Psychon Bull Rev (2017) 24:1949–1956
DOI 10.3758/s13423-017-1259-y

BRIEF REPORT

A diffusion decision model analysis of evidence variability in the lexical decision task

Gabriel Tillman[1] · Adam F. Osth[2] · Don van Ravenzwaaij[1,3] · Andrew Heathcote[1,4]

Published online: 6 March 2017
© Psychonomic Society, Inc. 2017

Abstract The lexical-decision task is among the most commonly used paradigms in psycholinguistics. In both the signal-detection theory and Diffusion Decision Model (DDM; Ratcliff, Gomez, & McKoon, *Psychological Review, 111*, 159–182, 2004) frameworks, lexical-decisions are based on a continuous source of word-likeness evidence for both words and non-words. The Retrieving Effectively from Memory model of Lexical-Decision (REM–LD; Wagenmakers et al., *Cognitive Psychology, 48*(3), 332–367, 2004) provides a comprehensive explanation of lexical-decision data and makes the prediction that word-likeness evidence is more variable for words than non-words and that higher frequency words are more variable than lower frequency words. To test these predictions, we analyzed five lexical-decision data sets with the DDM. For all data sets, drift-rate variability changed across word frequency and non-word conditions. For the most part, REM–LD's predictions about the ordering of evidence variability across stimuli in the lexical-decision task were confirmed.

Keywords Lexical-decision task · Diffusion decision model · REM-LD

Electronic supplementary material The online version of this article (doi:10.3758/s13423-017-1259-y) contains supplementary material, which is available to authorized users.

✉ Gabriel Tillman
gabriel.tillman@newcastle.edu.au

[1] School of Psychology, University of Newcastle, Callaghan, NSW 2308, Australia

[2] Melbourne School of Psychological Sciences, Melbourne, Australia

[3] Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, Netherlands

[4] School of Medicine, University of Tasmania, Hobart, Australia

The lexical-decision task involves identifying letter strings as words or non-words. It has been used extensively in psycholinguistic research to develop cognitive models of reading (e.g., Grainger & Jacobs, 1996; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Norris, 2006). Decisions in the lexical-decision task are typically understood using signal-detection theory (SDT, e.g., Norris, 1986; Balota & Chumbley 1984). In SDT, both words and non-words are assumed to have normally distributed evidence of word-likeness, with observers using a criterion on the evidence axis as a basis for their decision. For example, in the familiarity-recheck model (Balota & Chumbley, 1984), word and non-word stimuli have word-likeness values that represent the familiarity and meaning of letter strings. To make a decision, the observer sets decision criteria along the evidence axis, where word-likeness values lower than the lower criterion result in a non-word response and values higher than the upper criterion result in a word response.

Words of higher natural language frequency are more accurately discriminated and are responded to faster than words of lower frequency. In SDT, this effect occurs because high-frequency (HF) words have a higher mean on the evidence axis than low-frequency (LF) words and non-words. SDT modeling of choice proportions support the assumption that all evidence distributions have equal variance (Jacobs, Graf, & Kinder, 2003; Brown & Steyvers, 2005). However, consideration of response times (RTs) in

addition to choice often yields different conclusions than choice alone (Ratcliff & Starns, 2009), and some work using the Diffusion Decision Model (DDM; Ratcliff, 1978) — a process model that can account for both RTs and accuracy — has suggested higher variability for words than non-words (Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Dutilh, Krypotos, & Wagenmakers, 2011; Dutilh et al., 2012).

In this article, we test the equal evidence variability assumption and demonstrate that our results can be explained by a computational model of lexical retrieval, the Retrieving Effectively from Memory model of Lexical-Decision (REM–LD; Wagenmakers et al., 2004). REM–LD is based on the REM architecture (Shiffrin & Steyvers, 1997), which provides a comprehensive explanation of human memory, accounting for episodic recognition, cued and free recall, perceptual identification, and long-term and short-term priming (Shiffrin, 2003). It provides a detailed account of lexical decision data in "time-controlled" tasks, where participants must respond at a deadline specified by the experimenter. Specifically, the model accounts for word frequency, non-word lexicality, repetition priming, the interaction between these effects, and many other typical findings in the lexical-decision task (see Wagenmakers et al., 2004).

