Non-Decision Time Effects in the Lexical Decision Task

Christopher Donkin (Chris.Donkin@newcastle.edu.au)
Andrew Heathcote (Andrew.Heathcote@newcastle.edu.au)
Scott Brown (Scott.Brown@newcastle.edu.au)
Department of Psychology, University of Newcastle, NSW, 2308, Australia
Sally Andrews (sallya@psych.usyd.edu.au)
Department of Psychology, University of Sydney, NSW, 2006, Australia

Abstract

It has been argued that performance in the lexical decision task (LDT) does not provide a direct measure of lexical access because of the effect of decision processes. We re-examine LDT data and fits of the diffusion decision model reported by Ratcliff, Gomez and McKoon (2004) and show that they assumed too little role for non-decision processes in explaining the word frequency effect. Our analysis supports an effect of frequency on decision *and* non-decision time.

Keywords: Lexical decision task; diffusion model

Reading is one of the most remarkable abilities achieved by the human mind. One of the key aspects enabling reading is the ability to recognize a string of characters as being a word, a process called "lexical decision". The lexical decision task (LDT) is a paradigm for studying word identification in which participants are presented with a string of letters and they must quickly decide whether or not the letters form a word. If the letters presented do make a word, then the time taken to make a 'word' response is thought to give information about how long it took to retrieve the word from their database of words, a process referred to as lexical access.

The *word frequency* effect is one of the most robust findings from the LDT paradigm: words used less frequently in natural language take longer to indentify than higher frequency words. Historically, the word frequency effect has been reported as a difference in mean reaction time (RT) for correct responses between low and high frequency words. Mean RT from high and low frequency words usually differs by around 60-80ms. However, RT in the LDT is quite variable, typically having a standard deviation of greater than 100ms. Some of this variability is because of differences between words within a frequency class, but variability also occurs between the same word on different occasions. Variability in RT is also positively skewed, with a longer right (slow) than left (fast) tail in RT distribution, and the length of the right tail has been found to vary systematically in LDT experiments. Hence, researchers have begun to investigate differences in the entire RT distribution between high and low frequency words, rather than just the mean RT (Andrews & Heathcote, 2001; Balota & Spieler, 1999; Plourde & Besner, 1997). More recently, there have been lexical theories proposed that account for effects on all aspects of RT distribution (Ratcliff, Gomez and McKoon, 2004; Yap, Balota, Cortese & Watson, 2006).

RT distributions have been shown to be well characterized by the ex-Gaussian distribution (Luce, 1986). The ex-Gaussian distribution is produced by convolving (i.e., adding samples from) the Gaussian and Exponential distributions. It has three parameters, the mean (μ) and standard deviation (σ) of the Gaussian component and the mean of the exponential component (τ). These parameters give information about the shape of the RT distribution. In particular, the μ parameter is affected by the speed of the fastest responses made by participants. Similarly, the τ parameter is affected by the length of the right tail of the RT distribution.

Differences in parameter estimates from fits of the ex-Gaussian to high and low frequency RT distributions indicate that there are changes in the very fastest and slowest responses made by participants. Changes in μ of approximately 20-30ms have been reported (Andrews & Heathcote, 2001; Balota & Spieler, 1999; Plourde & Besner, 1997). These changes indicate that the entire RT distribution shifts to be slower for less frequent words, independently of any changes in the shape of the distribution. In the same applications of the ex-Gaussian, changes in τ of approximately 35-45ms were observed, suggesting that the right tail is longer when the words to be identified are less frequent.

Balota and Chumbly (1984) argued that the data from LDT tasks come from a combination of the lexical process *and* the decision process. Ratcliff et al. (2004) furthered this line by arguing information about lexical access can only be obtained from RT *after* accounting for the decision process. In other words, even studying the full range of behavioral data in the LDT (i.e., accuracy and RT distributions for correct and error responses) does not by itself provide clear information about lexical access. To address this issue they fit a model of the decision process, the diffusion model, to their LDT data and used estimates of its parameters, and the parameters of a simple characterization of non-decision processes, to examine lexical access. When Yap et al. (2006) compared the diffusion account with a hybrid two-stage model of the LDT based on Balota and Chumbly's work, they concluded in favor of the diffusion model.

