, $F(2, 26) = 102.97$, $MSE = .0015$. In Experiment 2, with random strings of letters as nonwords, responses were about 100 ms faster overall and between −.004 (high-frequency words) and .127 (very low-frequency words) more accurate relative to Experiment 1. The differences in mean RTs for high- and lower frequency words were reduced to 40 ms and 20 ms (cf. James, 1975; Neely, 1977) but were still significant, $F(2, 26) = 50.28$, $MSE = 237$. The decrease in accuracy from high- to low- and very low-frequency words was also reduced to about .04, which was still significant, $F(2, 26) = 16.21$, $MSE = .00028$.

Experiments 3 and 4, which did not include very low-frequency words, showed the same patterns of results as Experiments 1 and 2. The RT difference between high- and low-frequency words was 66 ms with pseudowords as the nonwords, $F(1, 14) = 119.94$, $MSE = 415$, and 38 ms with random letter strings as the nonwords, $F(1, 14) = 68.68$, $MSE = 172$; the accuracy rates differences were .127 with pseudowords, $F(1, 14) = 55.06$, $MSE = .0017$, and .017 with random letter strings, $F(1, 14) = 26.84$, $MSE = .0000934$.

The manipulation of the proportion of high- versus low-frequency words had little effect on the patterns of results in Experiments 5 and 6 compared with Experiments 1 and 3 except that for Experiment 5, there were greater differences between RTs for high-, low-, and very low-frequency words. In Experiment 5, with a high proportion of high-frequency words, mean RTs and accuracy rates were similar to those in Experiment 1. The differences in mean RTs and accuracy across the three levels of word frequency were 89 ms and .072 between high- and low-frequency words and 61 ms and .117 between low- and very low-frequency words, $F(2, 26) = 85.20$, $MSE = 915$, for RT; and $F(2, 26) = 103.56$, $MSE = .0012$, for accuracy. In

Experiment 6, the result of using a high proportion of very low-frequency words was to produce slower responses in all conditions relative to Experiments 1 to 5 and to reduce the difference between high- and very low-frequency words relative to Experiment 5. The difference between high- and very low-frequency words in mean response time was 100 ms and in accuracy rates was .130, both significant: $F(1, 8) = 58.09$, $MSE = 1005$; and $F(1, 8) = 28.32$, $MSE = .0028$, respectively.

The manipulation of the proportion of high-frequency words in Experiments 5 and 6 was similar to a manipulation that has been labeled *frequency blocking*. In frequency blocking, blocks of trials that include only high-frequency words are compared to blocks that include equal proportions of high- and low-frequency words. Generally, RTs for high-frequency words are shorter in blocks that include only high-frequency words (Glanzer & Ehrenreich, 1979; Gordon, 1983; G. O. Stone & Van Orden, 1993). This result contrasts with the results presented here: RTs for high-frequency words were only about 30 ms shorter in Experiment 5, where there was a large proportion of high-frequency words, as in Experiment 6, where there was a low proportion. Also, RTs for high-frequency words in Experiment 5 were longer than in Experiment 1, where the proportions of high- and lower frequency words were about equal. However, the difference in RTs between high- and very low-frequency words was larger in Experiment 5 than in Experiments 1 and 6, suggesting something like a frequency blocking effect on the difference between high- and low-frequency RTs instead of on the RT for high-frequency words alone.

### Correct Responses for Nonwords

In general, correct nonword responses had about the same RTs or were a little shorter than the slowest word responses, which were responses for the low- or very low-frequency words. Also, nonword RTs were shorter for random letter strings than for pseudowords (by about 70 to 200 ms) and were more accurate (by about .02 to .04). In Experiments 1 through 6, type of nonword was manipulated between experiments. Experiment 7 compared responses to random letter strings and pseudowords in the same experiment.

### Error RTs

The effects of the two main variables—word frequency and type of nonword—on RTs for word stimuli were generally the same for error responses as for correct responses. Just as for correct RTs, error RTs decreased as word frequency increased, and error RTs were shorter when the nonwords were random letter strings than when they were pseudowords. Error RTs for nonwords were about the same as error RTs for low- and very low-frequency words.

The aspect of error responses that strongly constrains the diffusion model is their RT relative to the RT for correct responses. In the experiments with random letter strings as nonwords (Experiments 2 and 4), which were those with highest overall accuracy rates, the pattern was clear: Error RTs were shorter than correct RTs for both words and nonwords.

For Experiments 1, 3, 5, and 6—the experiments with pseudowords—the pattern was complex because there were individual differences (a complexity that is not shown in Table 3 because the RTs reported in the table are means across subjects). For some subjects in the conditions with highest accuracy (high-frequency words), there were no error responses or very few error responses. These subjects tended to be the slower and more accurate subjects. Because these subjects had few errors in the high-accuracy conditions, the error RTs in these conditions tended to come from fast, lower accuracy subjects (and so the entries in Table 3 reflect these subjects). In addition to this speed–accuracy effect, we noticed that the fast subjects tended to produce errors faster than correct responses, and the slow subjects tended to produce errors slower than correct responses. To show this, Table 4 combines Experiments 1, 3, and 5 (Experiment 6 did

not have low-frequency words) and splits the data so that accuracy and RTs for the fast and slow subjects are presented separately. Subjects put into the fast group ($n = 24$) had mean RTs shorter than the overall mean RT, and subjects put into the slow group ($n = 21$) had mean RTs longer than the overall mean RT. This split shows that error RTs for the fast subjects were shorter than correct RTs, whereas error RTs for the slow subjects were longer or about the same as correct RTs (the correct RT – error RT difference was about 30 ms for fast subjects and about −20 ms for slow subjects, averaged over high- and low-frequency words and pseudowords in Experiments 1, 3, and 5). The fast subjects were also somewhat less accurate than the slow subjects (see Ratcliff et al., 1999, Experiment 1, for discussion of similar speed–accuracy differences across individual subjects).

The differing patterns for fast versus slow subjects' correct and error RTs provided one of the main constraints on fitting the diffusion model. The only means the model had to account for the differing patterns was to allow boundary separation and variability in starting point to vary between fast and slow subjects. If we required variability in starting point to be the same for fast and slow subjects, the range of starting points would be a larger proportion of the total boundary separation when the boundary separation was small than when it was large. This can produce fast errors relative to correct responses when boundary separation is small (fast subjects) and slow errors relative to correct responses when boundary separation is large (slow subjects). This pattern of fast versus slow errors also depends on the other parameters; for different combinations of parameter values, the pattern also could be all fast errors or all slow errors for both speed and accuracy conditions (see Ratcliff & Rouder, 1998; Ratcliff et al., 1999).

## RT Distributions

To examine RT distributions, we used the RTs of each subject to calculate five quantile RTs: the .1, .3, .5, .7, and .9 quantiles. Then we averaged the quantiles across subjects to form the average quantiles shown in Table 5. (The average quantiles are not Vincent averages [Vincent, 1912], which are the averages of means of the RTs within bins [Ratcliff, 1979]; instead, the averages used here are averages over individual subjects' quantile RTs. Average quantiles were used because it was more efficient for the model to generate predictions for quantiles.)

Figure 3 shows the five quantiles for correct responses for the various types of word stimuli and for nonwords for Experiments 1 and 2. The data are represented by the crosses. The dark gray dots are the output of Monte Carlo simulations of the diffusion model, the +s are best-fitting values from the diffusion model, and the light gray dots are bootstrap samples designed to show the range of the data if the experiment was repeated; these are discussed later. The leading edges of the RT distributions are represented by the .1 quantiles (the lowest cross in each column), whereas their skews are represented by the spread of the higher quantiles. Overall, the RT distributions were positively skewed (i.e., larger separation among the higher quantiles than among the lower quantiles), the typical result in RT studies. When mean RT increased across the word frequency conditions, the distributions moved both in leading edge and spread, with the larger part of the increase in the mean coming from increasing spread of the longer quantiles. Although the effects in the leading edges of the distributions were small, they were significant. When the nonwords were pseudowords (Experiments 1, 3, 5, and 6), the leading edge shifted among the different word frequency conditions more than when the nonwords were random letter strings (Experiments 2 and 4), as is shown in Figure 3 for Experiments 1 and 2 and in Table 3 for Experiments 3 through 6. With pseudowords, averaging across experiments, the .1 quantile RT for high-frequency words was about 40 ms shorter than the .1 quantiles for low- and very low-frequency words. With random letter strings as nonwords, this difference in the .1 quantile RTs was smaller: 13 ms in Experiment 2 and 14 ms in Experiment 4 (see Table 3). The leading edges as measured by the .1 quantile RT varied

significantly as a function of word frequency: Experiment 1, $F(2, 30) = 81.37$, $MSE = 105.40$; Experiment 2, $F(2, 26) = 23.21$, $MSE = 55.23$; Experiment 3, $F(1, 14) = 112.68$, $MSE = 126.73$; Experiment 4, $F(1, 16) = 34.67$, $MSE = 47.12$; Experiment 5, $F(2, 26) = 35.78$, $MSE = 292.67$; Experiment 6, $F(1, 8) = 27.79$, $MSE = 385.31$.

### Summary

There were six main features of the data for modeling:

1. For words, accuracy increased and RT decreased (for both correct and error responses) as word frequency increased, and this was true whether the nonwords were random letter strings or pseudowords. The differences between the high- and low-frequency conditions were larger when the nonwords were pseudowords.
2. For words, RTs were shorter and accuracy was higher when the nonwords were random letter strings than when they were pseudowords.
3. For nonwords, correct responses had about the same RTs as correct responses for the slowest words. Responses were faster for random letter strings than for pseudowords, and accuracy was a little higher.
4. Most of the differences in RTs that occurred with increased word frequency were due to decreased skew of the RT distribution.
5. However, when the nonwords were pseudowords, there was a moderately large effect of frequency on the leading edge of the RT distribution: The leading edge for high-frequency words was shorter by 40 ms than the leading edges of the RT distributions for lower frequency words and nonwords. When the nonwords were random letter strings, the differences were considerably smaller, about 13–14 ms, but still significant.
6. With random letter strings, error RTs were shorter than correct RTs. Error RTs were also shorter than correct RTs with pseudowords but only for fast subjects; for slow subjects, error RTs were about the same as or longer than correct RTs (we present error RT distributions later).

Overall, these six features of the data provide severe constraints on fitting the diffusion model. With only six parameters plus one value of drift rate for each type of word (high-, low-, and very low-frequency) and each type of nonword (pseudowords and random letter strings), the model is required to fit the effects of word frequency and type of nonword on the complete sets of data: mean correct and error RTs for words and nonwords; accuracy rates for words and nonwords; the shapes of the RT distributions, including both skew and leading edge for correct responses for words and nonwords; and the relative speeds of correct versus error responses for words and nonwords.

### Method for Fitting the Diffusion Model to Data

To fit the diffusion model to the data, we formed a chi-square statistic and minimized it by adjusting the parameter values using a general SIMPLEX minimization routine. The data that were entered into the minimization routine for each experimental condition were the five quantile RTs averaged across subjects for both correct and error responses and the accuracy values. The quantile response times were fed into the diffusion model, and for each quantile, the cumulative probability of a response by that point in time was generated from the model. Subtracting the cumulative probabilities for each successive quantile from the next higher quantile yields the proportion of responses between each quantile. For the chi-square computation, these are the expected values, to be compared with the observed proportions of responses between the empirical quantiles. The expected values were multiplied by the number

of observations to produce expected frequencies. The observed proportions of responses for the quantiles are the proportions of the distribution between successive quantiles (i.e., the proportions between the 0, .1, .3, .5, .7, .9, and 1.0 quantiles are .1, .2, .2, .2, .2, and .1) multiplied by the probability correct for correct response distributions or the probability of error for error response distributions (multiplied by a number proportional to the number of observations in the condition). In a few cases, there were too few error RTs (less than five) to compute error RT quantiles for high-frequency words for more than one or two subjects. In these cases, these error RTs did not contribute to the fit; that is, no value of chi-square was computed for these conditions for error responses. Summing over (observed [O] − expected [E])$^2$/E for correct and error responses for each type of word and nonword gives a single chi-square value to be minimized:

$$\chi^2 = \Sigma(\mathrm{O} - \mathrm{E})^2 \Big/ \mathrm{E}.$$

In research on fitting the diffusion model to data with the chi-square method (Ratcliff & Tuerlinckx, 2002), it was found that parameter values could not be recovered accurately when there were enough long or short outlier RTs in the data to seriously affect the quantile RTs. In fitting the data reported here, we removed short outliers by trimming out responses shorter than 350 ms (e.g., Swensson, 1972), and we also removed very long outliers (longer than 2,000 ms). Ratcliff and Tuerlinckx explicitly modeled remaining contaminants by assuming that the contaminants in each experimental condition came from a uniform distribution that had maximum and minimum values corresponding to the maximum and minimum RTs in the condition. We performed the fits for the data here both with and without this assumption about contaminants. We found little difference between the two sets of fits and report the fits without the assumptions about contaminants.

The quality of fits of a model to data can sometimes be compromised by averaging. For example, in most of the experiments presented here, there were large differences among subjects in their overall accuracy values (e.g., 10%–15%). Pooling all of the data from all the subjects together can provide a picture that is not representative of any subject. To check for this problem, we computed the accuracy, mean RT, and .1 quantile values for each subject (in each condition) and averaged these values across subjects. The resulting averages looked much like the typical subject, providing reassurance that averaging over subjects as we did for the fits reported here did not introduce biases. This replicated a finding from three earlier studies (Ratcliff, Thapar, & McKoon, 2001, 2003; Thapar, Ratcliff, & McKoon, 2003) in which the diffusion model was fit both to data pooled over all subjects and to individual subjects; there were no systematic differences between parameter values in the two cases.