We show that REM–LD predicts that evidence-strength variability differs across stimulus classes, and in particular that variability is highest in HF words, then LF words, then non-words. We use fits of the DDM model to data from "information-controlled" tasks, where responding is under the participant's control, to test these predictions. In what follows, we describe the REM–LD model and obtain predictions about evidence variability. We then describe the DDM model of lexical decision and discuss how it can be used to perform a strong inferential test (Platt, 1964) of REM–LD's predictions about the ordering of variability across item types.

Retrieving effectively from memory – lexical decision

In REM–LD, lexical representations for words are vectors of features, which encode semantic, phonemic, and orthographic information about the words that are experienced. During a lexical-decision task the features of the probe are matched in parallel to features of lexical traces in memory. As the probe is perceived for longer, more features are sampled in the probe vector to cue lexical memory. Not having all features available for the matching process results in the existence of mismatching features as well as matching features, even when the probe is the same as the trace. The probe is matched against each trace in lexical memory and a ratio is calculated reflecting the relative likelihood that the trace is the probe or not. These likelihood ratios are then averaged over all lexical traces, yielding a posterior odds ratio that the probe is present in lexical memory. For non-words, none of the traces yield a strong match to the probe vector, resulting in a low posterior odds of responding word. The trace corresponding to the probe vector yields a strong match, which increases the posterior odds of a word response. Each time a word is encountered in natural language more features of the probe are encoded to the corresponding lexical trace. A discrimination advantage for higher over lower frequency words occurs because higher frequency lexical traces have more features that can match the probe.

Over many trials, the model computes a posterior odds ratio distribution of responding word. The logarithm of these distributions (the log-odds distribution) are normally distributed and are analogous to SDT evidence distributions. For non-words, the variance of the log-odds distribution decreases as the evidence mean increases, so random-letter (RL) strings, which have a lower mean, will have higher evidence variability than pseudo-words (PW), which have a higher mean. For words, on the other hand, the strong match between the probe and one of the lexical traces skews the log-odds distribution, increasing the variability of the distribution relative to the non-word distribution. Because HF words produce a stronger match, this results in a greater skew and hence larger variability relative to LF words. Overall, REM–LD generally predicts lower variability with a larger evidence mean for non-word stimuli and higher variability with a larger evidence mean for word stimuli.

We used the equations from Wagenmakers et al. (2004) to derive deadline predictions of REM–LD about the log-odds distributions of the stimulus classes used in Ratcliff et al. (2004) and Wagenmakers, Ratcliff, Gomez, and McKoon (2008). These classes include words of HF, LF, and very-low-frequency (VLF), along with two types of non-words, PW and RL strings. The parameter settings of the model were based on Wagenmakers et al. (2008) (for details see Supplementary Materials: https://osf.io/9ngaw).

The mean (top left panel) and standard deviation (SD; top right panel) of the log-odds distributions are plotted in Fig. 1. We allowed the REM–LD model to respond at 4 different times (i.e., deadlines), which represent how long the system has been processing the stimulus before it responds. For all deadlines greater than 250ms, which was the starting point for the decision process, the SD is largest for HF, followed by LF, RL, VLF, and finally PW stimuli. To empirically test REM–LD's predictions we turn to the DDM.