The diffusion model account of RT is composed of two parts – a decision time and a non-decision time. The account of LDT starts by assuming that a stimulus is perceived and encoded. This is followed by lexical access, which gives an estimate of how much evidence the stimulus provides for each response (word and non-word in an LDT). This evidence determines the rate at which information is accumulated, called *drift rate*, and drives the decision part of the diffusion model. The time taken for the initial

perceptual, encoding and lexical access processes, plus the time to execute the motor response after the decision process is completed, makes up the non-decision time. The non-decision time, T_{er} in the diffusion model determines the smallest possible RT and, therefore, changes in T_{er} shift the entire RT distribution. The ex-Gaussian evidence reviewed above might have suggested that the word frequency effect would, in part, be explained by differences in T_{er} for high and low frequency words. However, when Ratcliff et al. (2004) applied the diffusion model to data from nine LDT experiments they concluded that only drift rate differed between high and low frequency words. In other words, word frequency effects in the LDT were simply due to how 'wordlike' the string of letters was, and not caused by other aspects of the non-decision processing, such as the time required for lexical access. Ratcliff et al. claimed that the shift of the RT distribution due to word frequency is captured by the inclusion of trial-to-trial variability in T_{er} and not due to systematic differences in T_{er} determined by the frequency of the word being identified.

In the current paper we reanalyze Ratcliff et al.'s (2004) data and demonstrate that their fits of the diffusion model systematically fail to account for the word frequency effect on both fast and slow responses. We then show that the misfit is greatly reduced by allowing T_{er} to differ for words of different frequency. We finish by discussing the implications of our results and possible extensions. First, however, we begin by describing the diffusion model.

The Diffusion Model

The diffusion model with trial-to-trial variability in parameters is the most successful model of choice and reaction time for simple decisions between two alternatives (Ratcliff, 1978) and has been applied repeatedly to LDT data since Ratcliff et al.'s (2004) initial work (Gomez, Ratcliff & Perea, 2007; Ratcliff, Perea, Colangelo, & Buchanan, 2004; Wagenmakers, Ratcliff, Gomez & McKoon, 2008). The diffusion model assumes that participants sample evidence from the stimulus continuously, and this evidence stream updates an evidence total, say x, illustrated as a function of time by the irregular line in Figure 1. The accumulator begins the decision process in some intermediate state, say $x=z$. Evidence that favors the response "word" increases the value of x, and evidence that favors the other response ("non-word") decreases the value of x. The evidence accumulation process continues until sufficient evidence favors one response over the other, causing the total to reach one of its two boundaries (the horizontal lines at $x=0$ and $x=a$ in Figure 1). The choice made by the model depends on which boundary is reached (a for a "word" response or 0 for a "non-word" response) and decision time equals the accumulation time.

Depending on the stimulus, evidence tends to accumulate more towards one boundary or another, and the average rate of this accumulation is called the "drift rate", which we will label v. Larger positive or negative drift rates cause faster and more accurate responses as evidence heads towards the correct boundary at a faster rate. The evidence accumulation process also varies randomly from moment-to-moment during the accumulation process, and the amount of this variability is another parameter of the model, s. The diffusion model used in Ratcliff et al. (2004) also includes three extra variability parameters, the distribution of drift rates is assumed to vary from trial-to-trial according to a normal distribution with mean v and standard deviation η. Start point is also assumed to vary from trial-to-trial according to a uniform distribution with centre z and range s_z. Finally, non-decision time is assumed to vary between trials according to a uniform distribution with centre T_{er} and range s_T. Critically, non-decision variability enables the diffusion model to better account for shifts in RT distribution between conditions that differ only in drift rate. When there is no non-decision variability a change in drift rate almost exclusively slows RT by lengthening the right tail of the distribution, with only a small effect on the fastest RTs. When non-decision variability is added the effect of a drift rate change on fast RTs is increased sufficiently so that Ratcliff et al. (2004) were satisfied with an account of the word frequency effect in terms of a pure selective influence on drift rate.