In fitting the model to the data from each experiment, all the parameters were held constant across the conditions of the experiment except drift rate. The parameters held constant were as follows: the starting point of the diffusion process ($z$), the across-trial variability in the starting point ($s_z$), the boundary separation ($a$), the nondecision time ($T_{er}$), the across-trial variability in the nondecision time ($s_t$), and the across-trial variability in drift rate ($\eta$). Drift rate ($v$) varied for words of the three different frequencies and for the two types of pseudowords. Variability within a trial $s$ is a scaling parameter (this means that the same fits could be obtained with another value of $s$ by rescaling the rest of the parameters), and its value was fixed at $s = .1$ for consistency with other published fits of the diffusion model to data.

The best-fitting parameter values are shown in Table 6. The values of the boundary separation and starting point parameters ($a$ and $z$) were highly consistent across the experiments. The type of nonword—pseudowords in Experiments 1, 3, 5, and 6 versus random letter strings in Experiments 2 and 4—produced a small difference in $T_{er}$ and in $s_t$: The values were smaller with random letter strings. However, it was not clear whether this was a systematic or random

effect; it was not obtained for Experiments 7, 8, and 9. The parameters for variability in drift ($\eta$) and variability in starting point ($s_z$) showed no systematic differences due to pseudowords versus random letter strings (except that $\eta$ and $s_z$ were a little higher for Experiments 2 and 4).

The only large and reliable effects on parameter values were the effects on drift rates of word frequency and type of nonword. Not surprisingly, the drift rate was higher for high-frequency words than for low-frequency words and was higher for low-frequency words than for very low-frequency words; in addition, the drift rate for random letter strings had a larger negative value than the drift rate for pseudowords.

Of the drift rate effects, the one that might be thought surprising in the context of some models (e.g., G. O. Stone & Van Orden, 1993) was that drift rate alone captured almost all of the effect on word RTs of the type of nonword. The differences in RTs among words of different frequencies were larger with pseudowords than with random letter strings, and this was accounted for by differences in drift rates: Differences among the drift rates were smaller with pseudowords than with random letter strings. In particular, although the drift rates for high-frequency words were about the same in all the experiments, the drift rates for the low- and very low-frequency words were lower when the nonwords were pseudowords. This is clearly observable in Table 6, where the drift rates for low- and very low-frequency words in Experiments 1, 3, and 5 are always numerically smaller than their drift rates in Experiments 2 and 4.

Differences in drift rates also accounted for the shift in the leading edge of the RT distribution for high-frequency words relative to lower frequency words. The leading edge of the high-frequency word distribution was shifted about 40 ms shorter relative to the leading edges of the low-frequency and very low-frequency word distributions when the nonwords were pseudowords. The diffusion model accurately captured this with only differences in drift rate as a function of word frequency (see Table 3). The diffusion model accommodated the shift in leading edge for the reasons discussed in the introduction.

Table 3 shows fits of the model to the data for correct and error mean RTs, accuracy values, and .1 quantile RTs for correct responses, along with standard errors in these quantities. Figure 3 is designed to show the model's goodness of fit graphically for the data from Experiments 1 and 2. In the figure, the *X*s are the experimental data and the +s are the predicted values from the model with the best-fitting parameter values.

We calculated two different estimates of variability, one using a graphical Monte Carlo method (Ratcliff & Tuerlinckx, 2002) to show variability in the model's predictions and the other using a bootstrap method to show variability in the data. For the graphical Monte Carlo method, for each experiment, we first generated sets of simulated data, each set with the same number of observations as for each subject in the experiment. We repeated this to produce the same number of data sets as there were subjects. For each data set, quantile RTs and accuracy values were calculated, and these were averaged over data sets (in the same way the experimental data were averaged over subjects). This was repeated 100 times, and the dark gray dots in Figure 3 plot the quantile RTs and accuracy values for each of the 100 replications (see Ratcliff & Tuerlinckx, 2002). Variability in accuracy values is represented by the scatter of the dots along the *x*-axis and variability in the quantile values is represented by scatter along the *y*-axis.

The bootstrap method we used allows an estimate of the variability that would result if the experiment were rerun with new subjects. We used two levels of random selection. First, for each subject in the experiment, we sampled with replacement from the experimental data for that subject to generate a new set of bootstrap data for the subject. Second, we sampled with replacement from the subjects to produce a new set of subjects, and for each of these subjects, we used the bootstrap data we had generated (as just described). The idea was to represent what

would happen with a different random sample of subjects than those that actually participated in the experiment (see Efron, 1982). For each of the simulated subjects, we calculated their quantile RTs and accuracy values and averaged these across subjects in the same way as for the experimental data. We repeated this 100 times, and the resulting values are plotted as the light gray dots in Figure 3.

Figure 3 shows that, for Experiments 1 and 2, the two types of simulated data overlap each other, which means that the model predictions vary in the same ways as would be expected from the experimental data. Although only the results for Experiments 1 and 2 are displayed in the figure, we did the same simulations for Experiments 3 through 6. Across all six experiments, only 9 out of the 105 quantile RTs predicted from the model with the best-fitting parameter values (the +s) are outside 2 *SE* confidence intervals for the bootstrap simulated data (light gray dots), and only 24 out of 105 of the data points (the *X*s) are outside 2 *SE* confidence intervals for the Monte Carlo simulated predictions for the model (dark gray dots).

Table 3 provides values of mean RT and accuracy for the data, standard error values in each, and the predicted values from the model with best-fitting parameter values. These standard errors supplement the Monte Carlo and bootstrap studies. All except two differences between the data and the predicted accuracy values were within .025, all except two differences between the data and the predicted mean RTs were within 25 ms, and all except one difference between the data and the predicted .1 quantile RTs were within 16 ms. Error RTs are more variable because they are based on many fewer observations; all except one difference between the predicted and data values were within 40 ms.

## Discussion

In lexical decision, the discriminability of words from nonwords is reflected in both accuracy and RT, and so a model must account for both dependent measures. A measure based on accuracy alone, such as d', ignores the other dependent variable, RT. Measures of RT alone ignore trade-offs in accuracy that accompany changes in speed. The diffusion model provides an integrated account of both speed and accuracy. The RT and accuracy data from Experiments 1 through 6 are translated by the model into drift rates, which are measures of discriminability for the various experimental conditions.

The important result of the experiments is that variations in drift rate account for all the observed effects of word frequency and type of nonword on RT and accuracy. Discrimination of words from nonwords, measured by drift rate, is better for high- than lower frequency words, and it is better when the nonwords are random letter strings than when they are pseudowords. The interaction between these two variables is also a matter of discrimination: The difference in discriminability between high- and lower frequency words is larger with pseudowords than with random letter strings.

Figure 4 (top panel) shows how discriminability can be represented in a two-dimensional signal-detection framework. Two is the smallest number of dimensions that can adequately accommodate the relative drift rates for the various classes of stimuli. Distances among the classes of stimuli in the space represent differences in their wordness values. How the values from the figure enter into the diffusion decision process can be understood as follows: Suppose that the decision process has multiple sources of information feeding into it from the lexicon, including semantic information, phonological information, orthographic information, and other kinds of lexical information. With random letter strings as the nonwords in an experiment, all of these sources of information are valid indicators of whether a stimulus is a word. In effect, if a string of letters looks like a word, if its phonemes are wordlike, and if it has meaning, these sources of evidence combine to produce a high value of wordness and hence a high drift rate toward the "word" decision boundary. But with pseudowords in the experiment instead of

random letter strings, some of the sources of information, especially orthographic information, are less reliable and so are not used (or they are weighted much less) in determining drift rate. For lack of better insight into what the two dimensions might unambiguously represent, we label them *lexical strength* and *orthographic wordlikeness*. Two-dimensional representation suggestions like the one proposed here have been made previously, for example, by G. O. Stone and Van Orden (1993).

The two-dimensional representation can be embedded for illustrative purposes in a two-dimensional signal-detection framework (e.g., Ashby, 2000). In the top panel of Figure 4, pseudowords (PWs) are considerably lower than words on the lexical strength dimension but only a little lower on the orthographic dimension. Random letter strings (RL) are considerably lower on both dimensions than words and pseudowords. High-frequency words are higher on the lexical strength dimension than low-frequency words, which are higher than very low-frequency words. Differences in drift rates between the different item types are given by distances between them in the two-dimensional space. When the nonwords are random letter strings, both dimensions figure into a determination of the distances. The distance between words and nonwords is represented by $x$ in the figure and the distance among the high-, low-, and very low-frequency words is represented by $u$, where $u$ is determined by projecting the differences among the high-, low-, and very low-frequency words onto the diagonal (as shown by the dashed lines in the figure). When the nonwords are pseudowords, the orthographic dimension is not reliable and so distances are computed on the lexical strength dimension alone. The distance between words and nonwords is represented by $y$, and the distance among the three types of words is represented by $v$. The relative distances determine the relative values of the drift rates that enter the diffusion model, as shown in the bottom panel of Figure 4.

## Experiment 7

In Experiments 1 through 6, each experiment included only one type of nonword. Experiment 7 included both types. According to the hypotheses about the lexical decision process just outlined, with pseudowords in an experiment, orthographic "wordlikeness" information should be less reliable in making a decision than it is in an experiment with random letter strings as nonwords. In other words, the data should look much like the data from Experiments 1, 3, 5, and 6. Responses to high-frequency words should be faster than responses to lower frequency words, and this advantage should show up in both mean RTs and as a shift in the leading edge of the high-frequency RT distribution relative to the distributions for low- and very low-frequency words. Also, correct RTs to random letter strings should be considerably shorter than correct RTs to low-frequency and very low-frequency words because difficult negative items have been introduced into the experiment (e.g., Ratcliff, 1985; Ratcliff & Hacker, 1981).

### Method

**Subjects—**Fifteen Northwestern University students participated in this experiment for credit in an introductory psychology class.

**Materials—**New pools of words were used in this experiment (these were developed to examine neighborhood effects, which are not reported here): 558 high-frequency words, with frequencies of 100 or more per million ($M$ = 249, $SD$ = 623.30; number of letters, $M$ = 5.76, $SD$ = 1.58); 501 low-frequency words, with frequencies of 3 to 6 per million ($M$ = 4.3, $SD$ = .76; number of letters, $M$ = 6.11, $SD$ = 1.76); and 381 very low-frequency words, with frequencies of 1 per million (number of letters, $M$ = 5.46, $SD$ = 1.60; Kučera & Francis, 1967).

**Procedure—**The lexical decision procedure was the same as in Experiments 1 to 6. There were 15 blocks of 100 trials, each block consisting of 18 high-frequency words, 18 low-frequency words, 14 very low-frequency words, 25 pseudowords, and 25 random letter strings, all ranging in length from 4 to 9 letters.

## Results

Responses shorter than 350 ms and longer than 2,000 ms were eliminated from the analyses (about 0.5% of the responses). Table 7 shows a summary of the results.

As expected, for correct responses, responses to high-frequency words were shorter, by 60 ms, than responses to low-frequency words and shorter by 115 ms than responses to very low-frequency words, $F(2, 28) = 65.05$, $MSE = 671.83$. Responses to high-frequency words were also more accurate by .04 than responses to low-frequency words and were more accurate than responses to very low-frequency words by .14, $F(2, 28) = 91.76$, $MSE = .00088$.

For correct responses for nonwords, responses to random letter strings were more accurate by .10, $F(1, 14) = 63.24$, $MSE = .00012$; and their mean correct RT was 150 ms shorter than responses to pseudowords, $F(1, 14) = 161.21$, $MSE = 1246.43$.

The pattern of RTs for error responses to word stimuli relative to correct RTs fell in about the middle of the results from the earlier experiments in which the nonwords were pseudowords (Experiments 1, 3, 5, and 6). For high-frequency words, errors were faster than correct responses, whereas for low- and very low-frequency words, errors were slower than correct responses. Error responses to nonwords were faster than correct responses.

Table 8 displays the RT distributions for correct responses. Just as in Experiments 1, 3, 5, and 6, which also used pseudowords, the .1 quantile RT (the leading edge of the RT distribution) for high-frequency words began to rise above zero earlier (20 ms and 34 ms earlier, respectively) than the .1 quantile of the RT distributions for the low- and very low-frequency words, $F(2, 28) = 34.03$, $MSE = 133.70$. In the earlier experiments with pseudowords, the average difference in the .1 quantile between the distributions for high- and very low-frequency words was larger, about 45 ms, and in the earlier experiments with random letter strings, there was a smaller difference (e.g., 13 ms in Experiment 2). Here, with both pseudowords and random letter strings in the same experiment, the high frequency/very low-frequency .1 quantile difference was intermediate in value. There was also a 59-ms difference in the .1 quantile for correct responses to pseudowords versus random letter strings, $F(1, 14) = 89.86$, $MSE = 281.41$.

## Fits of the Diffusion Model

The diffusion model was fit to the data from Experiment 7 using the same method as for the previous experiments, and the results were consistent with those experiments (see Table 7). The differences among the experimental conditions were all well-captured by variations in drift rate. The drift rates for words increased as a function of word frequency, and the drift rate for random letter strings had a larger absolute value than the drift rate for pseudowords. The fits were good, with the predicted mean RT for correct responses within 15 ms of the observed mean RT in all except two conditions (very low-frequency words and pseudowords, which missed by 25 and 26 ms, respectively, both less than 2 *SE*s). The predicted accuracy values were all within 3% of the data. Fits for error RTs were not as good, although all were within 2 *SE*s of the data.