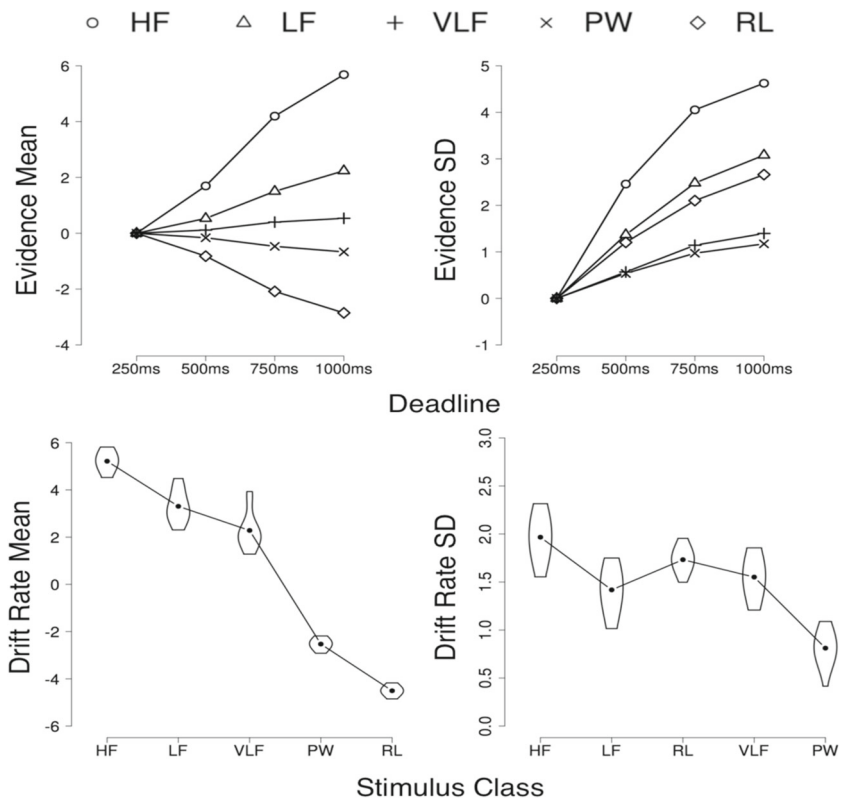

Fig. 1 The *top row plots* the mean (*top left panel*) and SD (*top right panel*) of the log-posterior-odds-ratio distributions from the REM–LD model at four different deadlines (250 ms, 500 ms, 750 ms, 100 0ms). HF = high-frequency, LF = low-frequency, VLF = very-low-frequency, and PW = pseudo-word, and RL = random letter strings. The minimum processing time was 250 ms, and the rate of increase in probability of activation was .0025. The probability of a feature match when encoding the same item was HF = .85, LF = .75, and VLF = .65. The probability of a feature match when encoding a different item was PW = .5, and RL = .35. The *bottom row plots* the drift-rate mean (v, *bottom left panel*) and SD (η, *bottom right panel*) group-level mean posterior distributions from DDM fits to the five lexical-decision experiments. For visualization purposes, the distributions for each stimulus class are the concatenation of the posterior distributions across all 5 experiments. The posterior distributions are displayed as violin plots, which show the median of the posterior (*black dot*) and a rotated kernel density mirrored on either side. The violin plots are truncated to contain the 95% highest density interval. The stimulus class labels along the x-axis are ordered from *left-right* in the same order as REM–LD's predicted ordering from highest to lowest

The diffusion decision model

Simple elaborations of SDT (e.g., Balota & Spieler 1999) do not correctly predict the shapes of RT distributions (Yap, Balota, Cortese, & Watson, 2006) in an information controlled lexical-decision task. For this reason, researchers have used the DDM (Ratcliff et al., 2004) an evidence accumulation model that can account for both choice proportion and RTs in information-controlled responding.

In the DDM, decisions between two alternatives are based on the accumulation of evidence from a stimulus until one of two decision boundaries is reached. Evidence begins to accumulate at the starting point z, which is sampled from a uniform distribution with width s_z. Evidence accumulation is noisy within a trial, and has a mean rate, the "drift-rate", that is sampled from a normal distribution with mean v and standard deviation η on each trial. Ratcliff (1978) introduced inter-trial drift-rate variability to model item differences in a recognition-memory task and this variability is analogous to the continuously distributed variability of evidence in SDT (Ratcliff, 1978, 1985).

Evidence accumulates until it hits either an upper boundary (a, corresponding to a 'word' response) or a lower boundary (0, corresponding to a 'nonword' response). The boundary that is reached first determines the decision, and the time taken to reach the boundary is the decision time. Non-decision time, T_{er}, which quantifies the time taken to encode stimuli and execute a motor response, is estimated as the remainder of each RT. Non-decision time is assumed to have a uniform distribution with range s_t. Figure 2 illustrates the DDM account of lexical-decision.

