Journal of Experimental Psychology:
Learning, Memory, and Cognition
2008, Vol. 34, No. 3, 495–513

0278-7393/08/$12.00 DOI: 10.1037/0278-7393.34.3.495

On the Additive Effects of Stimulus Quality and Word Frequency in Lexical Decision: Evidence for Opposing Interactive Influences Revealed by RT Distributional Analyses

Melvin J. Yap and David A. Balota
Washington University

Chi-Shing Tse
University at Albany, State University of New York

Derek Besner
University of Waterloo

The joint effects of stimulus quality and word frequency in lexical decision were examined in 4 experiments as a function of nonword type (legal nonwords, e.g., BRONE, vs. pseudohomophones, e.g., BRANE). When familiarity was a viable dimension for word–nonword discrimination, as when legal nonwords were used, additive effects of stimulus quality and word frequency were observed in both means and distributional characteristics of the response-time distributions. In contrast, when the utility of familiarity was undermined by using pseudohomophones, additivity was observed in the means but not in distributional characteristics. Specifically, opposing interactive effects in the underlying distribution were observed, producing apparent additivity in means. These findings are consistent with the suggestion that, when familiarity is deemphasized in lexical decision, cascaded processing between letter and word levels is in play, whereas, when familiarity is a viable dimension for word–nonword discrimination, processing is discrete.

Keywords: additive effects, distributional analysis, stimulus quality, word frequency, lexical decision

One of the most robust and intriguing findings in the visual word recognition literature is the observation that stimulus quality and word frequency produce additive effects in lexical decision performance (Balota & Abrams, 1995; Becker & Killion, 1977; Plourde & Besner, 1997; Stanners, Jastrzembski, & Westbrook, 1975; Yap & Balota, 2007). When discriminating between words and nonwords, participants respond faster to frequently encountered, compared to rarely encountered, words; they also respond faster to visually clear, compared to visually degraded, words. However, these two factors are strongly additive, that is, two main effects with no interaction. Using Sternberg's (1969a) additive-factors logic, Yap and Balota (2007) argued that the additive effects of stimulus quality and word frequency can be interpreted as being consistent with an early encoding stage where stimuli are perceptually normalized and a later stage where lexical retrieval takes place. Perceptual normalization here is defined as a process that refines a degraded representation sufficiently to allow subsequent lexical retrieval processes to work equally efficiently for clear and degraded stimuli (Sternberg, 1969b). In the context of lexical decision, this allows degraded words that have been normalized to be matched to representations stored in memory.

Although separate stages imply additive effects, additive effects do not necessarily imply separate stages. For example, in a cascade model (McClelland, 1979), all processes are operating continuously, with the partial output of one process immediately available as an input for the next process. Importantly, the cascade model does not assume separate stages, and yet is able to produce approximately additive effects. However, we suggest that the additive effects of stimulus quality and word frequency are easier to reconcile with a stage than a cascade architecture. Specifically, Roberts and Sternberg (1993) evaluated the predictions of these two classes of models across a variety of experimental domains (e.g., detection, identification, and classification), and concluded that the stage model provided a better account of empirical additive effects and that the cascade model incorrectly predicted the relations between means and variances (see Roberts & Sternberg, 1993, and Yap & Balota, 2007, for more in-depth discussion).

Melvin Yap, and David Balota Department of Psychology, Washington University, Chi-Shing Tse Department of Psychology, University at Albany, State University of New York; Derek Besner, Department of Psychology, University of Waterloo, Ontario, Canada.

Melvin Yap is now at the Department of Psychology, National University of Singapore, Republic of Singapore; Chi-Shing Tse is now at the Department of Psychology, Washington University.

This research was supported by National Institute on Aging Grant AG03991 and National Science Foundation Grant BCS0001801 to D. A. Balota, and Natural Sciences and Engineering Research Council of Canada Grant A0998 to D. Besner.

We thank Ken Forster, Sachiko Kinoshita, Jim Neely, and Guy Van Orden for helpful comments on an earlier version of this article and Saul Sternberg for further clarifying the different predictions from stage and cascade models. We also extend thanks to Viviana Benitez, Garrett Broshuis, Geri Gower, and Marg Ingleton for their help with data collection.

Correspondence concerning this article should be sent to David Balota, Department of Psychology, Washington University, St. Louis, Box 1125, One Brookings Drive, St. Louis, MO, 63130. E-mail to dbalota@artsci.wustl.edu

Additive Effects and Models of Word Recognition

Importantly, the additive pattern between word frequency and stimulus quality has been invoked as support for multistage models of word recognition (e.g., Borowsky & Besner, 1993; Forster, 1976; Paap, Newsome, McDonald, & Schvaneveldt, 1982), and is troublesome for models where stimulus quality and frequency exert effects at a common locus. For example, the classic logogen model (Morton, 1969) predicts an interaction between the two variables, with larger stimulus quality effects for low-frequency words. In order to illustrate this prediction, consider a high-frequency word that is 10 cycles away from recognition threshold and a low-frequency word that is 20 cycles away from threshold, that is, a word frequency effect of 10 cycles. If one degrades the input such that the input on each cycle decreases by half in the degraded condition (compared to the clear condition), then it should take twice as many cycles for the low-frequency word to reach threshold. Specifically, the high-frequency word will be recognized in 20 (2×10) cycles, whereas its low frequency counterpart will only be recognized in 40 (2×20) cycles. Hence, the original word frequency effect of 10 cycles now yields a word frequency effect of 20 cycles in the degraded condition, producing a stimulus quality $\times$word frequency interaction.