The diffusion model fit the .1 quantile reaction times to within 7 ms for each stimulus type except pseudowords, which missed by 15 ms, which was not a significant difference (the width of the confidence interval was greater than 30 ms; for typical spreads in Experiments 1 and 2,

see Figure 3). The difference in the .1 quantile between the two types of nonwords was fit with a large difference in drift rate between random letter strings and pseudowords (.237) just as for high-frequency versus lower frequency words in Experiments 1, 3, 5, and 6 (Table 3).

In Experiment 7, the values of the drift rates for word stimuli were a little larger than in the earlier experiments with pseudowords (Experiments 1, 3, 5, and 6; see Table 6) but smaller than those for Experiment 2, which used random letter strings. This finding indicated that mixing random letter strings with pseudowords made the decision process only a little less difficult compared with when all the nonwords were pseudowords. The other parameters for the fits of the diffusion model were within the ranges of the values for Experiments 1 through 6 (see Table 6).

## Experiments 8 and 9

In lexical decision, responses to repeated words are faster than responses to first presentations of the words. Repetition has a larger effect on low-frequency words than on high-frequency words, so it reduces the difference in mean RT between them (Balota & Spieler, 1999; Duchek & Neely, 1989; Forster & Davis, 1984; Scarborough, Gerard, & Cortese, 1979). To provide more data against which to test the diffusion model, we collected repetition data—in Experiment 8, in the context of pronounceable nonwords, and in Experiment 9, in the context of random letter strings.

### Method

Twenty-one Northwestern undergraduates participated in Experiment 8 and 21 in Experiment 9 for credit in an introductory psychology course. Items were selected from the same pools of words and nonwords as in Experiments 1 through 6, and the same lexical decision procedure was used. There were 20 blocks, each with 14 high-, 14 low-, and 14 very low-frequency words and 42 nonwords. Each stimulus was presented twice. The repetitions were in adjacent blocks so the lag between presentations had a mean of 84 items.

### Results

Responses longer than 2,000 ms and shorter than 350 ms were eliminated from the analyses (less than .5% of the responses in both experiments). The results are shown in Table 9.

In Experiment 8 (with pseudowords as the nonwords), for words presented for the first time, accuracy was reduced as a function of word frequency by .10 from high- to low-frequency words and by .11 from low-frequency to very low-frequency words. For the second presentations of words, these differences were reduced to .04 and .05. Averaged over word frequency, correct responses to words were about .06 more accurate with repetition. The main effects of frequency and repetition were both significant, $F(2, 40) = 105.61$, and $F(1, 20) = 158.15$, as was their interaction, $F(2, 40) = 47.22$, $MSE = .00070$.

For words presented for the first time, mean correct RT was increased by 74 ms from high- to low-frequency words and by 37 ms from low- to very low-frequency words. For the second presentations of words, these differences were reduced to 46 ms and 28 ms. Averaged over word frequency, correct RTs to words were about 46 ms shorter with repetition. The main effects of frequency and repetition were both significant, $F(2, 40) = 119.34$, and $F(1, 20) = 80.56$, as was their interaction, $F(2, 40) = 11.06$, $MSE = 340.20$.

With repetition, correct responses to nonwords slowed by a nonsignificant 3 ms, $F(1, 20) = 3.29$, $MSE = 80.27$. Repetition significantly increased accuracy but also by a small amount, .02, $F(1, 20) = 26.74$, $MSE = .00014$.

Mean RT for error responses to the first presentations of high-frequency words was 607 ms, but there were too few error responses for the second presentations to compute a mean (6 subjects had no errors, and 4 had only 1). For low- and very low-frequency words, error RTs were shorter than correct RTs on the first presentations and longer on the second presentations. Error RTs were longer than correct RTs on both the first and second presentations for pseudowords.

For words presented for the first time, the leading edges of the RT distributions shifted as a function of word frequency, much as in the earlier experiments: The .1 quantile was 40 ms shorter for high- than low-frequency words and 48 ms shorter for high- than very low-frequency words. The leading edge differences were reduced on the second presentation, to 25 ms and 41 ms, respectively. The main effects of frequency and repetition were both significant, $F(2, 40) = 96.57$, and $F(1, 20) = 145.24$, as was their interaction, $F(2, 40) = 4.31$, $MSE = 145.28$.

In Experiment 9 (with random letter strings as the nonwords), for words presented for the first time, accuracy was reduced as a function of word frequency by .02 from high- to low-frequency words and by .03 from low-frequency to very low-frequency words. For the second presentation of words, these differences were reduced to .01 and .01. Averaged over word frequency, with repetition, correct responses to words became about .01 more accurate. The main effects of frequency and repetition were both significant, $F(2, 40) = 37.50$, and $F(2, 40) = 4.47$, as was their interaction, $F(2, 40) = 2.64$, $MSE = .00036$.

For words presented for the first time, mean correct RT was increased by 23 ms from high- to low-frequency words and by 28 ms from low-frequency to very low-frequency words. For the second presentations of words, these differences were reduced to 10 ms and 13 ms. Averaged over word frequency, correct responses to words were about 20 ms shorter on the second presentation. The main effects of frequency, repetition, and their interaction were significant, $F(2, 40) = 51.41$, $F(1, 20) = 74.05$, and $F(2, 40) = 11.1$, $MSE = 186.40$. Repetition affected neither RT for nonwords, $F(1, 20) = .14$, $MSE = 69.84$, nor accuracy for nonwords, $F(1, 20) = 1.59$, $MSE = .00015$.

Error responses for both words and nonwords were faster than correct responses in all conditions, with RTs from 18 ms to 55 ms shorter, and changed little as a function of repetition. However, accuracy in this experiment was above 93% correct in all conditions, and so there were few responses contributing to the error RT means.

There were small, but significant, effects on the leading edges of the RT distributions as a function of word frequency, $F(2, 40) = 28.26$, and repetition, $F(1, 20) = 17.70$. The interaction between the two was not significant, $F(2, 40) = 0.85$, $MSE = 154.03$. The difference in the .1 quantile RTs between high- and low-frequency words was 16 ms, and between high- and very low-frequency words it was 17 ms. Table 8 shows the response time quantiles for the data and the fits of the model.

As in Experiments 1 to 6, the effects of word frequency were larger in Experiment 8, with pseudowords, than in Experiment 9, with random letter strings. Repetition interacted with word frequency and the type of nonword; the effect of repetition produced a larger decrease in RT and a larger increase in accuracy in Experiment 8 than in Experiment 9.

### Fits of the Diffusion Model

The diffusion model was fit to the data from Experiments 8 and 9 using the same method as in the previous experiments, and the results, fits, and parameter values are shown in Tables 9 and 10. The differences among the experimental conditions were all reasonably well-captured by variations in drift rate. Drift rates for words increased as a function of word frequency and

as a function of repetition (except for high-frequency words in Experiment 9, which changed little as a function of repetition). Drift rates for nonwords did not change as a function of repetition and had a smaller absolute value for Experiment 8 (pseudowords) than for Experiment 9 (random letter strings). The other parameters of the model were within the ranges of those from Experiments 1 through 7.

Predicted correct RTs matched the experimental data within 20 ms, and accuracy values matched the data within .04. Predictions for error RTs sometimes missed by as much as 40 ms, but some of the error RTs for which theory and data missed were based on less than five observations for some of the subjects. For the second presentations for high-frequency words in Experiment 8, 6 subjects had no error responses (and hence no experimental value is reported). The largest miss between the predicted and experimental .1 quantile RTs was 17 ms (only 2 out of 17 were outside the 95% confidence intervals). We performed the same analyses as in Figure 3 with confidence intervals from model simulations and from bootstrap computations from the experimental data. For the Monte Carlo simulations from the model, the data points significantly missed the predictions in 15 out of 105 correct responses for Experiments 7, 8, and 9. For the bootstrap confidence intervals from the experimental data, 7 out of 105 predictions missed.

## Error RT Distributions

Error RTs are much more variable than correct RTs because the number of observations is smaller. For example, if accuracy is .9, there is only a little more than one tenth the number of observations for errors as correct responses. As a consequence, standard errors on mean error RTs (presented in earlier tables) were between two and five times larger than standard errors for correct RTs. If we displayed the error RT quantiles as in Figure 3, the standard error bars would be much greater for errors and adjacent Monte Carlo or bootstrap quantiles would overlap each other. Table 11 gives the experimental and predicted error RT distributions for the more reliable of the error RT distributions, those conditions in which the number of observations for each subject was greater than six so that five quantiles could be computed. Most of the .1 quantile RTs predicted by the model for errors were within 15 ms of the empirical ones, but 5 out of 23 missed by more than 15 ms. However, generally, the diffusion model captured the shapes of the error RT distributions adequately as well as the mean error RTs presented in the earlier tables.

## General Discussion

### The Diffusion Model

The diffusion model has enjoyed considerable success across a range of cognitive paradigms, and here we have shown it to be successful with the lexical decision task. The model did a good job of fitting the data from all nine experiments.

Not only did the model fit the mean RTs for correct responses, it also fit accuracy rates, mean RTs for error responses, the relative speeds of correct and error responses, and the shapes of the RT distributions and their leading edges. Too often, error responses have been ignored in the development of models for the lexical decision task, so the application of a model to error responses is relatively novel in this domain (but see Grainger & Jacobs, 1996). It is unlikely that a model could be developed to fit only RTs for correct responses and then, serendipitously, turn out to be able to fit RT distributions and error RTs. Models must be developed to address all of the data simultaneously.

The parameters of the diffusion model correspond to components of processing. When the model fits the data well, as it does for the experiments presented here, the parameter values

that provide the best fits to the data can be interpreted in terms of task variables, showing how the variables affect various processing components. Overall, we found that the effects of word frequency, type of nonword, and repetition on correct and error RTs and accuracy were all handled by drift rate. Only drift rates, not any of the other parameters, were systematically affected across conditions and experiments. As would be expected, drift rate increased with word frequency and with repetition, and drift rate had a larger negative value for random letter strings than for pseudowords.

In the data, the leading edge of the RT distribution (measured by the .1 quantile RT) for words was shifted for high-frequency words relative to very low-frequency words over the experiments by about 13 ms when the nonwords were random letter strings and about 45 ms when the nonwords were pseudowords. In the model, the leading edge shift comes about through an interaction between variability in the nondecision component of RT and differences in drift rates between the high-frequency and lower frequency words. With a large drift rate, the leading edge depends only on the shortest values of the nondecision component of RT, whereas with smaller drift rates, it depends on all the values in the distribution.

Drift rates for words were larger when the nonwords in an experiment were random letter strings than when they were pseudowords. We explain this by assuming that the decision process can make use of multiple sources of information, such as orthographic, phonemic, and semantic information, as sketched in Figure 4. To decide whether a test string is a word, the sources of information are combined to provide a single quantity, the degree of wordness, which maps into drift rate in the diffusion model. With random letter strings, orthographic information, phonemic information, semantic information, and so on are all valid indicators of whether a stimulus is a word and so contribute to the value of wordness, whereas with pseudowords, the usefulness of orthographic wordlikeness information is considerably reduced (see Figure 4).

The values of the parameters of the diffusion model other than drift rate varied little across all the experiments. This finding indicates that the components of processing represented by boundary positions, starting point, across-trial and within-trial variability in drift rate, and the nondecision component of RT and its variability all were not systematically affected by type of nonword, the proportions of high- versus low-frequency words, or whether items were repeated or not.

A general issue that arises is whether the parameter values of the fits we report are unique. One question concerns drift rates: In the fits we present, the effects of all the experimental variables are accommodated in drift rate, not in any of the other parameters. The question is whether the model could still fit well if the effects of the experimental variables were all taken up by other parameters of the model. To address this question, we refit the data from Experiment 2 (random letter strings) with drift rates fixed at the values derived from Experiment 1 (pseudowords). The result was that the RT values were well predicted, but the accuracy values missed by .05 for nonwords, by less than .01 for high-frequency words, by .06 for low-frequency words, and by .16 for very low-frequency words. Except for high-frequency words, these misses are unacceptably large.

Another way to test whether changes in drift rate are all that is needed to account for the data is to hold all the other parameters exactly constant across the experiments. If drift rate can account for the effects of all the independent variables, then the model should still fit well. For this test, the model was refit to the data from Experiment 2 with all the parameters except drift rate fixed at their values from Experiment 1. The result was that drift rates differed significantly between the two experiments, in about the same way as for the fits presented above (shown in Table 6). The fits were reasonably good; the main systematic miss was in the .1 quantile RT,

where the model consistently overpredicted the data by about 20 ms. The fact that the fits were reasonably good reflects the fact that parameter values other than drift rate do not need to be systematically different between the two experiments. The original fits that we reported are better mainly because of a 28-ms difference in the $T_{er}$ parameter between the two experiments, a difference that could reflect individual differences between the two groups of subjects (cf. Ratcliff et al., 2001).

The questions just addressed are part of the general issue of model flexibility. To address this issue, Ratcliff (2002) generated a number of plausible but fake data sets and attempted to fit the diffusion model to them. If it were true that the diffusion model is flexible enough to fit any pattern of data, then the model should have provided good fits for the fake patterns, but it could not do so, showing that it is extremely constrained by real data.