Using the DDM, Ratcliff et al. (2004) modeled the effects of different stimulus classes in the lexical-decision task using differences in mean drift-rate alone. Higher

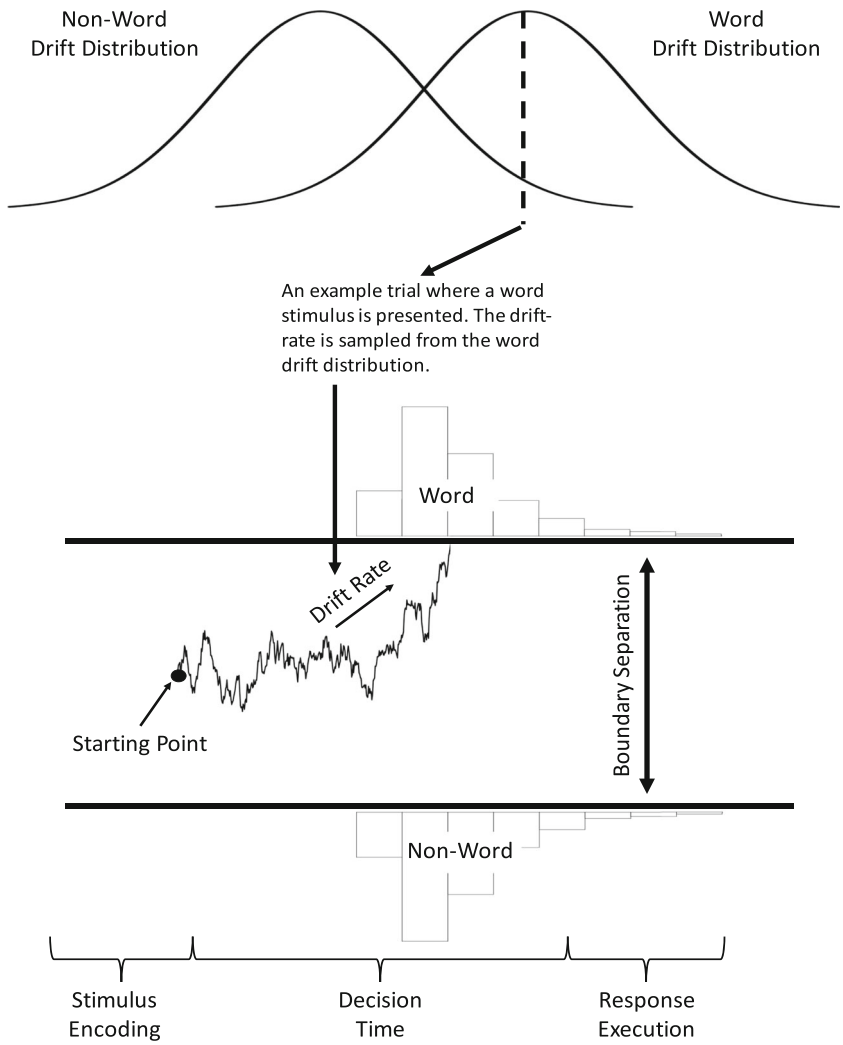

Fig. 2 DDM conceptualization of a two choice decision between 'word' and 'non-word' in the lexical-decision task. The *top panel* shows distributions of drift-rates across trials for both words and non-words. The *bottom panel* shows an example accumulation path for a trial with the displayed drift rate

frequency words had larger drift-rates, which accounted for their greater accuracy and faster RT. However, all item types were assumed to have the same drift-rate variability.

Testing the predictions of REM–LD with DDM

The log-posterior odds ratio distributions of REM–LD and the inter-trial drift rate distributions are both analogous to the evidence distributions of SDT. Under this assumption, we will test the predictions made by REM–LD about the SD of evidence distributions in a qualitative way. Using model selection methods, we test whether a DDM model with separate evidence variability (η) parameters for each stimulus class accounts for data better than a model with only one η for all stimulus classes. Then we will extract the estimates of rate means (v) and η from the former model, and compare them to REM–LD's predicted ordering.