Therefore, additivity seems incompatible with influential computational models that are built on the logogen framework, such as the interactive activation model (McClelland & Rumelhart, 1981), the computational dual route cascaded model of word recognition (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), the multiple read-out model (Grainger & Jacobs, 1996), and the CDP + model (Perry, Ziegler, & Zorzi, 2007). Indeed, simulations of the dual route cascaded model confirm our intuitions; stimulus quality and word frequency interacted in pronunciation performance, with larger stimulus quality effects for low-frequency words (Reynolds & Besner, 2004). The connectionist framework of lexical processing (Plaut, McClelland, Seidenberg, & Patterson, 1996), the major alternative to the dual route approach, also appears to yield the same interaction. Specifically, Plaut (personal communication, January 18, 2005), using the attractor network model described in Simulation 3 of Plaut et al., demonstrated that word frequency effects in pronunciation were larger for degraded items than for clear items. In summary, it appears that well-studied computational models of word recognition produce evidence for interactive effects of stimulus quality and word frequency, a finding that is difficult to reconcile with the empirical observation of additivity.

It is also unclear if additive effects of stimulus quality and word frequency can be accommodated by models that specifically target lexical decision performance. For example, the diffusion model (Ratcliff, Gomez, & McKoon, 2004) proposes that lexical decisions are driven by the accumulation of noisy information from a stimulus over time and that variations in performance can be modeled by a small number of parameters, such as drift rate (rate at which information accumulates) and boundary separation (the criterion that must be reached before a decision is made). It is not easy to envision how the diffusion model would handle strongly additive effects. For example, if we assume that stimulus degradation decreases drift rates for both high- and low-frequency words, this would predict larger stimulus quality effects for low-frequency words, given that they take longer to reach the word boundary. In fact, the geometry of the diffusion process ensures that the same change in drift rate for slow and fast response times (RTs) will lead to larger effects for the slow RTs compared to the fast RTs (Ratcliff & Rouder, 1998). Of course, the diffusion model fits a large number of parameters from data sets, including mean RTs for correct and error responses, the relative speeds of correct and error responses, the distribution of RTs, and error rates. Ultimately, whether the diffusion model can handle these effects is an empirical question that can only be answered by running the appropriate simulations on the actual model. Interestingly, when lexical decision data from a study investigating stimulus quality and frequency effects (Yap & Balota, 2007) were fitted to the diffusion model, the model parameters indicated that degradation had multiple effects. Specifically, drift rates went down, stimulus encoding took longer, and participants set more conservative response criteria (R. Ratcliff, personal communication, December 30, 2005). However, the model's fit for degraded words compared to clear words was associated with a much larger chi-square statistic, suggesting that it had difficulty capturing the empirical pattern.

Are Additive Effects Task-Specific?

The work examining the combined effects of stimulus quality and word frequency have almost always been based on the lexical decision task (LDT). Balota and Chumbley (1984) have argued that the LDT reflects both word identification processes and the word–nonword discrimination processes specific to that task. Specifically, to carry out a lexical decision, participants can use familiarity-based information to discriminate between words, which are familiar, and nonwords, which are unfamiliar. Familiarity-based information, in this context, refers to a global multidimensional quantity that reflects the orthographic and phonological similarity of a letter string to real words and is associated with labels such as *familiarity/meaningfulness* (Balota & Chumbley) and *wordness* (Ratcliff et al., 2004). With respect to extant models of visual word recognition, familiarity maps most obviously onto global lexical activity. Specifically, in both the dual route cascaded model (Coltheart et al., 2001) and the multiple read-out model (Grainger & Jacobs, 1996), word lexical decisions are produced when either local lexical activity (activation level of a single lexical representation) or global lexical activity (summed activation levels of all lexical representations) exceeds their respective thresholds. Because words will typically produce more global lexical activity than nonwords, global activity/familiarity is clearly a useful dimension for the word–nonword discrimination.

It is possible that the additive effects of stimulus quality and word frequency observed in the LDT are a by-product of lexical decision's emphasis on familiarity-based information (Balota & Chumbley, 1984; Besner, 1983). That is, because visual degradation undermines the familiarity-based information of a stimulus, participants may be compelled to perceptually normalize degraded stimuli in an additional early encoding stage, allowing familiarity-based information to be recovered and then used to discriminate familiar words from unfamiliar nonwords. If this hypothesis is correct, then additive effects should not be observed in lexical processing tasks that do not place such a high degree of emphasis on familiarity-based information. These tasks include speeded pronunciation, where participants read aloud visually presented words, and semantic classification, where participants decide whether items are exemplars of some category (e.g., Is BIRD animate?). Indeed, Yap and Balota (2007) have recently reported that for the same set of words that produces clear additive effects in lexical decision performance, speeded pronunciation and semantic classification yielded interactive effects of these two

factors, with larger stimulus quality effects for low-frequency words. Interestingly, using the same stimuli but a different degradation manipulation, O'Malley, Reynolds, and Besner (2007) obtained precisely the same pattern in speeded pronunciation. This is noteworthy because it confirms that the interaction is not an artifact of the degradation manipulation, which was contrast reduction in O'Malley et al. and rapid alternation of a letter string and a mask in Yap and Balota.

These findings suggest that the additive effects of stimulus quality and word frequency and the attendant implication of an early normalization stage are specific to lexical decision. Interestingly, in Sternberg's (1967) short-term memory scanning task, where participants decide whether a probe digit appeared in a previously presented set, additive effects of stimulus quality and set size are also realized. In this classic work, Sternberg interpreted these additive effects as being consistent with a stage model of memory search, where stimulus quality influences an early encoding stage and set size influences a subsequent serial comparison stage. This comparison is particularly germane because memory scanning can be viewed as conceptually very similar to lexical decision in that it also involves a binary decision that can be driven by familiarity-based information (Atkinson & Juola, 1974). In fact, Yap and Balota (2007) speculated that, in experimental contexts where familiarity is useful for driving binary decisions, normalization is necessary for recovering the familiarity-based information.