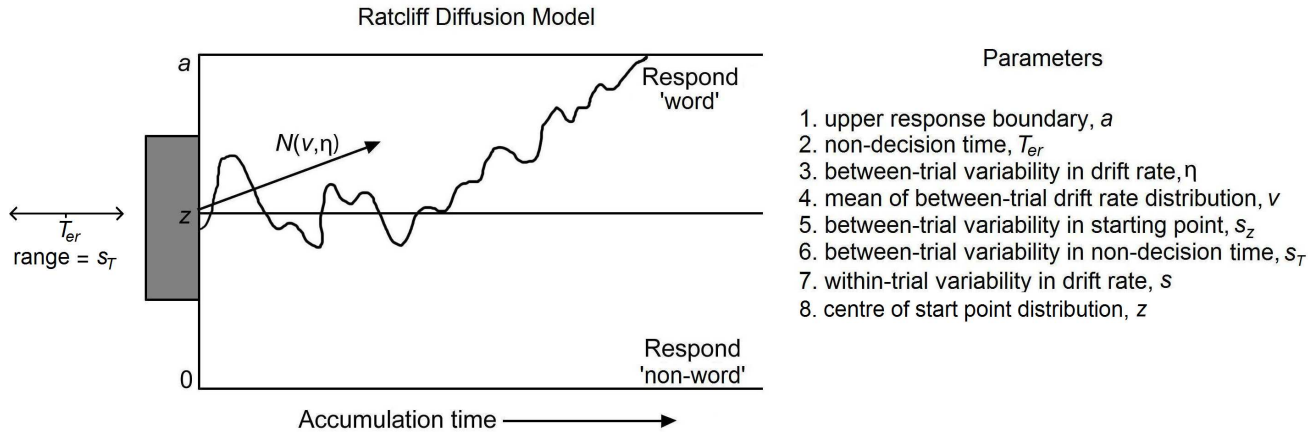

Figure 1: A graphical representation of a single diffusion model decision in an LDT task.

Ratcliff et al.'s (2004) LDT Data

Fits reported in the original paper

Ratcliff et al.'s (2004) fits to all experiments were accomplished by allowing only drift rate to vary between word frequency conditions. This is common practice when applying the diffusion model. Differences in non-decision process parameters cannot be the sole account for word frequency effects, as these processes cannot influence error rates. However, although less parsimonious, there is no reason why non-decision processes might not be affected by word frequency in addition to drift rates. Indeed, Ratcliff et al.'s (2004) application of the diffusion model to the LDT was one of the first occasions on which non-decision variability was used, with most earlier applications assuming a constant non-decision time (e.g., Ratcliff, 1978).

When we looked closely at Ratcliff et al.'s (2004) published fits of the diffusion model to their LDT data averaged over participants, we found a systematic pattern of misfit that was highly consistent across all of the nine experiments which they report. In particular, despite the inclusion of between-trial variability in T_{er}, the diffusion model consistently under-predicted the magnitude of the word frequency effect on the .1 quantile results for correct responses reported by Ratcliff et al.. The .1 quantile characterizes the fastest responses from the RT distribution (i.e., it is the RT below which the fastest 10% of responses occur). Changes in the .1 quantile indicate a shift in the entire RT distribution. Averaging over their nine experiments, the .1 quantile estimate for high frequency words was 27ms and 33ms faster relative to low and very low frequency words respectively, whereas for the model it was only 16ms and 22ms faster. Although the under-prediction is relatively small (11 ms on average), it is highly consistent, occurring in every one of the 19 fits reported in their Tables 3, 7 and 9 - a highly significant result using a binomial test (p<.001 for both low and very low frequency words). In contrast to results for the fast .1 quantile, the diffusion model consistently over-predicted the word frequency effect for the slow .9 quantile, for nine of ten fits comparing high and low frequency words (p<.001) and seven of nine fits comparing high and very low frequency words (p<.02).