DDM analysis

We used hierarchical Bayesian methods to estimate the parameters of the DDM; the fitting routine, the specific model parameterization for each data set, and the results of a parameter-recovery study validating our estimates are provided in Supplementary Materials.

Data sets

Table 1 provides data set details. All data sets contained a word frequency manipulation. In data set 1 (Experiment 1; Wagenmakers et al., 2008), participants were instructed to respond either quickly or accurately. Data set 2 (Experiment 2; Wagenmakers et al., 2008) contained 25% and 75% proportions of word stimuli. Data sets 3–5 (Experiments 1, 2, and 4, respectively, from

Table 1 Data sets

Data Set	Source	N	Obs.	Variables
1	Wagenmakers et al. (2008) Exp1.	17	1844	Emphasis (Speed or Accuracy)
				Word Frequency (high, low, very low, pseudo-words)
2	Wagenmakers et al. (2008) Exp.2	19	1915	Proportion (25% Word or 75% Word)
				Word Frequency (high, low, very low, pseudo-words)
3	Ratcliff et al. (2004) Exp.1	16	2057	Word Frequency (high, low, very low, pseudo-words)
4	Ratcliff et al. (2004) Exp.2	14	2070	Word Frequency (high, low, very low, random letter strings)
5	Ratcliff et al. (2004) Exp.4	17	1477	Word Frequency (high, low, random letter strings)

N = number of participants; Obs. = the mean observations for each participant

Ratcliff et al., 2004) all contained a word frequency manipulation, but changed the characteristics of the non-word stimuli, using either pseudo-words (pronounceable letter strings in data-set 3, and created by randomly replacing all vowels in words by other vowels, in data sets 1 and 2), or unpronounceable random-letter strings (data-sets 4 and 5).

Model selection

We selected among models using WAIC, a measure of out-of-sample prediction error (Watanabe, 2010; Gelman, Hwang, & Vehtari, 2014), where lower values indicate better out-of-sample prediction. We compared two versions of the DDM: the "equal model", which had one η parameter for all different stimulus conditions, and the "unequal model", which had a separate η parameter for each stimulus condition. Table 2 shows that the unequal model is preferred for data sets 4–5 (unequal model having WAIC values 6.9 and 7.5 less than the equal model, respectively) and strongly preferred for data sets 1–3 (unequal model having WAIC values 297, 125, and 176 less than the equal model, respectively). Note that a difference in WAIC of greater than 3 provides positive evidence and a difference of 10 or more strong evidence, and so, every data set provided either positive or strong evidence for the unequal model. We now examine whether the preferred models provide a good account of the data.

Model fit

We checked fit by generating posterior-predictive data from the unequal models, simulating 100 data sets of the same size as the empirical data from 100 parameter-vector samples from joint-posterior distributions for each participant in each experiment. Figure 3 plots summaries of the observed and predicted data. To summarize the RT distributions, we present five quantile estimates (10%, 30%, 50%, 70% and 90%). The 10%, 50%, and 90% quantiles represent the leading edge, median, and tail of the distribution, respectively. These plots also indicate the proportion of correct (green) and incorrect (red) responses along the y-axis.

The top two panels in Fig. 3 show empirical and predicted values for data sets 1 and 2 from Wagenmakers et al. (2008); the unequal model fits both well. The middle two panels and the bottom panel of Fig. 3 displays the same for data sets 3–5 (Ratcliff et al., 2004). The fits are good except for consistent misses of the tail of the error RT distribution. This misfit is likely due to the low rate of errors and relatively high variability in the 90% quantile for error RTs. However, the key finding that error RTs are on average slower than correct RTs is captured well.