It is important to recognize, however, that the LDT is not a unitary task, but may recruit different types of information for performing the word–nonword discrimination, depending on the nature of the nonword foils used (e.g., Shulman, Hornak, & Sanders, 1978). Interestingly, all studies that have examined the joint effects of stimulus quality and word frequency in lexical decision (Balota & Abrams, 1995; Becker & Killion, 1977; Borowsky & Besner, 1993; O'Malley et al., 2007; Plourde & Besner, 1997; Stanners et al., 1975; Yap & Balota, 2007) have used orthographically legal nonwords, like FLIRP. Such nonwords are orthographically similar to but phonologically distinct from real words, making familiarity a viable dimension for word–nonword discrimination. The question addressed in the present study is what happens when the discrimination is made more difficult, via the inclusion of pseudohomophones (i.e., nonwords that sound like real words, e.g., BRANE) as the nonword foil items. In this situation, because of the increased word–nonword overlap, familiarity or global lexical activity becomes a less reliable metric for lexical decisions, and participants may be compelled to individuate the letter strings in a more fine-grained manner, as they appear to do in speeded pronunciation and semantic classification. For example, within the dual route model, the decision making mechanism may now weight local lexical activity more heavily than global lexical activity during lexical decision. More importantly, because the utility of familiarity is undermined, the perceptual normalization stage, which is used to recover familiarity-based information, may now be diminished in lexical decision. This implies that when pseudohomophones are used, one may no longer observe the signature additive effects of stimulus quality and word frequency.

In addition to examining the influence of stimulus quality and word frequency in the context of pseudohomophones (nonwords that sound like real words, e.g., BRANE), we also analyzed RT performance at the level of distributional characteristics. It is becoming increasingly clear that the analysis of mean RTs alone is not only inadequate but can sometimes be misleading (Andrews & Heathcote, 2001; Balota & Spieler, 1999; Heathcote, Popiel, & Mewhort, 1991; Plourde & Besner, 1997; Spieler, Balota & Faust, 1996; Yap, Balota, Cortese, & Watson, 2006). In particular, failing to take the shape of the distribution into account may obscure more subtle aspects of performance. In their seminal article, Heathcote and colleagues (1991) investigated Stroop color-naming performance. Examining mean RTs, they observed no difference between the congruent (RED displayed in red) and baseline (XXX displayed in red) conditions. However, distributional analyses revealed that congruency shifted the leading edge of the RT distribution leftwards while simultaneously increasing its skew and that these opposing effects cancelled each other out, spuriously producing a null effect (see Spieler et al., 1996, for a replication of this pattern).

Distributional shifting and skewing were quantified using ex-Gaussian analyses, which characterize RT distributions by fitting RT data to an explicit model, the ex-Gaussian distribution. The ex-Gaussian distribution is a convolution of the normal (Gaussian) and exponential distributions, and contains three parameters: μ, the mean of the normal distribution; σ^2, the variance of the normal distribution; and τ, the mean and standard deviation of the exponential distribution (mathematically, the mean and standard deviation of the exponential distribution are identical). Fitting data to the ex-Gaussian yields parameter estimates (μ, σ, τ) for the data of interest and involves iteratively searching for a set of parameters that maximizes the goodness of fit between the empirical data and the theoretical ex-Gaussian distribution (see Press, Flannery, Teukolsky, & Vetterling, 1988, pp. 274–334, for more information). One interesting property of ex-Gaussian analyses is that the mean is the algebraic sum of μ and τ, allowing differences in means to be partitioned into μ (distributional shifting) and τ (distributional skewing).

In this article, we explore RT distributional characteristics using ex-Gaussian analyses and a nonparametric technique called *vincentizing*. Vincentizing creates composite RT distributions by averaging RT distributions across participants and involves computing a predefined number of vincentiles as a function of participant and experimental condition, where a vincentile is defined as the mean of observations between neighboring percentiles. For example, to obtain 10 vincentiles, the RT data within each condition for a participant is first sorted (from fastest to slowest responses), and the first 10% of the data are then averaged, followed by the second 10%, and so on. Individual vincentiles are then averaged across participants. Importantly, unlike ex-Gaussian fitting, vincentizing represents raw data directly and makes no assumptions about the shape of the underlying RT distribution. Plots of mean vincentiles are useful for investigating how different variables influence different regions of the RT distribution and provide an important graphical complement to ex-Gaussian analyses (see Balota, Yap, Cortese, & Watson, in press, for further discussion of the relation between vincentile plots and ex-Gaussian parameters).

Most relevant for the purposes of this study, Plourde and Besner (1997) and Yap and Balota (2007) observed additivity in the means, higher order cumulants[1] (i.e., variance and the third cumulant), and the ex-Gaussian parameters in their distributional analyses of lexical

[1] Cumulants are statistics that describe the characteristics of a distribution and are closely related to its moments. The first cumulant is the mean, the second the variance, and the third estimates the skew of a distribution. The third cumulant is zero for symmetric distributions, is negative for negatively skewed distributions, and is positive for positively skewed distributions.

decision performance. In the present study, we evaluate whether the predicted additivity persists when the utility of familiarity is undermined via a pseudohomophone nonword context. Indeed, it is even possible that one may observe trade-off effects in the underlying distributional characteristics that are not apparent at the level of the mean, like those reported by Heathcote et al. (1991) in Stroop performance. It is also useful to examine higher order cumulants because additivity in these estimates provides additional support for separate and independent processes (Roberts & Sternberg, 1993). It is important to point out, however, that estimates for higher order cumulants are associated with extremely large variances and are also very sensitive to outliers (Ratcliff, 1979). For example, Ratcliff estimated that 10,000 observations were necessary before the standard deviation of the fourth cumulant could be brought down to 10% of the size of the fourth cumulant. Given the relatively modest cell sizes used in this study ($n = 50$), it is likely that estimates of the variance and the third cumulant will be highly variable and unstable. However, if interactivity is still observed in these very noisy estimates, this provides compelling evidence against independent stages.