Another set of issues concerns variability in the parameter values of the model. First, although parameter values certainly vary across subjects, Ratcliff et al. (2001) and Thapar et al. (2003) showed that they are not biased but instead appear to be symmetrically distributed around their means. Second, the variability in the parameter values that is due to sampling variability from the data is not large, nor is it biased (Ratcliff & Tuerlinckx, 2002). To put it another way, given a set of parameter values for the model, random samples of data generated from the model will produce variable RTs and accuracy values. When the model is fit to these data, the resulting parameter values are slightly different from the parameters used to generate the simulated data, but the deviations are small and unbiased. This means that when the model is applied to the data from an experiment, it is possible that slightly different parameter values would also produce adequate fits, but the differences would not be large enough to change any conclusions.

In summary, the diffusion model is able to account for all facets of the experimental data presented here, including correct and error RT distributions and accuracy. The parameter values are consistent across the experiments, and they are similar to those reported with other experimental paradigms. The drift rates obtained from the fits are interpretable in terms of a two-dimensional signal-detection framework. We now turn to models of lexical access to explore how they relate to the diffusion model framework.

## Lexical Decision

The diffusion model analysis offers a new and simple view of the effects of independent variables on processing in the lexical decision task. Other models have used the effects of type of nonword, word frequency, or repetition to support hypotheses about different lexicons or different processing pathways for different stimulus types. From the diffusion model point of view, the effects of these variables are simply to alter the amount and kind of information contributing to the degree of wordness that drives the decision process and nothing more. The lexical system that feeds information to the decision process may have many facets, but once information is output from the system, it can be considered unidimensional (see Balota & Chumbley, 1984; Seidenberg & McClelland, 1989; and G. O. Stone & Van Orden, 1993, for similar views of the role of lexical access versus decision processes, though their models are considerably different from ours).

In comparison to previous models, the diffusion model forces consideration of all aspects of the experimental data. Below we discuss how some other models fail to deal with the full range of data. We attempt to show for each model what could be changed to allow its output to feed into a diffusion decision process in such a way as to appropriately explain the data. This exercise helps to illustrate exactly what our modeling does and does not contribute to an explanation of lexical decision. The model can account for the data, and it can provide an explanation of the

decision process, but it does not provide insights into lexical representations or how they are accessed.

It is important to note in the discussion below that, even if a lexical model appears to produce the qualitatively correct behavior of some quantity of how wordlike a stimulus is, there is no guarantee that the combination of the lexical model and the diffusion decision model will work. What would be needed to see if the combination works is a quantitative examination of whether the values of wordness produced by the lexical model map into drift appropriately (cf. criticism of Seidenberg & McClelland's [1989] model by Besner, Twilley, McCann, & Seergobin [1990]). Thus, the following discussion suggests only possible beginning points for research.

Although many models have been developed to account for performance in the lexical decision task, we can find none that has been successfully quantitatively fit to all aspects of the experimental data. This means that we cannot be sure that the models can account for the data they have been used to explain. This failing is exacerbated by the fact that only one of the models (Grainger & Jacobs, 1996) attempts to deal with all of the dependent variables in the data, namely, correct and error RT distributions and accuracy. For some of the other models, if they did make predictions about all the aspects of the data, it is likely that they would be wrong, and for others of the models, completely new assumptions would be needed.

### Serial Search Models

Rubenstein, Lewis, and Rubenstein's (1971) model assumes that words are ordered in the lexicon by frequency and that search starts with high-frequency words and proceeds to low-frequency words. In Forster's (1976) model, the orthographic representations of words are ordered in a peripheral access file that provides the address of a lexical entry in a master file. The orthographic representations are organized in bins based on the similarity of their first few letters, and within a bin, the words are ordered by frequency. Searches begin with the highest frequency words in a bin. For both of these models, RTs are predicted to be shorter for high- than lower frequency words. For Rubenstein et al.'s model, nonword responses are produced when the serial search terminates, and in Forster's model, nonword responses are produced when the search of a bin has produced no matching string. The models are mute as to how errors occur.

A major problem with serial search models is that they cannot accommodate a finding of correct nonword responses faster than correct word responses, because nonword responses can be made only after the serial search is terminated. For the same reason, serial search models cannot account for errors being faster than correct responses. Hence, serial search models would have to be changed substantially to accommodate the data presented here. The serial access assumption could be turned into a parallel access assumption about the relative availability of information from the various peripheral access files, and with such a move, much of the lexical structure could be retained. But this would be such a radical change that it is not clear whether the result could still be called the same model, and it is not clear whether it could quantitatively fit the data.

### Logogen and Parallel Search Models

One of the earliest and most influential models of the parallel search class is the logogen model (Morton, 1969, 1979). According to this model, lexical entries are represented as logogens—an auditory input logogen, a visual input logogen, and an output logogen. When a string of letters is input visually to the system, all the logogens that contain a feature that is in the stimulus letter string are incremented in parallel. Word identification occurs when evidence from the input reaches a threshold amount in a word's logogen. The threshold level for identification is a function of frequency: Less information is needed to identify a high-frequency word than a

low-frequency word. The model was developed mainly to explain data from paradigms in which a word must be identified for production, that is, naming tasks. It was not developed to explain the binary, word–nonword responses required by the lexical decision task. Mechanisms to produce nonword responses were added later as extensions to the model.

The logogen model is similar to random walk and diffusion models (see Ratcliff & McKoon, 1997) in that both kinds of models assume evidence accumulates over time toward response criteria (thresholds in the case of the logogen model). Perhaps the key difference between the logogen and diffusion models is that, because the logogen model was mainly concerned with word production tasks and not the binary lexical decision task, it uses the thresholds of individual logogens to implement the effects of word frequency (e.g., Ratcliff & McKoon, 1997). In contrast, the diffusion model is concerned with lexical decision, and the effects of word frequency are implemented in drift rates. The reason is that, for lexical decision, the specific word represented by a stimulus does not need to be identified; the only information needed is how wordlike the stimulus is. So it is assumed that evidence is accumulated on the basis of the degree of wordness, evaluated against two response criteria. The criteria are the same for all stimuli, and they cannot be altered as a function of the frequency of the stimulus (because this would require first identifying the frequency of the stimulus, which would make the decision process superfluous). The assumption that only drift rates are affected by word frequency is supported by the fact that the model fits the experimental data well with this assumption.

Coltheart, Davelaar, Jonasson, and Besner (1977) extended the logogen model to produce nonword responses in lexical decision by assuming a time deadline for negative responses, with a nonword response being produced when processing takes longer than the deadline. The deadline is adjusted as a function of the total activation in the logogen system. When the nonwords are random letter strings and a nonword is presented, the total activation summed over logogens is low, so a short deadline is set and RTs for nonwords are fast. Errors arise from variability in the flow of information to the logogen system. At a qualitative level, Coltheart et al.'s extension of the logogen model correctly predicts the effects of word frequency and type of nonword on mean RTs for correct responses to words and nonwords. However, it is unclear whether it could make accurate quantitative predictions for accuracy values, for error RTs, or for the shapes or leading edges of RT distributions.

Perhaps the most problematic aspect of the data for Coltheart et al.'s (1977) model is the fact that error responses are sometimes faster than correct responses. When a high-frequency word is presented, the total activation in the logogens should be high, and so the response deadline should be long and nonword error responses should be slow. Also, the RT distribution for nonword responses should be the same as the distribution of deadline times, which in turn is a function of the total activation in the system. The deadline times are usually assumed to be normally distributed; so the distributions of nonword RTs should be normal (cf. Grainger & Jacobs, 1996). However, empirical RT distributions are never normally distributed.

### Gordon's Parallel Resonance Model

Gordon's (1983) resonance model shares with the logogen model the assumption that the internal representations of all words are activated when a string of letters is presented to the system (see also Ratcliff, 1978). In Gordon's model, presentation of a letter string causes the internal representations of the words to resonate as a function of the degree to which they match the test string. The strength of the resonance drives the rate of accumulation of evidence in a single boundary decision process.

For lexical decision, the strength of resonance is largely determined by the frequency of the test word. The model can correctly predict word frequency effects on mean RT for correct

responses for words. But there is no mechanism to produce error responses or nonword responses; Gordon's experiment was a go/no-go task in which nonword responses were not made. Also the model was never explicitly fit to experimental data, and so it was never determined whether it could produce RT distributions that quantitatively match empirical ones.

Gordon's resonance model could be made compatible with the diffusion model: A criterion could be placed on the amount of resonance such that if the amount was above the criterion, the drift rate in a diffusion process would be positive, and if it was less than the criterion, the drift rate would be negative (cf. a drift criterion, Ratcliff, 1985; Ratcliff et al., 1999). The drift rates would determine response times and accuracy values just as described for the data from the experiments reported here.

### Multiple Read-Out Model

In the multiple read-out model (Grainger & Jacobs, 1996), word identification for visually presented strings of letters is accomplished via an orthographic lexicon structured as a localist connectionist network derived from McClelland and Rumelhart's (1981) interactive activation model. Words are represented as collections of the orthographic features contained in their letters. When a string of letters is input to the system, the representations of all words orthographically similar to the input are activated. Lateral inhibition among words that share features causes them to inhibit each other so the strongest beats down its competitors.

A word decision is based on two sources of information: global activity, the summed activation of all the words in the lexicon, and local activity, the activation values of individual words. A word response is made if either exceeds criterial values. The criterial values are variable across trials, and the mean of the global activity criterion is adjustable by strategy or experimental context. Nonword responses are based on a time deadline that is variable across trials with its mean adjustable in the same way as the global activity criterion. If global activity is small, the deadline is set short and nonword responses are fast. Errors occur because of variability in the criteria across trials.

Because high-frequency words produce more global and more local activation, the model predicts the correct qualitative effects of word frequency on response probabilities and mean RTs for word responses. However, it is not clear whether correct predictions could be produced for the shapes of RT distributions and the differences in their leading edges as a function of word frequency. The model cannot make correct predictions about RT distributions for nonwords because the deadline time is assumed to be normally distributed across trials (see data and simulations reported by Grainger & Jacobs, 1996).

It is possible for the multiple read-out model to produce errors to words that are faster than correct responses but only under unusual circumstances that do not match experimental data. If the distribution of deadlines is made extremely wide (i.e., large standard deviation) and the mean of the deadline distribution is relatively high (longer than the .9 quantile RT for correct responses), then errors can be faster than correct responses. However, this situation results in the leading edge of the RT distribution for errors being very short relative to correct responses. We simulated this situation and obtained error responses that were 30 ms shorter than correct RTs, but the .1 quantile RT for errors was 100 ms shorter than that for correct responses. In the experimental data, the .1 quantile RTs for correct and error responses are much closer together (within 10 ms).

A prediction of the multiple read-out model is that as word frequency increases, global activation increases and so the deadline for nonword responses increases. In another simulation, we decreased the mean of the distribution of word RTs and increased the deadline distribution mean to mimic the effect of increased activation. This slowed error responses, making them

slower than correct responses, unlike the data in which error responses speed up as correct responses speed up.

The results of the simulations show that it is likely to be impossible to get the deadline model to produce (a) shorter RTs for error responses than correct responses while at the same time producing the correct behavior of .1 quantile RTs, (b) a decrease in error RTs as word frequency and accuracy increase and correct RT decreases, and (c) RT distributions for correct responses to nonwords that are just as skewed as correct word RT distributions.

The multiple read-out model also has a meta-level problem in the assumption that the deadline criterion is adjusted within a trial on the basis of global activation (a problem that also applies to Coltheart et al.'s [1977] model). The output of the interactive activation part of the model is deterministic, which is why early global activation can be used to set the deadline. The question is why the system does not base its decision on this early information. Why wait to have the decision depend on a noisy criterion when there is accurate information available in the first few milliseconds of processing? If the model were altered so that the information output from interactive activation was noisy (i.e., not deterministic) and so could not be used to make an accurate early decision, then neither could it give information accurate enough for deadline setting. Thus, the model would be unable to produce fast negative responses, especially when the nonwords were random letter strings.

The multiple read-out model could be made consistent with data by assuming that information from interactive activation enters a noisy diffusion decision process. Variability in the decision process would have to be assumed, but there could also be variability in the activation process. However, there is no guarantee (without a comprehensive study) that an integration of interactive activation and the diffusion model could provide adequate quantitative fits to data.

### Activation–Verification Model

The activation–verification model (Paap & Johansen, 1994; Paap, McDonald, Schvaneveldt, & Noel, 1987; Paap, Newsome, McDonald, & Schvaneveldt, 1982) is a two-phase model. When a string of letters is input to the system, letter and then word representations are activated, with the amount of activation based on confusion matrices among letters. If the activation value of a word passes a criterial value, it is considered in the second, serial search, phase. In the serial search, words are considered one at a time in decreasing order of their frequency. If the match between the input letter string and a word reaches some criterial value, a "word" response is generated. If the match exceeds the criterion for none of the words, a "nonword" response is generated. Errors come from nonwords activating words that then pass the criterion in the serial search (false alarms) or cases in which all the candidate words activated by a word fall below the criterion used in the serial search (misses). Variability in processing comes about from item differences. The same item on each trial will produce the same activation value and the same candidate list of words for the serial search, but different items from the same word class will produce different activation values and different lists of candidate words on different trials. (There is a second possible decision process in which a decision is based on letter representations alone, but this has not been implemented and so we do not consider it here.)

The serial search based on frequency gives the model word frequency effects on mean RTs, but it shares the problems described earlier for serial search models; for example, it has no mechanism to account for errors faster than correct responses.