Drift rate parameters

The bottom panel of Fig. 1 shows the mean of the posterior distributions of the group-level mean drift-rate and η

Table 2 WAIC results for the equal and word frequency DDMs

Data set	Source	Equal model	Unequal model	Equal - unequal
1	Wagenmakers et al. (2008) Exp1.	−16549	**−16846**	297
2	Wagenmakers et al. (2008) Exp.2	−21897	**−22022**	125
3	Ratcliff et al. (2004) Exp.1	−20441	**−20617**	176
4	Ratcliff et al. (2004) Exp.2	−33395	**−33402**	6.9
5	Ratcliff et al. (2004) Exp.4	−29059	**−29067**	7.5

Bold WAIC values indicate the preferred model for the corresponding data set

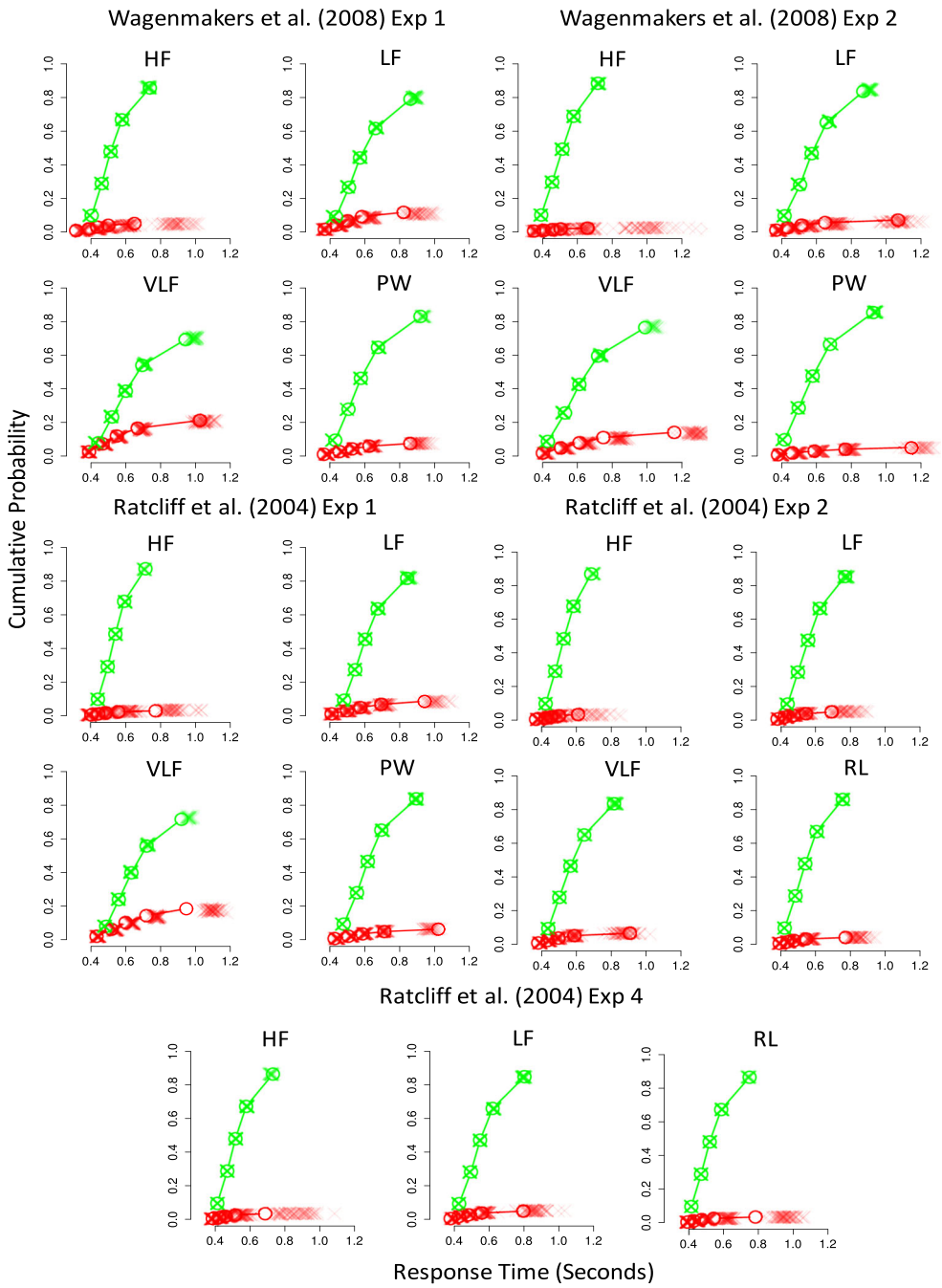

Fig. 3 Defective cumulative distribution plots of the predicted RTs from the unequal model and empirical RTs for each stimulus condition. HF = high-frequency, LF = low-frequency, VLF = very-low-frequency, and PW = pseudo-word, and RL = random letter strings. The *circles* represent the empirical data and the crosses represent the predicted data. Note the predicted data consists of 100 separate data sets superimposed on the empirical data. The *green points* are correct responses and the *red points* are incorrect responses

estimates from the unequal DDM. For visualization purposes, the distributions for each stimulus class are the concatenation of the posterior distributions across all 5 experiments. The ordering of mean drift-rates are in agreement with REM–LD's evidence means within words and within non-words. The ordering of η for the DDM is mostly in agreement with the evidence variability predictions of REM–LD, with the exception of LF words. Drift variability was highest for HF words, followed by RL, VLF, LF, then PW. REM–LD predicted that evidence variability was highest for HF words, followed by LF, RL, VLF then PW.

We used Bayesian predictive p-values to assess the probability that the difference between two posteriors is equal to or less than 0 (Meng, 1994). Small p-values in Table 3

Table 3 Bayesian predictive *p*-values for drift variance ordering

Data Set	HF ≤ LF	HF ≤ RL	HF ≤ VLF	HF ≤ PW	LF ≤ RL	LF ≤ VLF	LF ≤ PW	RL ≤ VLF	VLF ≤ PW
1	< .001	–	< .001	< .001	–	.985	< .001	–	< .001
2	< .001	–	.01	< .001	–	.972	.008	–	< .001
3	.012	–	.040	< .001	–	.679	< .001	–	.001
4	.011	.1	.015	–	.875	.556	–	.556	–
5	< .001	.074	–	–	.978	–	–	–	–

Low *p*-values suggest that the DDM and REM–LD are in agreement