Overview of Experiments

The joint effects of stimulus quality and word frequency were assessed in four lexical decision experiments. Legal pronounceable nonwords (e.g., FLIRP) were used in Experiment 1 whereas all the nonwords in Experiments 2, 3, and 4 consisted of pseudohomophones (e.g., BRANE). We expected to see additive effects of stimulus quality and word frequency in Experiment 1 in both means and distributional characteristics, given the prior work by Yap and Balota (2007) and Plourde and Besner (1997).[2]

In Experiment 2, the presence of pseudohomophones should increase the size of the word frequency effect, because it has been demonstrated that as word–nonword discrimination difficulty increases, word frequency effects increase in magnitude. The interaction between nonword type and word frequency can be accommodated within the random-walk framework, which, like the diffusion model described earlier, assumes that lexical decision involves the accumulation of noisy information over time (Stone & Van Orden, 1993; Yap et al., 2006). Like the diffusion model, the random-walk model produces larger effects in RT latencies when signal strength (the analogue of drift rate) is low. Broadly speaking, when pseudohomophones are used, discrimination difficulty is high and evidence accumulates less rapidly. This translates to a lower signal strength and larger effects. The critical question is whether we now obtain evidence for additivity in means and RT distributional characteristics when word–nonword discrimination is more difficult. Because of the intriguing distributional results obtained in Experiment 2, we attempted to replicate this pattern in two additional pools of participants from two different universities, who have different profiles of performance in the LDT. Because the experiments are very similar, we provide a single overall General Method section, and report the results separately for each experiment.

General Method

Sample and Procedures

Participants. A total of 144 participants participated in the four experiments for course credit or $10 (see Table 1 for a summary of participant characteristics). All participants had normal or corrected-to-normal vision and were recruited from the Washington University (Experiments 1 and 2); the University at Albany, State University of New York (Experiment 3); and the University of Waterloo (Experiment 4) participant pools.

Table 1
Mean Age, Years of Education, and Vocabulary Age (in Years) of Participants (Standard Deviations in Parentheses)

Experiment	*N*	Age	Years of Education	Vocabulary Age[a]
1	28	20.54 (5.73)	13.50 (1.43)	18.70 (0.86)
2	56	20.87 (5.44)	13.78 (1.37)	18.72 (0.97)
3	28	19.37 (2.73)	12.22 (1.22)	16.86 (1.81)
4	32	20.94 (1.90)	Not available	17.72 (1.18)

Note. [a]Shipley (1940) vocabulary scores.

Apparatus. An IBM-compatible computer running E-prime software (Schneider, Eschman, & Zuccolotto, 2001) was used to control stimulus presentation and to collect data. The stimuli were displayed on a 17-in. Super VGA monitor, and participants' responses were made on a computer keyboard.[3]

Stimuli. The stimuli for the LDT consisted of 200 words, 200 length-matched pronounceable nonwords (Experiment 1), and 200 length-matched pseudohomophones (Experiments 2, 3, and 4; see Appendix for a full list of stimuli). One hundred words were designated high frequency (mean counts per million = 1,227) and 100 words low frequency (mean counts per million = 44; Lund & Burgess, 1996). For high-frequency words, the mean length was 4.73 letters ($SD = .96$), and the mean orthographic neighborhood size (N) (Coltheart, Davelaar, Jonasson, & Besner, 1977) was 4.77. For low-frequency words, the mean length was 4.78 letters ($SD = .85$), and the mean N was 4.82. There was no significant difference between high- and low-frequency words with respect to length ($t < 1$) and orthographic neighborhood size ($t < 1$). For the legal nonwords and pseudohomophones, the mean Ns were 3.38 and 4.66, respectively, $p = .002$.[4] For the pseudohomophones, the mean baseword fre-

[2] The purpose of Experiment 1 was to provide a necessary baseline, specifically a within-subjects experiment utilizing the degradation manipulation (rapid interleaving of mask and target) described in Yap and Balota (2007). Yap and Balota used a between-participants manipulation; and, although Plourde and Besner (1997) had a within-subjects design, their degradation manipulation was based on contrast reduction.

[3] One reviewer was concerned that collecting response latencies from the keyboard may not be as accurate as using the PST serial response box. According to the timing simulations described in Schneider et al. (2001), E-prime programs running on appropriately tuned desktop Pentium class machines running at 120MHz or faster can collect millisecond precise real-time input from either a keyboard or the PST response box.

[4] The fact that the pseudohomophones and legal nonwords were not matched on orthographic N suggests that any observed nonword type effects cannot be unambiguously attributed to pseudohomophony; it is quite possible that orthographic N is also making a contribution. However, in Experiments 2–4, we are not claiming that pseudohomophony per se is responsible for the observed effects. Rather, we are primarily interested in manipulating the familiarity of the nonword context and examining how foils that are more similar to targets (i.e., pseudohomophones) influence the response to words. Clearly, familiarity in this instance is a multidimensional quantity that could encompass pseudohomophony and/or orthographic neighborhood size.

quency (i.e., the frequency of the word BRAIN for the nonword BRANE) was 647 counts per million.