Figure 2 is a graphical summary of these analyses of data and model fits for high and low frequency words averaged over experiments from Ratcliff et al. (2004). Though it was excluded for brevity, the plot of the difference between high and very low frequency words looks almost identical. The vertical axis shows the difference in RT between low and high frequency words. Note that the positive value of this difference means that participants were slower to respond to low frequency words – the standard word frequency effect. The horizontal axis represents the quantile values of the RT distribution. The average model predictions (shown by the solid line) for the.1 quantile fall below the observed data averaged across all experiments. Note also that the opposite is true for the .9 quantile – the average model predictions sit higher than the data in both plots. The systematic and opposite misfit for fast and slow responses resulted in over prediction of the effect of word frequency on variability (i.e., a much larger range between the 10% and 90% quantiles than observed in data).

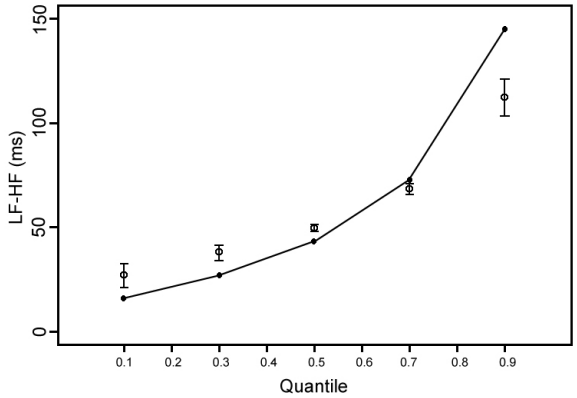

Figure 2: Word frequency effect quantile function based on responses to high frequency (HF) and low frequency (LF) words in Ratcliff et al.'s (2004) experiments 1-9. Average model fits across experiments and conditions are plot as lines, and data as symbols. Standard error bars indicate variability across experiments and condition

The diffusion model has clearly raised the bar for accounts of LDT performance by simultaneously fitting accuracy and RT distribution for both correct and error responses. Although we agree that the diffusion model provides an impressively comprehensive account of many aspects of performance in the LDT, the systematic misfit of the word frequency quantile functions indicates that there may be reason to re-examine the assumptions made by Ratcliff et al. (2004) in their application of the diffusion model.

The diffusion model appears to have misfit Ratcliff et al.'s (2004) data largely because the assumptions underlying the mapping of the diffusion model to the LDT task are too simple. Although simplicity is a virtue in quantitative modeling, identifying word frequency effects entirely with drift rate may represent an over-application of Occam's razor. Most models of reading assume that lexical access is accomplished more quickly as the frequency of a word increases (see Andrews & Heathcote, 2001, for a discussion). In the diffusion model framework, this could be interpreted as a faster non-decision time for high than low frequency words. Allowing for such a possibility might reduce the underestimation of the word frequency effect at the .1 quantile apparent in Figure 2. In other words, perhaps the diffusion model would provide a better account of the word frequency effect in LDT data if it were to also allow for changes in T_{er} for words of different frequency. We explore this possibility in the next section.

Exploring frequency effects on non-decision time

We fit four different versions of the diffusion model to data averaged over participants from Experiments 3, 4 and 5 from Ratcliff et al. (2004). All experiments were of nearly identical procedure, with differences being in the type of words used: Experiment 3 used high frequency, low frequency and pseudo-words, Experiment 4 was identical but used random letter strings instead of pseudo-words, and Experiment 5 was the same as Experiment 3 but also included very-low frequency words. Our re-analyses was limited to these three experiments because Ratcliff et al. did not publish critical information for fitting (e.g., quantiles for error RT) for the remaining experiments.

The four versions of the diffusion models differ according to how non-decision time, T_{er}, varied. There were two ways in which T_{er} was allowed to vary – randomly between trials (cf. Ratcliff et al., 2004) or systematically between word frequency conditions. Between-trial variation was uniformly distributed with mean T_{er} and range s_T. Between-condition variation in T_{er}, like between-condition variation in drift rate, meant that each of the word conditions had its own T_{er} value. The between-trial variability in T_{er} requires one parameter, s_T, whereas between-condition variability in T_{er} requires the estimation of an additional k-1 parameters, where k is the number of word frequency conditions in the experiment being fit. The four different models were factorial combinations of these two methods: 1) neither between-trial nor between-conditions variability in T_{er}, 2) only between-trial variability in T_{er}, 3) only between-conditions variability in $T_{er,}$ and 4) both between-trial and between-conditions variability in T_{er}.