suggest that the DDM and REM-LD are in agreement and larger *p*-values suggest that the two models are in disagreement. They mostly agree, except in regards to LF, RL, and VLF stimuli. The predicted order is reversed between LF and RL, and either reversed or equivocal between VLF and LF or RL.

General discussion

The lexical-decision task has often been conceptualized as a specific case of signal detection theory (SDT; Norris, 1986; Balota and Chumbley, 1984), with decisions based on a continuously distributed evidence variable (i.e., word-likeness). The outcomes of decisions depend on both the mean and variance of evidence, but previous studies have assumed that these evidence distributions are equally variable for words and non-words (Ratcliff et al., 2004; Wagenmakers et al., 2008) with some supporting evidence from choice data (Jacobs et al., 2003; Brown & Steyvers, 2005). This implies that performance depends purely on evidence-distribution means. However, the latter investigations did not consider response times (RTs), which could potentially support different conclusions (Ratcliff & Starns, 2009). This turned out to be the case, with our analysis based on both RTs and accuracy clearly rejecting the equal variance assumption (see also Dutilh et al., 2009, 2011, 2012). These results imply that researchers should take account of factors that affect the variability in evidence as well as its mean. For example, number of letters, orthographic neighborhood size, average base-word frequency, and average base-word number of syllables are factors known to affect between-item variability in response times and accuracy (Yap, Sibley, Balota, Ratcliff, & Rueckl, 2015). Estimation of inter-trial drift variability is sensitive to variability in RT and accuracy, and so, it seems likely that these item level differences will be influential on the magnitude of inter-trial evidence (i.e., drift rate) variability.

We also investigated the Retrieving Effectively from Memory model of Lexical-Decision (REM–LD), which has previously been used to account for data from a deadline lexical-decision task. REM–LD is based on a general model architecture that provides a comprehensive explanation of human memory. In REM–LD, stronger matches between the probe and trace skew the evidence distribution, which produces greater evidence variability for words than non-words, particularly for higher frequency words. Using typical parameter settings, we showed that REM–LD makes the prediction that the evidence variability will be largest for high-frequency words, followed by low-frequency, random letter strings, very-low-frequency, and finally pseudo-words.