Procedure. Participants were tested individually in sound-attenuated cubicles under standard lighting conditions, sitting about 60 cm from the computer screen. After completing a computer-administered 40-item Shipley (1940) vocabulary task, participants were told that letter strings would be presented at the center of the screen and that they were to indicate as quickly and accurately as possible via a button press on the keyboard whether the letter string spelled a word they knew or not. Twenty practice trials were then presented, followed by five experimental blocks of 80 trials, with mandatory breaks between blocks. The order in which stimuli were presented was randomized anew for each participant, with clear and degraded trials randomly intermixed within each block. All stimuli were uppercase and were presented in the 14-point Courier font. For the degradation condition, letter strings were rapidly alternated with a randomly generated mask of the same length. For example, the mask &?# was presented for 14 ms, followed by DOG for 28 ms; and the two were repeatedly alternated until the participant responded. The mask was generated from random permutations of the symbols @, #, $, %, &, ?, and *, with the proviso that the mask be the same length as the string and that symbols not be repeated within a mask. Each trial consisted of the following order of events: (a) a fixation point (+) at the center of the monitor for 400 ms, (b) a blank screen for 200 ms, and (c) a stimulus centered at the fixation point's location. The stimulus remained on the screen until a keyboard response was made. Participants pressed the apostrophe key for words and the *A* key for nonwords. Each correct response was followed by a 1,600 ms delay. If a response was incorrect, a 170 ms tone was presented simultaneously with the onset of a 450 ms presentation of the word "Incorrect" (displayed slightly below the fixation point). In order to keep the delay between the response to a stimulus and the presentation of the next stimulus constant across correct and incorrect trials, each incorrect response was followed by a 1,150-ms delay.

Design. A 2 × 2 factorial design was used: both stimulus quality and word frequency were manipulated within participants. The stimulus quality of each item was also counterbalanced across participants, so that there were 50 observations for each of the four experimental conditions.

Results

For all experiments, errors and response latencies faster than 200 ms or slower than 3,000 ms were first excluded, and the overall mean and standard deviation of each participant's word and nonword latencies were then computed. Of the remaining latencies, any latencies 2.5 standard deviations above or below each participant's respective mean (across all conditions) were removed. Analyses of variance (ANOVAs) were then carried out on the mean response latencies, accuracy, cumulants, and ex-Gaussian parameters of the RT data. For RT means and accuracy, ANOVAs by participants and items were conducted. For the cumulants and ex-Gaussian parameters, only ANOVAs by participants were conducted.

Experiment 1

Results

In Experiment 1, the overall error rate was 6.6%, and data trimming eliminated an additional 2.4% of the trials. The mean response latencies, accuracy, variance, third cumulant, and ex-Gaussian parameters for Experiment 1 are presented in Table 2.

Mean response latencies and accuracy. The ANOVA on mean response latencies yielded significant main effects of stimulus quality, $F_p(1, 27) = 81.30$, $p < .001$, $MSE = 2{,}795.42$, $\eta^2 = .75$; $F_i(1, 198) = 285.10$, $p < .001$, $MSE = 3{,}167.84$, $\eta^2 = .59$, and word frequency, $F_p(1, 27) = 118.61$, $p < .001$, $MSE = 627.34$, $\eta^2 = .82$; $F_i(1, 198) = 52.36$, $p < .001$, $MSE = 6{,}545.71$, $\eta^2 = .21$. The stimulus quality × word frequency interaction was not significant, either by participants ($F < 1$) or by items ($p = .24$). Turning to accuracy, the main effects of stimulus quality, $F_p(1, 27) = 21.67$, $p < .001$, $MSE = .0012$, $\eta^2 = .45$; $F_i(1, 198) = 22.13$, $p < .001$, $MSE = .0040$, $\eta^2 = .10$, and word frequency, $F_p(1, 27) = 45.48$, $p < .001$, $MSE = .0014$, $\eta^2 = .63$; $F_i(1, 198) = 22.31$, $p < .001$, $MSE = .010$, $\eta^2 = .10$, were again significant. The interaction between stimulus quality and word frequency was not significant (F_p and $F_i < 1$).

Variance and third cumulant. For variance, only the main effect of stimulus quality was significant, $F(1, 27) = 11.02$, $p = .003$,

Table 2
Means of Participants' Lexical Decision Response Time Means, Accuracy, Variance, Third Cumulant, and Ex-Gaussian Parameter Estimates as a Function of Stimulus Quality and Word Frequency in Experiment 1, Legal Nonword Background

Stimulus quality/word frequency	*M*	% Errors	Mu	Sigma	Tau	Var	3rd cum
Clear							
High-frequency words	557	2	455	43	102	1.67E + 04	5.77E + 06
Low-frequency Words	605	6.8	486	53	121	1.93E + 04	4.68E + 06
Frequency effect	48	4.8	31	10	19	2.61E + 03	−1.09E + 06
Degraded							
High-frequency words	644	5	490	41	156	4.07E + 04	2.95E + 07
Low-frequency Words	699	9.8	522	47	180	4.37E + 04	1.97E + 07
Frequency effect	55	4.8	32	6	24	2.96E + 03	−9.78E + 06
Interaction	7	0	1	−4	5	3.43E + 02	−8.69E + 06
Nonwords							
Clear legal nonwords	683	7	567	64	117		
Degraded legal nonwords	797	7.6	583	68	216		

MSE = 1.50E + 09, η^2 = .29; the variance was larger for degraded words. The main effect of frequency and the interaction were not significant, *F*s < 1. For the third cumulant, only the main effect of stimulus quality was significant, $F(1, 27) = 5.20$, $p = .031$, *MSE* = 2.03E + 15, $\eta^2 = .16$; the third cumulant was larger for degraded words. The main effect of frequency and the interaction did not approach significance, *F*s < 1. As noted earlier, because of the high degree of variability in higher order cumulants, the lack of a reliable interaction cannot be taken as definitive support for additive stages, but it is consistent with this perspective.