The activation–verification model uses confusability matrices between letters (data from subjects' performance in letter identification) to generate activation values, and so it can make predictions about individual test items. If the activation values were used to directly determine drift rate in a diffusion process (deleting the serial search phase of the model), then it is possible

that the composite model could produce correct qualitative predictions for the dependent variables. However, some elaboration of the model would be needed to deal with the effects of type of nonword.

### Dual Dictionary Model

The dual dictionary model (Glanzer & Ehrenreich, 1979) explains word frequency effects by assuming two lexicons: One is a fast access lexicon that contains only high-frequency words and yields short RTs, and the second is a slower access lexicon that contains all words. A nonword response can be generated only after failure to find a match in either lexicon. No specific proposals have been made about how errors arise.

The data from Experiments 5 and 6 provide particular problems for the model. In Experiment 5, there was a large proportion of high-frequency words, whereas in Experiment 6, there was a large proportion of very low-frequency words. A large proportion of very low-frequency words should cause the system to search the all-words lexicon on all trials, and so there should be no advantage to high-frequency words. However, high-frequency words were significantly faster than very low-frequency words in both experiments, by amounts that were not too different.

### Information Pathways and Strategic Control Frameworks

G. O. Stone and Van Orden (1993) examined two frameworks for explaining performance in the lexical decision task. The first suggests that the lexical processing system is made up of several independent modules, each of which deals with a different kind of information (e.g., phonological or orthographic), and the system can select, for any given task, which modules (or “information pathways”) give productive output. In the lexical decision task, different modules could be selected for distinguishing words from random letter strings than for distinguishing words from pseudowords.

The second framework proposes a fixed processing system, and parameters are adjusted by strategic control for different tasks and different contexts within tasks, such as the use of random letter strings versus pseudowords. The decision process Stone and Van Orden considered was a random walk, and they examined the effects on RTs and accuracy values of altering the rate of accumulation of evidence and the response criteria settings. The diffusion model presented here can be viewed as an instantiation of this framework. The only required addition would be the use of either the two-dimensional signal-detection representation presented in Figure 4 or something that served the same purpose. This would provide differences in drift rate as a function of type of nonword. A combination of the pathways and random walk/diffusion models should be capable of dealing with all the aspects of the data for the experiments presented here, but as we have noted in discussing other models, quantitative fits to data would be necessary for complete evaluation.

### The Dual Route Cascaded (DRC) Model

Coltheart, Rastle, Perry, Langdon, and Ziegler (2001) proposed the DRC model as an updated version of Coltheart et al.’s (1977) model. The model implements word recognition processes with two routes between letter inputs and an output phonemic system (see also Coltheart, Curtis, Atkins, & Haller, 1993; Morton, 1979), one a direct grapheme to phoneme route and the other through orthographic and phonemic lexicons. The second route includes letter units, orthographic units, phonological units, and phonemic units, structured as a local connectionist network. The DRC model has been applied to a much wider domain than lexical decision, including word naming and the effects of various cognitive impairments on naming performance. The application to lexical decision is identical to Grainger and Jacobs’s (1996) model. Activation levels for individual words and a value of global activation are computed

from the orthographic output layer, with separate response criteria. Nonword responses are produced from a time deadline, and errors occur because of variability in the criteria across trials.

The computational framework of this model, excluding the decision mechanism, could be consistent with the diffusion model in the same way as Grainger and Jacobs's (1996) model. Degrees of wordness would be computed for words and nonwords, and wordness values would enter the diffusion decision process as drift rate. As in the two-dimensional representation scheme in Figure 4, wordness values for words would have to be reduced when the nonwords were random letter strings as opposed to when the nonwords were pseudowords. Note that as for all the earlier combinations of lexical models with the diffusion model, there is no guarantee that the combination of this model and the diffusion model would produce the correct quantitative behavior.

### Distributed Connectionist Models

Seidenberg and McClelland's (1989) system is made up of an input layer of orthographic nodes, a hidden layer of nodes, and an output layer of phonological nodes. During word learning, for each word input to the system, a pattern of activation is entered into the input nodes. Activation flows from the input orthographic nodes through the hidden nodes to the phonological nodes and back to the orthographic nodes. The output activation patterns at the orthographic and phonological layers are compared with "teacher" target patterns (one for the orthographic representation and one for the phonological representation), and the differences between them are used to adjust the weights of the connections among all the nodes using the back-propagation algorithm. High-frequency words are presented to the system more frequently during training and so are better learned. During test trials, activation from an input string of letters flows from the orthographic units to the phonological units and back to the orthographic units. The degree of match between the input and the internally generated output drives the decision process. If the match value is above a criterion, a "word" response is produced, and if not, a "nonword" response is produced. To account for the effects of type of nonword, Seidenberg and McClelland (1989) proposed that when the nonwords are pseudowords, phonological as well as orthographic output is assessed, but this was not implemented (see the critique by Besner et al., 1990, and the reply by Seidenberg & McClelland, 1990). Error responses come from high-match values when nonwords are presented and low-match values when a word is presented. The model does not have an explicit mechanism to produce RTs, although Seidenberg and McClelland (1989) suggested that degree of match might map onto rate of accumulation of evidence in a manner similar to Ratcliff's (1978) diffusion model. With such an output process, correct and error responses and their distributions could be predicted.

Plaut (1997) recently updated this approach by adding a layer of semantic nodes to the layers of orthographic and phonological nodes. He proposed that a measure of semantic stress based only on activity in the semantic nodes could be used as the basis for lexical decisions, with the value of semantic stress driving a stochastic decision process that provides accuracy and reaction time measures. However, the use of the semantic nodes alone to provide a value of drift rate would not produce the drop in drift rate for low- and very low-frequency words when pseudowords are used instead of random letter strings. If drift rate could also be partly determined by outputs from the orthographic or phonological layers, and the relative importance of semantic versus orthographic or phonological output could be adjusted, then the model could be qualitatively consistent with the effects of type of nonword. But as with all the other models discussed, the resulting model would have to be fit to data to determine whether it can produce correct quantitative predictions.

### Familiarity–Recheck Model

Balota and Chumbley (1984, 1990) have argued that word frequency effects are a by-product of the decision process in the lexical decision task and not a consequence of the way the mental lexicon is organized or accessed. They proposed a model inspired by signal-detection theory in which the familiarity values of words and nonwords are distributed normally, with higher frequency words having higher familiarity than lower frequency words. The two distributions are assumed to be separated but overlapping. For the decision process, two criteria are set on the familiarity dimension. If the familiarity value of a string of letters is above the upper criterion, a positive response is initiated, and if it is below the lower criterion, a negative response is initiated. If the familiarity value is between the two criteria, an extra, slow rechecking process is needed. Errors could arise from three sources: word familiarity below the lower criterion or nonword familiarity above the upper criterion; failure in the rechecking process; or guesses, which occur after a time deadline on the rechecking process. Word frequency effects are explained as shifts in the familiarity distributions. High-frequency words have a higher mean familiarity value and so they are more likely to exceed the upper criterion, which gives them faster RTs. This model was not tested quantitatively, and additional assumptions would have to be added to do so. Mainly, the model served as a vehicle with which to argue that qualitative explanations of lexical decision data do not have to depend on lexical structure or lexical access processes.

Recently, Balota and Spieler (1999; see also Andrews & Heathcote, 2001) examined RT distributions in the lexical decision task and attempted to account for their behaviors as a function of repetition and frequency with several two-stage models. The simple versions of the models were shown to be inadequate. For example, in one model, the time required for the familiarity stage was assumed to have a Gaussian distribution and the time required for the rechecking stage was assumed exponential, giving a combined RT distribution of the convolution of the Gaussian and exponential distributions. The model Balota and Spieler supported was a hybrid model in which familiarity values are assumed to be normally distributed. Words of different frequencies come from different familiarity distributions, with the means of the distributions shifted as a function of frequency. If the familiarity of a word is larger than the upper criterion, RT for that word is obtained from a normal distribution, with its mean a function of word frequency. If the familiarity value is below the lower criterion, RT is randomly selected from normal distribution (for nonword responses). If familiarity is between the upper and lower criteria, RT is sampled from an ex-Gaussian distribution, the same ex-Gaussian regardless of word frequency. With this model, Balota and Spieler were able to account for the shapes of RT distributions for correct responses to words and nonwords as a function of word frequency and repetition.

Balota and Spieler (1999) noted that the model did not provide an account of error rates or error RTs, and they discussed possible alternatives to it. Their general conclusion was that any model of the lexical decision task should address RT distributions. The diffusion model does that, with a more integrated approach than their hybrid model.

## Conclusion

Application of the diffusion model to the data from Experiments 1 through 9 produces a simple, perhaps even boring, picture: Performance in the lexical decision task is a matching process in which noisy evidence is accumulated over time, with the quality of the evidence for a string of letters derived from a two-dimensional representation of how wordlike the string is. In this framework, the lexical decision task does not provide a window into the complexities of lexical processing. Models of lexical processing make contact with the diffusion model by producing outputs that can be interpreted in terms of a two-dimensional representation. The interpretation

of data offered by the diffusion model is much simpler and much less mysterious than has been the case with many alternative theoretical accounts of processing.

The data presented in this article are similar to data reported in a number of articles on lexical decision. The data served as a base for testing the diffusion model, and the model provides good quantitative fits for all aspects of the data. Drift rates measure the degree to which a stimulus is wordlike—its wordness. With a simple two-dimensional signal-detection model, it is possible to qualitatively describe the relative wordness values for various types of words and nonwords and how those values are affected by whether the nonwords in an experiment are pseudowords or random letter strings. One aim for future research is to determine whether any current word recognition models can produce the right numerical values of degree of wordness to map into drift rates such that all the aspects of the data can be accommodated.

Coltheart et al. (2001) listed experimental results that they considered to be benchmark phenomena that any model of lexical decision must explain; all of them concerned the effects of experimental variables on mean RT for correct responses to words and nonwords. The only variables on the list that we have not addressed are those involving the size and makeup of a word's lexical neighborhoods. We expect that neighborhood variables, like word frequency and type of nonword, have their effects on drift rate. More important, we add to Coltheart et al.'s list the full range of the dependent variables in the lexical decision task: accuracy values, RT distributions, and the relative speeds of correct and error responses.

One of the major problems with models of the lexical decision task has been that none of them has been quantitatively fit to the full range of data. The nearest that any of the models has come is to provide simulations of patterns of results that qualitatively match some aspects of data (e.g., Grainger & Jacobs, 1996). Thus, we cannot be sure that any of the models are actually capable of accounting for even those aspects of the data that they have attempted to explain. And even if the predictions of the models for some aspects of the data are correct, predictions for RT distributions, accuracy, and error RTs likely would not match the data.

In summary, application of the diffusion model to lexical decision data shows that the major effects of word frequency, type of nonword, and repetition are all captured by a single component of processing: drift rate. The consequence is that RT data, which early in the development of models of lexical access appeared to provide information about lexical structure, do not provide such information after all. All that RT data, in conjunction with the other aspects of the data, provide is a single value of the degree of match of a stimulus to the lexicon, with experimental variables changing the degree of match value. From the perspective of modeling the lexicon, this means that we can focus on models that simply hand the decision process a single value of goodness of match and not on models that base complicated decision processes on various streams of information in competition with each other or models that search different parts of the lexicon, unless these models can ultimately provide a single value of goodness of match.

**Acknowledgements**

Portions of this research (data and modeling) were presented at the annual meeting of the Society for Mathematical Psychology, Bloomington, IN, July 1997. Preparation of this article was supported by National Institutes of Mental Health Grants R37-MH44640 and K05-MH01891 and by National Institute on Deafness and Other Communication Disorders Grant R01-DC01240. We thank David Balota, Max Coltheart, Jonathan Grainger, and Manuel Perea for comments on an early draft of the article.

## References


Andrews S, Heathcote A. Distinguishing common and task-specific processes in word identification: A matter of some moment? Journal of Experimental Psychology: Learning, Memory, and Cognition 2001;27:514–544.

Ashby FG. A stochastic version of general recognition theory. Journal of Mathematical Psychology 2000;44:310–329. [PubMed: 10831374]

Balota DA, Chumbley JI. 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 1984;10:340–357. [PubMed: 6242411]

Balota DA, Chumbley JI. Where are the effects of frequency in visual word recognition tasks? Right where we said they were. Comment on Monsell, Doyle, & Haggard (1989). Journal of Experimental Psychology: General 1990;119:231–237. [PubMed: 2141355]

Balota DA, Spieler DH. Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. Journal of Experimental Psychology: General 1999;128:32–55. [PubMed: 10100390]

Besner D, Twilley L, McCann R, Seergobin K. On the association between connectionism and data: Are a few words necessary? Psychological Review 1990;97:432–446.

Busemeyer JR, Townsend JT. Fundamental derivations from decision field theory. Mathematical Social Sciences 1992;23:255–282.

Busemeyer JR, Townsend JT. Decision field theory: A dynamic–cognitive approach to decision making in an uncertain environment. Psychological Review 1993;100:432–459. [PubMed: 8356185]

Coltheart M, Curtis B, Atkins P, Haller M. Models of reading aloud: Dual-route and parallel-distributed-processing approaches. Psychological Review 1993;100:589–608.

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, NJ: Erlbaum.

Coltheart M, Rastle K, Perry C, Langdon P, Ziegler J. DRC: A dual route cascade model of visual word recognition and reading aloud. Psychological Review 2001;108:204–256. [PubMed: 11212628]

Davelaar E, Coltheart M, Besner D, Jonasson JT. Phonological recoding and lexical access. Memory and Cognition 1978;6:391–402.