We fit the Diffusion Decision Model (DDM; Ratcliff, 1978) to free-response lexical-decision data and examined the parameter estimates of inter-trial drift rate variability, which is analogous to evidence variability in SDT and REM–LD (Ratcliff, 1978, 1985). We found that the predictions of REM–LD were comparable to the DDM's evidence variability estimates for all word frequency conditions except low-frequency words. Specifically, the DDM predicted drift variability was highest for high-frequency words, followed by random letter strings, very-low-frequency, low-frequency, then pseudo-words. Overall, our results are encouraging because two prominent models of lexical-decision mostly agreed about predictions of word and non-word evidence variability.

Evidence variability occurs because items in the same category do not have the same word-likeness value, or in terms of the DDM, the same drift rate. Intuitively, one might assume that higher frequency words are less variable than lower frequency words; perhaps because people might not know the definitions to some lower frequency words, making them more like non-words and inflating the variability. Despite this intuition, we observed that higher frequency words are more variable. Under REM–LD, the reason that higher frequency words are more variable is because of the way lexical retrieval operates by comparing a probe cue to all of the traces in the participant's lexical memory. When the probe cue is a word, it produces a strong match to its own trace and a weak match to all of the other traces in lexical memory. When these matches are averaged together, the contribution from the strong match skews the posterior odds

ratio distribution, producing greater variability for words than non-words and greater variability for higher frequency words relative to lower frequency words.

Our results parallel findings from the recognition memory literature, where inter-trial drift rate variability is higher for studied (i.e., stronger) items (Ratcliff & Starns, 2009; Starns & Ratcliff, 2014; Starns, Ratcliff, & McKoon, 2012; Osth, Dennis, & Heathcote, 2017). Models of recognition memory employ the same retrieval structure as REM–LD and predict higher variability for studied items for a similar reason: recognition is carried out by matching a cue vector against each memory, calculating the similarity, and making a decision based on either the summed or averaged similarity. Findings about evidence variability have played a crucial role in developing a theoretical understanding of recognition memory (Wixted, 2007; Osth & Dennis, 2015; Shiffrin & Steyvers, 1997), and our results suggest that they may play a similar role for theories of lexical memory.

References

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 340–357.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, *128*(1), 32–55.

Brown, S. D., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 587–599.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256.

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E. J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, *16*(6), 1026–1036.

Dutilh, G., Krypotos, A. M., & Wagenmakers, E. J. (2011). Task-related versus stimulus-specific practice. *Experimental Psychology*, *58*(6), 434–442.

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012). Testing theories of post-error slowing. *Attention, Perception, and Psychophysics*, *74*(2), 454–465.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*(3), 518–565.

Jacobs, A. M., Graf, R., & Kinder, A. (2003). Receiver operating characteristics in the lexical decision task: Evidence for a simple signal-detection process simulated by the multiple read-out model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 481–488.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 1142–1160.

Norris, D. (1986). Word recognition: Context effects without priming. *Cognition*, *22*(2), 93–136.

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*(2), 260–311.

Osth, A.F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*, *92*, 101–126.

Platt, J. R. (1964). Certain systematic methods of scientific thinking may produce much more rapid progress than others Platt JR. *Science (New York, NY)*, *146*(3642), 347–53.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review*, *92*(2), 212–225.

Ratcliff, R., Gomez, P., & McKoon, G. M. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159–182.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83.

Shiffrin, R. M. (2003). Modeling memory and perception. *Cognitive Science*, *27*(3), 341–378.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: Rem—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of roc functions: A diffusion model analysis. *Journal of Memory and Language*, *70*, 36–52.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zroc slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*(1), 1–34.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159.

Wagenmakers, E. J., Steyvers, M., Raaijmakers, J. G., Shiffrin, R. M., Van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*(3), 332–367.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152.

Yap, M. J., Balota, D. A., Cortese, M. J., & Watson, J. M. (2006). Single- versus dual-process models of lexical decision performance: Insights from response time distributional analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1324–1344.

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the english lexicon project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 597.