Ex-Gaussian analyses. Using the quantile maximum likelihood estimation (QMLE) procedure in the QMPE v2.18 program (Cousineau, Brown, & Heathcote, 2004; Heathcote, Brown, & Mewhort, 2002), ex-Gaussian parameters (μ, σ, τ) were obtained for each participant across the different experimental conditions. QMLE provides unbiased parameter estimates and has been demonstrated to be more effective than continuous maximum likelihood estimation for small samples (Heathcote & Brown, 2004; Speckman & Rouder, 2004). All fits successfully converged within 250 iterations.

For μ, the main effects of stimulus quality, $F(1, 27) = 45.61$, $p < .001$, *MSE* = 753.38, $\eta^2 = .63$, and word frequency, $F(1, 27) = 80.83$, $p < .001$, *MSE* = 336.74, $\eta^2 = .75$, were significant. The stimulus quality × word frequency interaction was not significant ($F < 1$). Turning to σ, only the main effect of word frequency approached significance, $F(1, 27) = 3.49$, $p = .073$, *MSE* = 480.11, $\eta^2 = .11$. Turning to τ, the main effects of stimulus quality, $F(1, 27) = 23.37$, $p < .001$, *MSE* = 3,808.02, $\eta^2 = .46$, and word frequency, $F(1, 27) = 12.08$, $p = .002$, *MSE* = 1,030.34, $\eta^2 = .31$, were significant. The stimulus quality × word frequency interaction was not significant ($F < 1$). The ex-Gaussian analysis is very clear; all parameters (except σ) produced main effects, but none of the parameters produced interactions.

Vincentile analyses. The mean vincentiles[5] for the different experimental conditions are plotted in the top two panels of Figure 1, whereas the bottom panel of Figure 1 presents the mean frequency effect as a function of stimulus quality and vincentiles. For the top two panels, the empirical mean vincentiles are represented by data points and standard error bars, and the estimated vincentiles for the respective best-fitting ex-Gaussian distribution are represented by lines. As the error bars indicate, each vincentile represents a different range of RTs for each participant. Presenting the data in this manner allows one to visually assess the goodness of fit between empirical and estimated vincentiles. Clearly, the data are fitted well by the ex-Gaussian distribution, and the divergence between mean vincentiles and theoretical ex-Gaussian vincentiles is typically smaller than one standard error in most cases. Note that the bottom panel, which presents difference scores, depicts only empirical vincentiles.

As one can see, although the frequency effect increases across vincentiles for both the clear and degraded conditions, word frequency effects are not different for clear and degraded items at any of the vincentiles. In summary, when legal nonwords are used in lexical decision, stimulus quality and word frequency produce robust additive effects in RT means and distributional characteristics, along with accuracy, a pattern that is consistent with the available literature. We now turn to experiments that explore this interaction in the context of pseudohomophones.

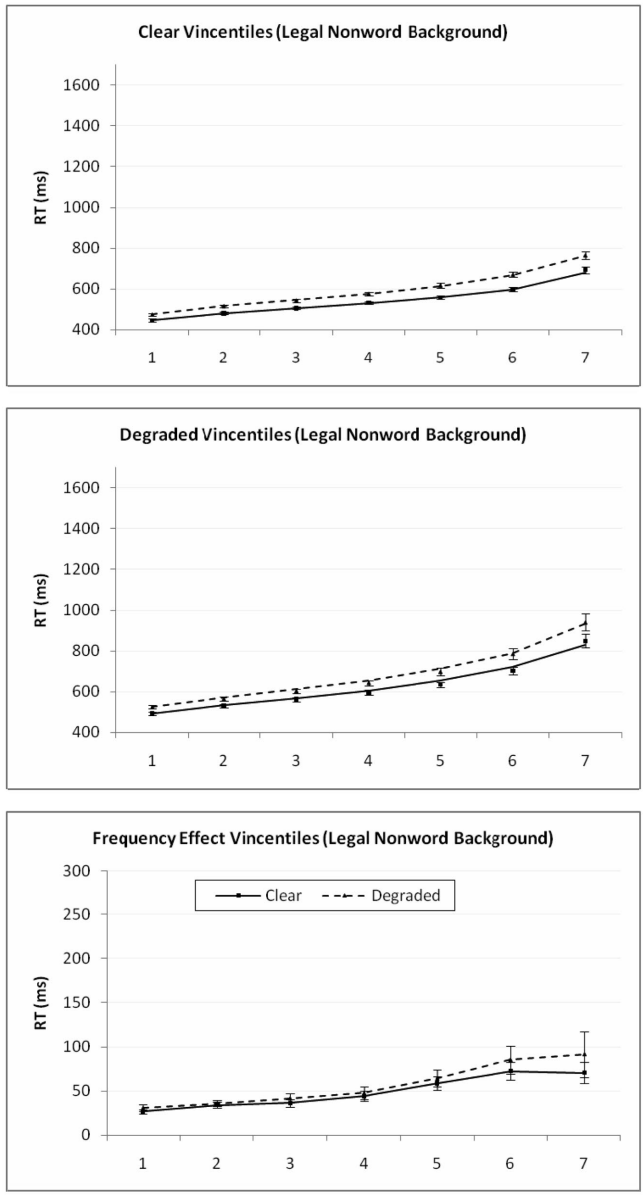

Figure 1. Lexical decision performance from Experiment 1 (legal nonword context) as a function of word frequency and vincentiles in the clear (top panel) and degraded (middle panel) conditions, along with frequency effects as a function of vincentiles (bottom panel). In the top and middle panels, participants' mean vincentiles (■ = high frequency, ▲= low frequency) are represented by data points and standard error bars. Best-fitting ex-Gaussian vincentiles are represented by lines (solid line = high frequency, dashed line = low frequency).