The data to be fit were accuracy and quantile values for correct and error responses averaged over participants from each experiment. We fit the diffusion model using an adaptation of Voss and Voss's (2008) diffusion model code to use quantile maximum likelihood estimation (Heathcote, Brown & Mewhort, 2002). The Bayesian information criterion (BIC) was calculated using the BIC statistic for N observations grouped into bins:

$$BIC = -2(\sum_i Np_i \ln(\pi_i)) + M \ln(N)$$

where p_i is the proportion of observations in the i^{th} bin, and π_i is the proportion of observations in the i^{th} bin as predicted by the model. M is the number of parameters of the model used to generate predictions. The BIC is composed of two parts, the first is a measure of misfit, and a second part, $M \ln(N)$, penalizes a model for its complexity as indicated by the number of estimated parameters. When comparing two models, the model with the smaller BIC is thought to have provided a better fit after complexity has been taken into account. Best fitting parameter estimates for each of the four models to all three experiments and their respective BIC values are given in Table 1.

Despite the complexity of the analysis, the pattern of results was relatively simple. Adding between-trial variability in T_{er} always improved the BIC value, and so too did adding between-condition variability in T_{er}. In all three experiments the model with both between-trial and between-condition variability in T_{er} had the lowest BIC. This implies that the improvement in fit due to the extra free parameters outweighed the penalty for added complexity. The next best fitting model in two out of three experiments was the model used to originally fit the data in Ratcliff et al. (2004) – the model with between-trial variability in T_{er}. In Experiment 5 not the model without between-trail variability in T_{er}, but with between-condition variability in T_{er} achieved the second best fit.

The model with neither between-trial nor between-condition variability in T_{er} consistently performed the worst of the four models. Inspection of the fits revealed that, as expected, this model predicted almost no change in the .1

Table 1: Parameter estimates from fits of four different versions of the diffusion model to Experiments 3-5. M1 was the model with no variability in Ter, M2 had variability between-trials, M3 had variability between-conditions and M4 had both. In all models starting point, z, was set at a/2.

	Model	a	s_z	η	v_h	v_l	v_o	v_v	s_t	T_{er}	T_{er} *HF*	T_{er} *LF*	T_{er} *O*	T_{er} *VLF*	BIC
Exp3	M1	.128	.059	.037	.348	.176	-.226			.404					91887
	M2	.122	.069	.108	.446	.219	-.282		.17	.444					91126
	M3	.127	.065	.052	.335	.188	-.243				.396	.421	.422		91449
	M4	.122	.076	.113	.412	.226	-.301		.16		.428	.451	.461		**90843**
Exp4	M1	.133	.08	.089	.367	.361	-.302			.378					98571
	M2	.126	.075	.101	.381	.361	-.366		.11	.39					98415
	M3	.132	.081	.093	.37	.319	-.358				.379	.391	.375		98453
	M4	.127	.078	.011	.391	.334	-.374		.105		.392	.404	.387		**98320**
Exp5	M1	.147	.069	.069	.354	.214	-.259	.128		.409					89190
	M2	.144	.075	.01	.394	.234	-.253	.141	.139	.431					89000
	M3	.144	.074	.074	.336	.243	-.217	.132			.402	.435	.429	.425	88693
	M4	.148	.093	.124	.404	.257	-.296	.163	.125		.422	.451	.461	.454	**88546**

quantile due to changes in word frequency. Because of this it was also unable to capture other aspects of the RT distribution. Hence, we do not consider the model without variability in T_{er} any further. Although, for brevity, we do not show the complete fits of the model to quantiles for correct and error responses for all word frequency conditions, these graphs clearly agree with our conclusions based on BIC values (they may be obtained by emailing the authors).