Diederich A. Dynamic stochastic models for decision making under time constraints. Journal of Mathematical Psychology 1997;41:260–274. [PubMed: 9325121]

Duchek JM, Neely JH. A dissociative word-frequency × levels-of-processing interaction in episodic recognition and lexical decision. Memory and Cognition 1989;17:148–162.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans* Philadelphia: Society for Industrial and Applied Mathematics.

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 257–287). Amsterdam: North-Holland.

Forster KI, Davis C. Repetition and frequency attenuation in lexical access. Journal of Experimental Psychology: Learning, Memory, and Cognition 1984;10:680–698.

Glanzer M, Ehrenreich SL. Structure and search of the internal lexicon. Journal of Verbal Learning and Verbal Behavior 1979;18:381–398.

Gordon B. Lexical access and lexical decision: Mechanisms of frequency sensitivity. Journal of Verbal Learning and Verbal Behavior 1983;22:24–44.

Grainger J, Jacobs AM. Orthographic processing in visual word recognition: A multiple read-out model. Psychological Review 1996;103:518–565. [PubMed: 8759046]

James CT. The role of semantic information in lexical decisions. Journal of Experimental Psychology: Human Perception and Performance 1975;1:130–136.

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

Laming, D. R. J. (1968). *Information theory of choice reaction time* New York: Wiley.

Link SW. The relative judgement theory of two choice response time. Journal of Mathematical Psychology 1975;12:114–135.

Link, S. W. (1992). *The wave theory of difference and similarity* Hillsdale, NJ: Erlbaum.

Link SW, Heath RA. A sequential theory of psychological discrimination. Psychometrika 1975;40:77–105.

McClelland JL, Rumelhart DE. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review 1981;88:375–407.

*Merriam-Webster's ninth collegiate dictionary* (1990). Springfield, MA: Merriam-Webster.

Meyer DE, Irwin DE, Osman AM, Kounios J. The dynamics of cognition and action: Mental processes inferred from a speed–accuracy decomposition technique. Psychological Review 1988;95:183–237. [PubMed: 3375399]

Morton J. The interaction of information in word recognition. Psychological Review 1969;76:165–178.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing visible language 1* (pp. 259–268). New York: Plenum.

Neely JH. Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited capacity intention. Journal of Experimental Psychology: General 1977;106:226–254.

Paap KR, Johansen L. The case of the vanishing frequency effect: A retest of the verification model. Journal of Experimental Psychology: Human Perception and Performance 1994;20:1129–1157.

Paap, K. R., McDonald, J. E., Schvaneveldt, R. W., & Noel, R. W. (1987). Frequency and pronounceability in visually presented naming and lexical decision tasks. In M. Coltheart (Ed.), *Attention and performance XII* (pp. 221–243). Sussex, England: Erlbaum.

Paap K, Newsome SL, McDonald JE, Schvaneveldt RW. An activation–verification model for letter and word recognition. Psychological Review 1982;89:573–594. [PubMed: 7178333]

Plaut DC. Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. Language and Cognitive Processes 1997;12:767–808.

Ratcliff R. A theory of memory retrieval. Psychological Review 1978;85:59–108.

Ratcliff R. Group reaction time distributions and an analysis of distribution statistics. Psychological Bulletin 1979;86:446–461. [PubMed: 451109]

Ratcliff R. A theory of order relations in perceptual matching. Psychological Review 1981;88:552–572.

Ratcliff R. Theoretical interpretations of speed and accuracy of positive and negative responses. Psychological Review 1985;92:212–225. [PubMed: 3991839]

Ratcliff R. Continuous versus discrete information processing: Modeling the accumulation of partial information. Psychological Review 1988;95:238–255. [PubMed: 3375400]

Ratcliff R. A diffusion model account of reaction time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. Psychonomic Bulletin & Review 2002;9:278–291. [PubMed: 12120790]

Ratcliff R, Hacker MJ. Speed and accuracy of same and different responses in perceptual matching. Perception & Psychophysics 1981;30:303–307. [PubMed: 7322806]

Ratcliff R, McKoon G. A counter model for implicit priming in perceptual word identification. Psychological Review 1997;104:319–343. [PubMed: 9127584]

Ratcliff R, Rouder JN. Modeling response times for two-choice decisions. Psychological Science 1998;9:347–356.

Ratcliff R, Rouder JN. A diffusion model account of masking in letter identification. Journal of Experimental Psychology: Human Perception and Performance 2000;26:127–140. [PubMed: 10696609]

Ratcliff, R., & Smith P. L. (in press). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*

Ratcliff R, Thapar A, McKoon G. The effects of aging on reaction time in a signal detection task. Psychology and Aging 2001;16:323–341. [PubMed: 11405319]

Ratcliff R, Thapar A, McKoon G. A diffusion model analysis of the effects of aging on brightness discrimination. Perception & Psychophysics 2003;65:523–535. [PubMed: 12812276]

Ratcliff R, Tuerlinckx F. Estimating the parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. Psychonomic Bulletin & Review 2002;9:438–481. [PubMed: 12412886]

Ratcliff R, Van Zandt T, McKoon G. Connectionist and diffusion models of reaction time. Psychological Review 1999;106:261–300. [PubMed: 10378014]

Roe RM, Busemeyer JR, Townsend JT. Multialternative decision field theory: A dynamic connectionist model of decision-making. Psychological Review 2001;108:370–392. [PubMed: 11381834]

Rubenstein H, Lewis SS, Rubenstein MA. Homographic entries in the internal lexicon: Effects of systematically and relative frequency of meanings. Journal of Verbal Learning and Verbal Behavior 1971;10:57–62.

Scarborough DL, Gerard L, Cortese C. Accessing lexical memory: The transfer of word repetition effects across task and modality. Memory & Cognition 1979;7:3–12.

Seidenberg MS, McClelland JL. A distributed, developmental model of word recognition and naming. Psychological Review 1989;96:523–568. [PubMed: 2798649]

Seidenberg MS, McClelland JL. More words but still no lexicon: Reply to Besner et al. Psychological Review 1990;97:447–452.

Shulman HG, Davison TCB. Control properties of semantic coding in a lexical decision task. Journal of Verbal Learning and Verbal Behavior 1977;16:91–98.

Smith PL. A note on the distribution of response time for a random walk with Gaussian increments. Journal of Mathematical Psychology 1990;34:445–459.

Smith PL. Psychophysically principled models of visual simple reaction time. Psychological Review 1995;102:567–591.

Smith PL, Vickers D. The accumulator model of two-choice discrimination. Journal of Mathematical Psychology 1988;32:135–168.

Stone GO, Van Orden GC. Strategic control of processing in word recognition. Journal of Experimental Psychology: Human Perception and Performance 1993;19:744–774. [PubMed: 8409857]

Stone M. Models for choice reaction time. Psychometrika 1960;25:251–260.

Strayer DL, Kramer AF. Strategies and automaticity: I. Basic findings and conceptual framework. Journal of Experimental Psychology: Learning, Memory, and Cognition 1994;20:318–341.

Swensson RG. The elusive tradeoff: Speed versus accuracy in visual discrimination tasks. Perception & Psychophysics 1972;12:16–32.

Thapar A, Ratcliff R, McKoon G. A diffusion model analysis of the effects of aging on letter discrimination. Psychology and Aging 2003;18:415–429. [PubMed: 14518805]

Vincent SB. The function of the vibrissae in the behavior of the white rat. Behavior Monographs 1912;1:1–81.

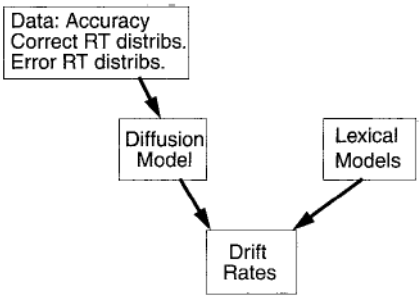


**Figure 1.**
The relationship between data, the diffusion model fits, drift rates, and models of word identification. The diffusion model fits the data and provides values of drift rate that represent how wordlike the stimulus is. The word identification models need to produce values of drift rate to provide a complete description of the data. A complete model would represent lexical processing, which would produce drift rates to feed into the diffusion model to produce predicted values of the dependent measures. RT distribs. = response time distributions.

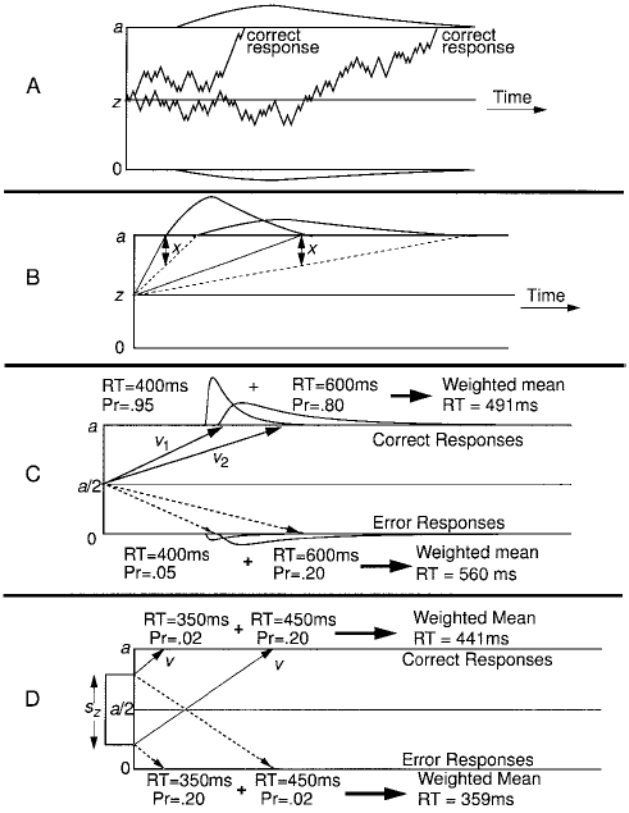


**Figure 2.**

An illustration of the diffusion model. A: Two sample paths that illustrate the variability in information accumulation from trial to trial. B: Illustration of how distribution shape changes as drift rates decrease. If the fastest and slowest processes are both slowed by *x*, the fastest responses are slowed a little (leading edge of the distribution) and the slowest responses are slowed a lot (tail). C: Effect of averaging response times (RTs) for two drift rates, $v_1$ and $v_2$, when the starting point is midway between the two boundaries. Error responses are slower than correct responses because more slow responses are averaged with fewer fast responses. D: Effect of averaging RTs for two values of starting points, $a/2 - s_z$ and $a/2 + s_z$. Error responses are faster than correct responses because more fast errors are averaged with fewer slow errors. $a$ = boundary separation; $z$ = starting point; $s_z$ = range of distribution of starting point; Pr = probability.

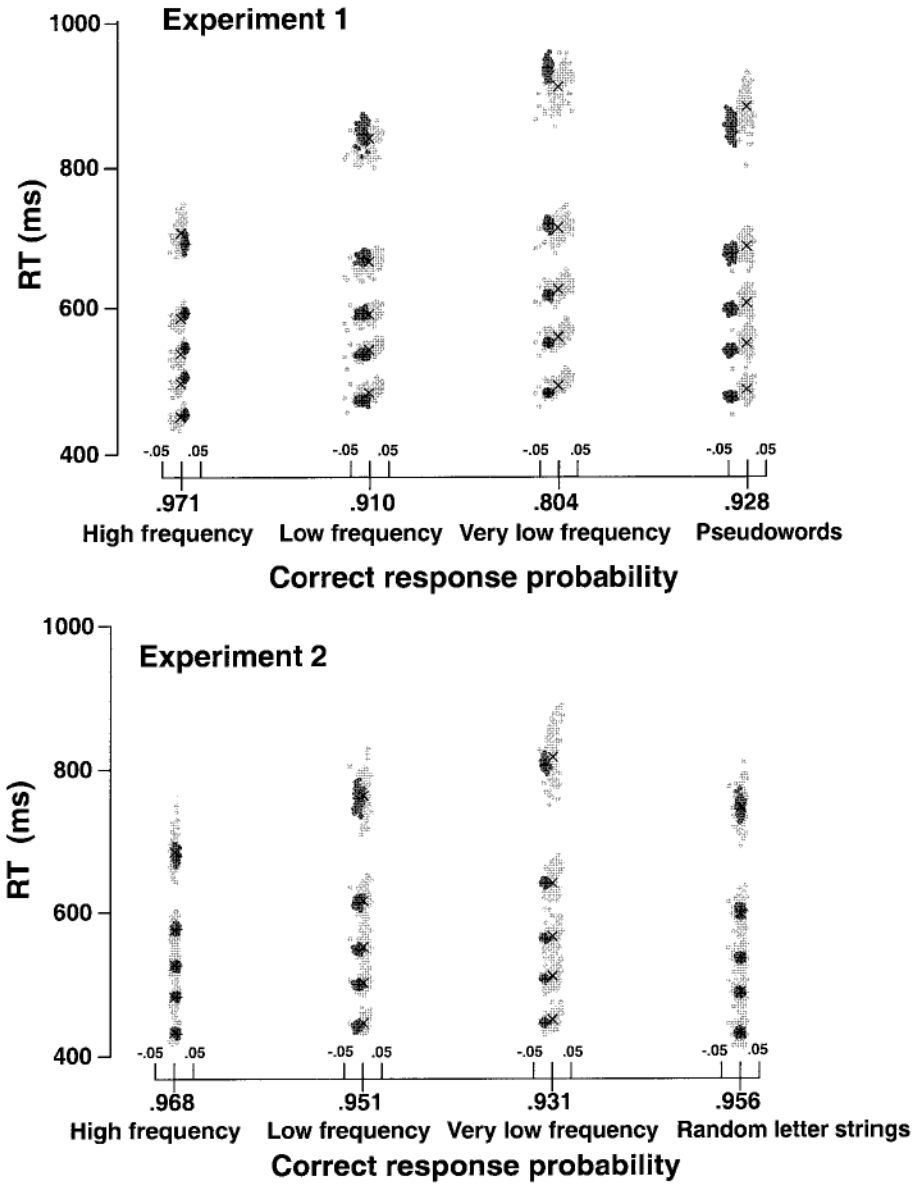


**Figure 3.**

Empirical and predicted .1, .3, .5, .7, and .9 quantiles for the response time (RT) distributions in Experiments 1 and 2. The ×s are quantile RTs plotted against accuracy values calculated from the data. The +s are the predicted values from the model with the best-fitting parameter values. The dark gray dots show variability from Monte Carlo simulations based on the model, and the light gray dots show variability from bootstrap simulations from the data. The light gray dots also represent variability that would be expected if the experiment was replicated with new subjects.