[5] In preliminary analyses, we plotted the vincentiles using 7, 10, and 15 vincentiles. The same trends emerged regardless of the number of vincentiles. We decided to use 7 vincentiles to maximize the reliability of each vincentile. For each participant, this allows each vincentile to be based on approximately 6 RT trials (after eliminating error and outlier trials), given that there are 50 trials per condition.

Experiment 2

Results

In Experiment 2, the overall error rate was 7.7%, and data trimming eliminated an additional 2.6% of the trials. The mean response latencies, accuracy, ex-Gaussian parameters, variances, and third cumulants for Experiment 2 are presented in Table 3.

Mean response latencies and accuracy. The ANOVA on mean response latencies yielded significant main effects of stimulus quality, $F_p(1, 55) = 158.22$, $p < .001$, $MSE = 5{,}850.34$, $\eta^2 = .74$; $F_i(1, 198) = 442.81$, $p < .001$, $MSE = 3{,}941.15$, $\eta^2 = .69$, and word frequency, $F_p(1, 55) = 79.54$, $p < .001$, $MSE = 3{,}225.16$, $\eta^2 = .59$; $F_i(1, 198) = 76.77$, $p < .001$, $MSE = 10{,}184.26$, $\eta^2 = .28$. The stimulus quality × word frequency interaction was not significant, either by participants or by items (F_p and $F_i < 1$). Turning to accuracy, the main effects of stimulus quality, $F_p(1, 55) = 24.68$, $p < .001$, $MSE = .0014$, $\eta^2 = .31$; $F_i(1, 198) = 36.33$, $p < .001$, $MSE = .0017$, $\eta^2 = .16$, and word frequency, $F_p(1, 55) = 86.12$, $p < .001$, $MSE = .0021$, $\eta^2 = .61$; $F_i(1, 198) = 28.48$, $p < .001$, $MSE = .014$, $\eta^2 = .13$, were significant. The interaction between stimulus quality and word frequency was not significant, either by participants ($p = .30$) or by items ($p = .23$).

Variance and third cumulant. For variance, the main effects of stimulus quality, $F(1, 55) = 38.17$, $p < .001$, $MSE = 2.20\text{E} + 09$, $\eta^2 = .41$, and word frequency, $F(1, 55) = 10.19$, $p = .002$, $MSE = 6.54\text{E} + 08$, $\eta^2 = .16$, were significant; the variance was larger for degraded words and low-frequency words. The interaction was not significant, $F < 1$. For the third cumulant, only the main effect of stimulus quality was significant, $F(1, 55) = 13.94$, $p < .001$, $MSE = 2.79\text{E} + 15$, $\eta^2 = .20$; the third cumulant was larger for degraded words. The main effect of frequency and the interaction were not significant, Fs < 1.

Ex-Gaussian analyses. For μ, the main effects of stimulus quality, $F(1, 55) = 84.35$, $p < .001$, $MSE = 1{,}738.92$, $\eta^2 = .61$, and word frequency, $F(1, 55) = 65.29$, $p < .001$, $MSE = 1{,}182.50$, $\eta^2 = .54$, were significant. Importantly, the stimulus quality × word frequency interaction was now significant, $F(1, 55) = 6.05$, $p = .017$, $MSE = 1{,}106.37$, $\eta^2 = .10$, with larger frequency effects for degraded words. Turning to σ, the main effect of word frequency was significant, $F(1, 55) = 11.62$, $p = .001$, $MSE = 547.65$, $\eta^2 = .17$; the interaction between stimulus quality and word frequency was marginal, $F(1, 55) = 3.88$, $p = .054$, $MSE = 849.55$, $\eta^2 = .07$, with larger frequency effects for degraded words. Turning to τ, the main effects of stimulus quality, $F(1, 55) = 65.70$, $p < .001$, $MSE = 5{,}034.61$, $\eta^2 = .54$, and word frequency, $F(1, 55) = 21.83$, $p < .001$, $MSE = 2{,}563.92$, $\eta^2 = .28$, were significant, but the stimulus quality × word frequency interaction was not significant ($p = .22$). Although the stimulus quality × word frequency interaction was not significant in τ, it is noteworthy that the effect of frequency was actually larger in the clear condition than in the degraded condition, which is precisely opposite to the pattern obtained in μ and σ.

Vincentile analyses. The magnitudes of the frequency effect across the different vincentiles are plotted in Figure 2 (bottom panel). In the presence of pseudohomophones, degraded words show a larger frequency effect than clear words, particularly in the early and modal vincentiles. This trend reverses in the final vincentile, where clear words show a larger frequency effect than degraded words. This is consistent with the nonsignificant tendency in τ, noted above. Hence, the opposing effects in the early and late vincentiles create additive effects at the level of the mean.