Our reason for investigating between-condition variability in T_{er} was based on the systematic misfit of the word frequency effect. Figure 3 shows that there is an improvement in the account of the word frequency effect when between-condition variability in T_{er} is added to the diffusion model. The plots in Figure 3 are like those in Figure 2, but are from individual experiments rather than averaged across all nine experiments in Ratcliff et al. (2004). Each of the three plots also now contains three sets of model predictions (represented by solid lines) rather than one. The filled black dots represent the difference between RTs from high and low frequency words at each of the .1, .3, .5, .7 and .9 quantiles from the data. For all experiments we again observe that the difference between low and high frequency words is positive at all quantile values. This suggests that the RT distribution for low frequency words is shifted above that of high frequency words.

The models with between-condition variability in T_{er} both provide a good account of the word frequency effect, while the model with only within-condition variability in T_{er} still systematically fails to capture the effect. The lines connected by plus signs (+) are the predictions of the diffusion model with only between-trial (within-condition) variability in T_{er} (i.e. the same as the model used in Figure 2 and Ratcliff et al., 2004). Note the systematic under-prediction of the .1 quantile in all experiments, and the over-prediction of the .9 quantile in Experiments 3 and 5. The predictions of the models with between-condition variability in T_{er} or both forms of variability in T_{er} (representing in Figure 3 by lines joined by crosses and triangles, respectively) provide a much better account of the word frequency effect. Indeed, the two models produce an almost identical account of the word frequency effect in Experiments 4 and 5. In these experiments both models provide an excellent account of the difference between RTs from high and low frequency conditions at all quantiles except for the .9 quantile in Experiment 5. In Experiment 3 the model with both types of variability provides and excellent account of all but the .9 quantile, whereas the two other models also provide a less accurate account at three of the four remaining quantiles. Though we do not show it here due to space restrictions, a plot like Figure 3, but comparing high and very low frequency words from Experiment 5, showed the same pattern of results (once again this plot may be obtained by emailing the authors).