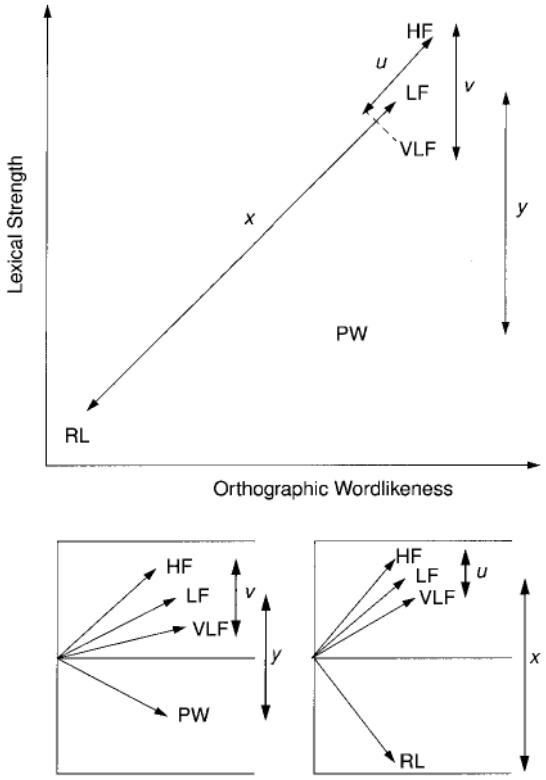


**Figure 4.**

Top: An illustrative two-dimensional signal-detection view of distances between the different kinds of stimuli in the experiments. HF = high-frequency words, LF = low-frequency words, VLF = very low-frequency words, PW = pseudowords, and RL = random letter strings. When the negative items are random letter strings, the distance between LF and negative items is *x* and is greater than *y*, the distance between LF and negative items when the negative items are pseudowords. Bottom: An illustration of how distances in the two-dimensional signal-detection space would map into drift rate in the diffusion model, the left panel with pseudowords and the right panel with random letter strings. In contrast, the distance between HF and VLF is larger when the negative items are pseudowords (distance *v*) than when they are random letter strings (distance *u*).

## Table 1

Parameters of the Diffusion Model

| Parameter | Description |
|---|---|
| $a$ | Boundary separation |
| $z$ | Starting point |
| $s_z$ | Variability in starting point across trials (range of a uniform distribution) |
| $T_{er}$ | Nondecision component of response time (e.g., encoding, response output) |
| $s_t$ | Variability in nondecision component of response time across trials (range of a uniform distribution) |
| $v$ | Drift rate (one for each experimental condition) |
| $\eta$ | Variability in drift rate across trials (standard deviation of a normal distribution) |
| $s$ | Variability in drift within each trial (standard deviation), a scaling parameter |

**Table 2**

Numbers of Words in the Stimulus Pools per Number of Letters and Type of Stimulus

| Type of stimulus | Number of letters | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| High-frequency | 228 | 190 | 142 | 115 | 66 | 43 | 15 | 800 |
| Low-frequency | 81 | 125 | 185 | 172 | 135 | 93 | 6 | 800 |
| Very low-frequency | 144 | 106 | 97 | 173 | 144 | 46 | 31 | 741 |
| Random letter strings | 226 | 210 | 212 | 230 | 172 | 91 | 26 | 1,167 |

## Table 3

Results From Experiments 1–6 as a Function of Stimulus Type

| | Error mean RT (ms) | | | Correct mean RT (ms) | | | Probability correct | | | Correct RT at .1 quantile (ms) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Type of stimulus | Data | *SE* | Model | Data | *SE* | Model | Data | *SE* | Model | Data | Model |
| Experiment 1: HF, LF, and VLF words and pseudowords | | | | | | | | | | | |
| Pseudowords | 698 | 24 | 668 | 661 | 14 | 650 | .928 | .011 | .889 | 492 | 484 |
| High-frequency | 636 | 72 | 593 | 571 | 9 | 570 | .971 | .008 | .984 | 453 | 458 |
| Low-frequency | 682 | 25 | 674 | 639 | 9 | 642 | .910 | .021 | .900 | 487 | 479 |
| Very low-frequency | 686 | 24 | 711 | 679 | 12 | 683 | .804 | .023 | .785 | 497 | 489 |
| Experiment 2: HF, LF, and VLF words and random letter strings | | | | | | | | | | | |
| Random letter strings | 591 | 39 | 606 | 575 | 15 | 582 | .956 | .010 | .957 | 432 | 432 |
| High-frequency | 497 | 10 | 514 | 549 | 12 | 556 | .968 | .006 | .973 | 433 | 434 |
| Low-frequency | 590 | 48 | 566 | 589 | 14 | 591 | .951 | .010 | .943 | 446 | 443 |
| Very low-frequency | 625 | 37 | 599 | 609 | 16 | 614 | .931 | .011 | .915 | 451 | 448 |
| Experiment 3: HF and LF words and pseudowords | | | | | | | | | | | |
| Pseudowords | 623 | 21 | 602 | 653 | 18 | 631 | .928 | .008 | .910 | 488 | 468 |
| High-frequency | 553 | 23 | 554 | 582 | 15 | 587 | .972 | .008 | .963 | 447 | 457 |
| Low-frequency | 647 | 22 | 653 | 657 | 16 | 675 | .845 | .020 | .821 | 487 | 476 |
| Experiment 4: HF and LF words and random letter strings | | | | | | | | | | | |
| Random letter strings | 555 | 13 | 557 | 563 | 18 | 552 | .962 | .005 | .957 | 422 | 423 |
| High-frequency | 519 | 17 | 512 | 557 | 15 | 550 | .962 | .006 | .957 | 429 | 430 |
| Low-frequency | 546 | 20 | 575 | 595 | 16 | 597 | .945 | .006 | .928 | 443 | 439 |
| Experiment 5: HF, LF, and VLF words and pseudowords (high proportion of HF words) | | | | | | | | | | | |
| Pseudowords | 713 | 29 | 691 | 714 | 20 | 713 | .940 | .009 | .914 | 513 | 498 |
| High-frequency | 595 | 31 | 575 | 618 | 20 | 606 | .967 | .003 | .987 | 466 | 470 |
| Low-frequency | 715 | 31 | 702 | 707 | 23 | 724 | .895 | .011 | .902 | 507 | 598 |
| Very low-frequency | 741 | 25 | 776 | 768 | 27 | 788 | .782 | .018 | .785 | 518 | 509 |
| Experiment 6: HF and VLF words and pseudowords (high proportion of VLF words) | | | | | | | | | | | |
| Pseudowords | 775 | 51 | 750 | 744 | 27 | 770 | .912 | .025 | .910 | 549 | 544 |
| High-frequency | 575 | 34 | 565 | 649 | 22 | 616 | .961 | .010 | .955 | 499 | 501 |
| Very low-frequency | 770 | 44 | 739 | 749 | 29 | 761 | .831 | .030 | .835 | 548 | 534 |

*Note.* RT = response time; HF = high-frequency; LF = low-frequency; VLF = very low-frequency.

**Table 4**

Fast and Slow Subjects' Mean RTs for Experiments 1, 3, and 5

| | Fast subjects | | | Slow subjects | | |
|---|---|---|---|---|---|---|
| **Type of stimulus** | **Probability correct** | **Correct RT (ms)** | **Error RT (ms)** | **Probability correct** | **Correct RT (ms)** | **Error RT (ms)** |
| Pseudowords | .925 | 614 | 601 | .937 | 728 | 748 |
| High-frequency | .949 | 543 | 505 | .972 | 628 | 623 |
| Low-frequency | .841 | 614 | 594 | .912 | 711 | 753 |

*Note.* RT = response time.

## Table 5

RT Distributions for Correct Responses, Experiments 1–6

| Type of stimulus | Data (quantile RTs; ms) | | | | | Model (quantile RTs; ms) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | .1 | .3 | .5 | .7 | .9 | .1 | .3 | .5 | .7 | .9 |
| Experiment 1: HF, LF, and VLF words and pseudowords | | | | | | | | | | |
| Pseudowords | 492 | 556 | 613 | 691 | 884 | 484 | 550 | 608 | 690 | 869 |
| High-frequency | 453 | 501 | 542 | 591 | 710 | 458 | 512 | 554 | 602 | 701 |
| Low-frequency | 487 | 547 | 597 | 670 | 841 | 479 | 544 | 601 | 681 | 858 |
| Very low-frequency | 497 | 565 | 632 | 717 | 912 | 489 | 561 | 629 | 729 | 950 |
| Experiment 2: HF, LF, and VLF words and random letter strings | | | | | | | | | | |
| Random letter strings | 432 | 489 | 536 | 597 | 745 | 432 | 489 | 538 | 604 | 749 |
| High-frequency | 433 | 482 | 526 | 575 | 684 | 434 | 486 | 528 | 580 | 683 |
| Low-frequency | 446 | 502 | 552 | 616 | 762 | 443 | 501 | 551 | 618 | 761 |
| Very low-frequency | 451 | 511 | 566 | 640 | 817 | 448 | 510 | 566 | 642 | 811 |
| Experiment 3: HF and LF words and pseudowords | | | | | | | | | | |
| Pseudowords | 488 | 557 | 612 | 691 | 865 | 468 | 536 | 596 | 677 | 844 |
| High-frequency | 447 | 505 | 550 | 608 | 746 | 457 | 518 | 567 | 627 | 746 |
| Low-frequency | 487 | 557 | 618 | 700 | 875 | 476 | 552 | 624 | 727 | 946 |
| Experiment 4: HF and LF words and random letter strings | | | | | | | | | | |
| Random letter strings | 422 | 475 | 523 | 585 | 737 | 423 | 474 | 521 | 585 | 721 |
| High-frequency | 429 | 481 | 523 | 577 | 724 | 430 | 480 | 525 | 584 | 705 |
| Low-frequency | 443 | 502 | 551 | 617 | 790 | 439 | 498 | 556 | 635 | 809 |
| Experiment 5: HF, LF, and VLF words and pseudowords (high proportion of HF words) | | | | | | | | | | |
| Pseudowords | 513 | 588 | 658 | 752 | 987 | 498 | 580 | 655 | 760 | 984 |
| High-frequency | 466 | 521 | 571 | 636 | 812 | 470 | 535 | 586 | 646 | 766 |
| Low-frequency | 507 | 583 | 651 | 745 | 972 | 498 | 581 | 659 | 769 | 1,006 |
| Very low-frequency | 518 | 613 | 699 | 826 | 1,087 | 509 | 602 | 698 | 839 | 1,143 |
| Experiment 6: HF and VLF words and pseudowords (high proportion of VLF words) | | | | | | | | | | |
| Pseudowords | 549 | 631 | 708 | 813 | 1,076 | 544 | 622 | 695 | 797 | 1,026 |
| High-frequency | 499 | 561 | 609 | 690 | 876 | 501 | 562 | 610 | 669 | 799 |
| Very low-frequency | 548 | 629 | 714 | 828 | 1,099 | 534 | 616 | 697 | 820 | 1,102 |

*Note.* RT = response time; HF = high-frequency; LF = low-frequency; VLF = very low-frequency.

**Table 6**

Diffusion Model Best-Fitting Parameters for Experiments 1–7

| Experiment | $a$ | $T_{er}$ | $\eta$ | $S_z$ | $v_r$ | $v_p$ | $v_h$ | $v_l$ | $v_v$ | $s_t$ | $z$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.110 | 0.435 | 0.070 | 0.004 | | −0.213 | 0.396 | 0.216 | 0.128 | 0.159 | 0.056 |
| 2 | 0.125 | 0.407 | 0.123 | 0.076 | −0.358 | | 0.477 | 0.368 | 0.312 | 0.133 | 0.066 |
| 3 | 0.112 | 0.433 | 0.014 | 0.054 | | −0.235 | 0.337 | 0.151 | | 0.164 | 0.056 |
| 4 | 0.120 | 0.394 | 0.101 | 0.072 | −0.350 | | 0.392 | 0.274 | | 0.110 | 0.063 |
| 5 | 0.127 | 0.438 | 0.035 | 0.046 | | −0.198 | 0.368 | 0.183 | 0.104 | 0.179 | 0.063 |
| 6 | 0.130 | 0.466 | 0.081 | 0.004 | | −0.215 | 0.360 | | 0.138 | 0.171 | 0.060 |
| 7 | 0.131 | 0.402 | 0.086 | 0.025 | −0.401 | −0.164 | 0.390 | 0.255 | 0.158 | 0.146 | 0.063 |

*Note.* $a$ = boundary separation; $T_{er}$ = nondecision component of response time (RT); $\eta$ = standard deviation in drift rate across trials; $s_z$ = range of distribution of starting point; $v_r$ = drift rate for random letter strings; $v_p$ = drift rate for pseudowords; $v_h$ = drift rate for high-frequency words; $v_l$ = drift rate for low-frequency words; $v_v$ = drift rate for very low-frequency words; $s_t$ = range of distribution in the nondecision component of RT; $z$ = starting point.