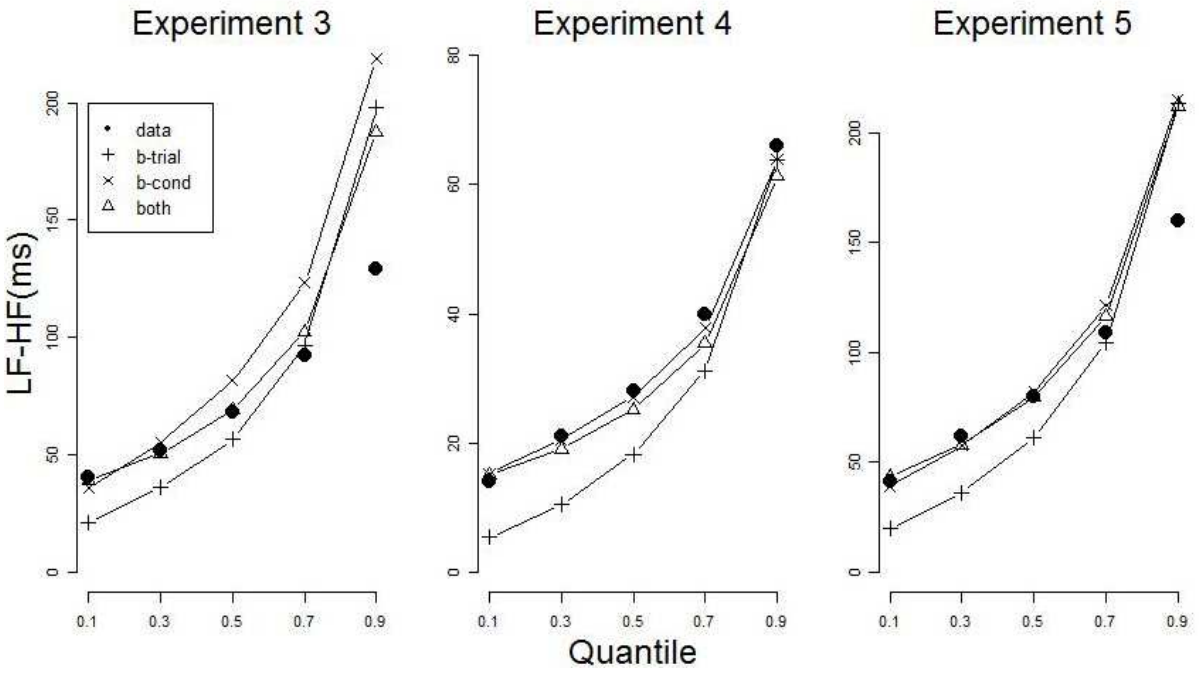

Figure 3: Word frequency effect quantile function based on responses to high frequency (HF) and low frequency (LF) words in Ratcliff et al.'s (2004) experiments 3-5. Data are shown as filled black dots and model predictions from a diffusion model with between-trial variability in T_{er}, a model with between-condition variability in T_{er} and a model with both forms of variability are shown by lines connected with a plus symbol (+), a cross (x), and a triangle, respectively.

Discussion

We were prompted to fit a diffusion model which allowed mean non-decision time (T_{er}) to vary as a function of word frequency because of a) results from previous analyses of RT distribution using the Ex-Gaussian distribution, b) systematic misfit of the word frequency effect by a diffusion model which allows only drift rate to vary between frequency conditions, and c) the fact that a shift is plausible according as most reading models, which assume that word frequency affects the time taken for lexical access. A diffusion model with both between-condition and between-trial variability provided a better fit to the data, even after accounting for this models increased parametric complexity. In particular, the model with both forms of variability provided an improved account of the word frequency effect compared to Ratcliff et al.'s (2004) original model with only between-trial variability in T_{er}, as it did not systematically under-predict the shift in the RT distribution between high and low frequency words.

A diffusion model with between-condition variability in T_{er}, but without between-trial variability in T_{er}, was also able to account for the shift effect. However, in terms of overall fit, this model did worse in two of three experiments than the Ratcliff et al. (2004) original model. A diffusion model with no variability in T_{er} either between-conditions or between-trials fit had a poor overall fit and account of the word frequency effect. These results together suggest that the addition of between-condition variability in T_{er} greatly improves the account of the shift in RT distribution due to changes in word frequency (see also Ratcliff & Tuerlinckx, 2002).

Even the diffusion model with both forms of variability in T_{er} still over-predicted the slowest differences between high and low frequency words in two of the three experiments we examined. This suggests that our current account of the word frequency effect and the LDT may not be complete. Indeed, given the intricacies of the lexicon, an even more complex model of the effects of frequency on non-decision time seems quite plausible and may account for these failings. However, it has been argued that the .9 quantile estimate is much more variable than the other quantile estimates, and most subject to the influence of slow outlier responses, so this misfit is not necessarily indicative of a failed model. An alternative possibility is raised by Donkin, Brown and Heathcote's (submitted) recent demonstration that the moment-to-moment variability parameter has been, without justification, over-constrained in all previous applications of the diffusion model. When we let this parameter vary across frequency conditions BIC improved and excellent fits were obtained to all quantiles of the word frequency effect, and all other aspects of the data. However, due to space restrictions, details concerning these fits will be reported elsewhere.

Acknowledgments

We acknowledge support from an ARC Discovery project grant to Andrews and Heathcote.

References

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Human Perception and Performance, 27,* 514-544.

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

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

Donkin, C., Brown, S., & Heathcoate, A. (submitted). The over-constraint of response time models. *Psychological Review*.

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology*: *General*, *136*, 389-413

Heathcote, A., Brown, S.D. & Mewhort, D.J.K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin and Review*, *9*, 394-401

Luce, R. D. (1986) Response times: Their role in inferring elementary mental organization. NY: Oxford University Press.

Plourde, C. E., & Besner, D. (1997). On the locus of the word frequency effect in visual word recognition. *Canadian Journal of Experimental Psychology, 51,* 181-194.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 88,* 552-572.

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

Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers, *Brain and Cognition*, *55*, 374-382.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438-481.

Voss, A., & Voss, J. (2008). A Fast Numerical Algorithm for the Estimation of Diffusion-Model Parameters. *Journal of Mathematical Psychology, 52,* 1-9

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,* 140-159.

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,* 1324-1344.