**Table 7**

Summary of Results and Fits of the Diffusion Model for Experiment 7

| | Error mean RT (ms) | | | Correct mean RT (ms) | | | Probability correct | | | Correct RT at .1 quantile (ms) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Type of stimulus** | **Data** | ***SE*** | **Model** | **Data** | ***SE*** | **Model** | **Data** | ***SE*** | **Model** | **Data** | **Model** |
| Random letter strings | 568 | 12 | 546 | 579 | 13 | 576 | .970 | .008 | .987 | 443 | 450 |
| Pseudowords | 708 | 29 | 754 | 743 | 19 | 716 | .867 | .015 | .841 | 501 | 486 |
| High-frequency | — | — | 621 | 575 | 13 | 571 | .980 | .007 | .989 | 443 | 444 |
| Low-frequency | 673 | 24 | 715 | 635 | 16 | 641 | .940 | .007 | .946 | 467 | 462 |
| Very low-frequency | 798 | 30 | 777 | 680 | 22 | 705 | .838 | .012 | .851 | 477 | 477 |

*Note.* Dashes indicate that there were no observations for many of the subjects. RT = response time.

## Table 8

RT Distributions for Correct Responses for Experiments 7–9

| Type of stimulus | Data (quantile RTs; ms) | | | | | Model (quantile RTs; ms) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | .1 | .3 | .5 | .7 | .9 | .1 | .3 | .5 | .7 | .9 |
| Experiment 7: HF, LF, and VLF words, random letter strings, and pseudowords | | | | | | | | | | |
| Random letter strings | 443 | 495 | 542 | 600 | 746 | 450 | 506 | 550 | 607 | 725 |
| Pseudowords | 501 | 592 | 680 | 798 | 1,061 | 486 | 565 | 646 | 765 | 1,033 |
| High-frequency | 443 | 497 | 537 | 593 | 730 | 444 | 499 | 544 | 599 | 720 |
| Low-frequency | 467 | 530 | 586 | 659 | 857 | 462 | 528 | 588 | 675 | 874 |
| Very low-frequency | 477 | 550 | 619 | 718 | 961 | 477 | 553 | 632 | 752 | 1,023 |
| Experiment 8: HF, LF, and VLF words and pseudowords (1 and 2 presentations) | | | | | | | | | | |
| Pseudowords (1) | 506 | 574 | 632 | 716 | 940 | 495 | 560 | 620 | 705 | 894 |
| Pseudowords (2) | 509 | 577 | 636 | 721 | 942 | 496 | 563 | 625 | 714 | 912 |
| High-frequency (1) | 472 | 525 | 570 | 635 | 779 | 479 | 535 | 581 | 641 | 769 |
| Low-frequency (1) | 512 | 580 | 639 | 725 | 913 | 498 | 566 | 631 | 726 | 936 |
| Very low-frequency (1) | 520 | 609 | 677 | 772 | 984 | 506 | 580 | 657 | 770 | 1,020 |
| High-frequency (2) | 462 | 509 | 552 | 602 | 736 | 474 | 527 | 569 | 621 | 729 |
| Low-frequency (2) | 487 | 543 | 591 | 653 | 840 | 488 | 550 | 603 | 678 | 843 |
| Very low-frequency (2) | 503 | 565 | 616 | 694 | 890 | 496 | 563 | 626 | 717 | 919 |
| Experiment 9: HF, LF, and VLF words and random letter strings (1 and 2 presentations) | | | | | | | | | | |
| Random letter strings (1) | 417 | 470 | 517 | 577 | 730 | 417 | 470 | 517 | 582 | 727 |
| Random letter strings (2) | 418 | 472 | 516 | 577 | 721 | 417 | 470 | 517 | 582 | 727 |
| High-frequency (1) | 415 | 465 | 506 | 564 | 688 | 420 | 470 | 512 | 567 | 684 |
| Low-frequency (1) | 431 | 487 | 533 | 593 | 735 | 425 | 479 | 528 | 593 | 737 |
| Very low-frequency (1) | 432 | 491 | 548 | 622 | 802 | 430 | 488 | 543 | 620 | 790 |
| High-frequency (2) | 412 | 464 | 507 | 554 | 676 | 419 | 468 | 509 | 562 | 674 |
| Low-frequency (2) | 421 | 471 | 516 | 570 | 696 | 423 | 475 | 521 | 581 | 712 |
| Very low-frequency (2) | 426 | 482 | 526 | 582 | 724 | 424 | 476 | 523 | 585 | 719 |

*Note.* RT = response time; HF = high-frequency; LF = low-frequency; VLF = very low-frequency. (1) = 1 presentation; (2) = 2 presentations.

## Table 9

Summary of Results and Fits of the Diffusion Model for Experiments 8 and 9

| | Error mean RT (ms) | | | Correct mean RT (ms) | | | Probability correct | | | Correct RT at .1 quantile (ms) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Type of stimulus** | **Data** | ***SE*** | **Model** | **Data** | ***SE*** | **Model** | **Data** | ***SE*** | **Model** | **Data** | **Model** |
| Experiment 8: HF, LF, and VLF words and pseudowords (1 and 2 presentations) | | | | | | | | | | | |
| Pseudowords (1) | 729 | 34 | 684 | 683 | 13 | 666 | .940 | .009 | .902 | 506 | 494 |
| Pseudowords (2) | 726 | 38 | 692 | 686 | 13 | 674 | .921 | .011 | .887 | 509 | 496 |
| High-frequency (1) | 607 | 31 | 625 | 610 | 16 | 608 | .966 | .005 | .970 | 472 | 479 |
| High-frequency (2) | — | — | 602 | 586 | 16 | 590 | .971 | .007 | .982 | 462 | 474 |
| Low-frequency (1) | 668 | 33 | 708 | 684 | 17 | 683 | .865 | .017 | .865 | 512 | 498 |
| Low-frequency (2) | 669 | 39 | 664 | 632 | 17 | 642 | .928 | .012 | .937 | 487 | 488 |
| Very low-frequency (1) | 707 | 34 | 741 | 721 | 19 | 720 | .751 | .021 | .752 | 520 | 506 |
| Very low-frequency (2) | 725 | 33 | 700 | 660 | 19 | 676 | .876 | .017 | .881 | 503 | 496 |
| Experiment 9: HF, LF, and VLF words and random letter strings (1 and 2 presentations) | | | | | | | | | | | |
| Random letter strings (1) | 537 | 16 | 554 | 556 | 12 | 558 | .953 | .006 | .949 | 417 | 418 |
| Random letter strings (2) | 534 | 20 | 555 | 554 | 11 | 559 | .949 | .008 | .949 | 418 | 418 |
| High-frequency (1) | 523 | 24 | 508 | 541 | 12 | 545 | .968 | .005 | .961 | 415 | 420 |
| High-frequency (2) | 489 | 16 | 504 | 535 | 12 | 541 | .961 | .005 | .965 | 412 | 419 |
| Low-frequency (1) | 521 | 21 | 535 | 564 | 12 | 568 | .947 | .007 | .936 | 431 | 425 |
| Low-frequency (2) | 490 | 17 | 523 | 545 | 11 | 557 | .952 | .005 | .948 | 421 | 423 |
| Very low-frequency (1) | 556 | 19 | 563 | 592 | 13 | 590 | .917 | .006 | .901 | 432 | 430 |
| Very low-frequency (2) | 523 | 16 | 526 | 558 | 13 | 561 | .938 | .007 | .945 | 426 | 424 |

*Note*. Dashes indicate that there were no observations for many of the subjects. RT = response time; HF = high-frequency; LF = low-frequency; VLF = very low-frequency; (1) = 1 presentation; (2) = 2 presentations.

**Table 10**

Diffusion Model Best-Fitting Parameters for Experiments 8 and 9

| Experiment | $a$ | $T_{er}$ | $\eta$ | $s_z$ | $v_r$ | $v_p$ | $v_h$ | $v_l$ | $v_v$ | $s_t$ | $z$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8, 1st presentation | 0.115 | 0.441 | 0.066 | 0.024 | | −0.218 | 0.334 | 0.181 | 0.109 | 0.003 | 0.057 |
| 8, 2nd presentation | 0.115 | 0.441 | 0.066 | 0.024 | | −0.203 | 0.384 | 0.260 | 0.195 | 0.003 | 0.057 |
| 9, 1st presentation | 0.117 | 0.389 | 0.076 | 0.061 | −0.320 | | 0.394 | 0.325 | 0.267 | 0.122 | 0.058 |
| 9, 2nd presentation | 0.117 | 0.389 | 0.076 | 0.061 | −0.320 | | 0.408 | 0.355 | 0.346 | 0.122 | 0.058 |

*Note*. $a$ = boundary separation; $T_{\mathrm{er}}$ = nondecision component of response time (RT); $\eta$ = standard deviation in drift rate across trials; $s_z$ = range of distribution of starting point; $v_r$ = drift rate for random letter strings; $v_p$ = drift rate for pseudowords; $v_h$ = drift rate for high-frequency words; $v_t$ = drift rate for low-frequency words; $v_v$ = drift rate for very low-frequency words; $s_t$ = range of distribution in the nondecision component of RT; $z$ = starting point.

## Table 11

RT Distributions for Error Responses

| Type of stimulus | Data (quantile RTs; ms) .1 | .3 | .5 | .7 | .9 | Model (quantile RTs; ms) .1 | .3 | .5 | .7 | .9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Experiment 1: HF, LF, and VLF words and pseudowords | | | | | | | | | | |
| Pseudowords | 467 | 548 | 629 | 747 | 1,008 | 485 | 555 | 619 | 712 | 917 |
| Low-frequency | 480 | 545 | 627 | 723 | 952 | 479 | 554 | 625 | 719 | 924 |
| Very low-frequency | 476 | 553 | 624 | 732 | 984 | 498 | 576 | 651 | 763 | 1,005 |
| Experiment 3: HF and LF words and pseudowords | | | | | | | | | | |
| Pseudowords | 444 | 517 | 585 | 701 | 891 | 450 | 514 | 566 | 640 | 803 |
| High-frequency | 429 | 465 | 526 | 570 | 731 | 437 | 492 | 536 | 587 | 697 |
| Low-frequency | 464 | 531 | 597 | 693 | 906 | 463 | 535 | 601 | 700 | 916 |
| Experiment 4: HF and LF words and random letter strings | | | | | | | | | | |
| Random letter strings | 406 | 436 | 481 | 534 | 661 | 416 | 464 | 510 | 583 | 764 |
| High-frequency | 394 | 432 | 481 | 588 | 766 | 401 | 443 | 478 | 531 | 672 |
| Low-frequency | 412 | 446 | 498 | 576 | 771 | 408 | 455 | 498 | 572 | 762 |
| Experiment 5: HF, LF, and VLF words and pseudowords (high proportion of HF words) | | | | | | | | | | |
| Pseudowords | 465 | 543 | 637 | 758 | 1,056 | 483 | 560 | 632 | 737 | 970 |
| High-frequency | 444 | 502 | 548 | 618 | 757 | 451 | 509 | 555 | 606 | 713 |
| Low-frequency | 476 | 560 | 635 | 762 | 1,020 | 486 | 565 | 640 | 751 | 997 |
| Very low-frequency | 483 | 577 | 675 | 790 | 1,104 | 503 | 594 | 692 | 839 | 1,162 |
| Experiment 6: HF and VLF words and pseudowords (high proportion of VLF words) | | | | | | | | | | |
| High-frequency | 499 | 551 | 600 | 670 | 856 | 539 | 614 | 682 | 778 | 989 |
| Very low-frequency | 565 | 675 | 777 | 913 | 1,204 | 571 | 670 | 775 | 929 | 1,263 |
| Experiment 7: HF, LF, and VLF words, random letter strings, and pseudowords | | | | | | | | | | |
| Pseudowords | 486 | 562 | 635 | 771 | 1,019 | 457 | 539 | 636 | 792 | 1,151 |
| Low-frequency | 445 | 527 | 588 | 709 | 972 | 457 | 531 | 612 | 738 | 1,037 |
| Very low-frequency | 510 | 629 | 734 | 870 | 1,162 | 478 | 572 | 681 | 849 | 1,221 |
| Experiment 8: HF, LF, and VLF words and pseudowords (1 and 2 presentations) | | | | | | | | | | |
| Pseudowords (1) | 492 | 564 | 649 | 776 | 1,064 | 495 | 563 | 629 | 726 | 943 |
| High-frequency (1) | 488 | 539 | 591 | 626 | 748 | 480 | 539 | 589 | 657 | 807 |
| Low-frequency (1) | 498 | 567 | 617 | 692 | 878 | 501 | 574 | 646 | 754 | 992 |
| Very low-frequency | 508 | 579 | 659 | 750 | 950 | 509 | 587 | 670 | 794 | 1,065 |
| Pseudowords (2) | 492 | 581 | 649 | 787 | 1,028 | 494 | 560 | 619 | 705 | 894 |

*Note.* RT = response time; HF = high-frequency; LF = low-frequency; VLF = very low-frequency. (1) = 1 presentation; (2) = 2 presentations.