In summary, at the level of the mean, large additive effects of stimulus quality and word frequency were obtained in both Experiments 1 and 2. Importantly, at the level of distributional characteristics, the two experiments yielded different patterns. In Experiment 1, which used legal nonword foils, additivity was demonstrated in both means and distributional characteristics. In Experiment 2, which used pseudohomophone foils, stimulus quality and word frequency interacted in μ, with larger frequency effects for degraded words. The vincentile analyses also confirmed that frequency effects were larger for degraded words in the early vincentiles. Collectively, these results suggest that when distributional characteristics are considered, one does not obtain unambiguous additivity between stimulus quality and word frequency in the presence of pseudohomophones. Instead, for the fastest and modal RTs, a larger frequency effect is obtained for degraded words; this trend is offset by an opposing effect in the slowest RTs. As noted above, such tradeoffs in distributional characteristics have been observed before (see Heathcote et al., 1991, and Spieler

Table 3
Means of Participants' Lexical Decision Response Time Means, Accuracy, Variance, Third Cumulant, and Ex-Gaussian Parameter Estimates as a Function of Stimulus Quality and Word Frequency in Experiment 2, Pseudohomophone Background

Stimulus quality/word frequency	*M*	% Errors	Mu	Sigma	Tau	Var	3rd cum
Clear							
High-frequency words	612	2.3	476	44	138	3.01E + 04	1.94E + 07
Low-frequency Words	678	7.5	502	47	177	4.40E + 04	2.45E + 07
Frequency effect	66	5.2	26	3	39	1.39E + 04	5.11E + 06
Degraded							
High-frequency words	738	4.3	516	40	222	7.19E + 04	5.11E + 07
Low-frequency Words	809	10.4	564	58	246	7.97E + 04	4.55E + 07
Frequency effect	71	6.1	48	18	24	7.87E + 03	−5.59E + 06
Interaction	5	0.9	22	15	−15	−6.07E + 03	−1.07E + 07
Nonwords							
Clear pseudohomophones	749	7.9	564	52	187		
Degraded pseudohomophones	912	10.5	625	74	288		

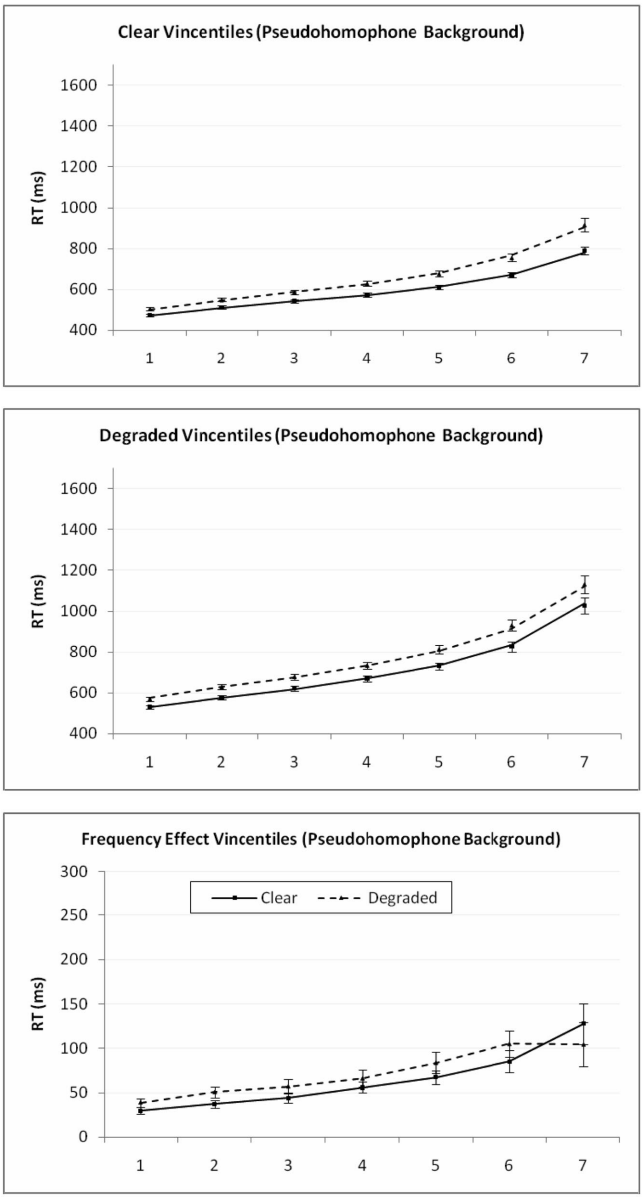

Figure 2. Lexical decision performance from Experiment 2 (pseudohomophone context) as a function of word frequency and vincentiles in the clear (top panel) and degraded (middle panel) conditions, along with frequency effects as a function of vincentiles (bottom panel). In the top and middle panels, participants' mean vincentiles (■ = high frequency, ▲ = low frequency) are represented by data points and standard error bars. Best-fitting ex-Gaussian vincentiles are represented by lines (solid line = high frequency, dashed line = low frequency).

et al., 1996). However, because of the theoretical importance of such a pattern, we decided to establish its stability by attempting to replicate the results from Experiment 2 in two distinct populations (Experiment 3, University at Albany, and Experiment 4, University of Waterloo), which produced varying levels of performance in the LDT.

Experiment 3

Results

The overall error rate was 11.3%, and data trimming eliminated an additional 4.3% of the trials. The mean response latencies, accuracy, ex-Gaussian parameters, variances, and third cumulants for Experiment 3 are presented in Table 4.

Mean response latencies and accuracy. The ANOVA on mean response latencies yielded significant main effects of stimulus quality, $F_p(1, 27) = 92.89$, $p < .001$, $MSE = 12{,}773.26$, $\eta^2 = .78$; $F_i(1, 198) = 299.50$, $p < .001$, $MSE = 13{,}001.93$, $\eta^2 = .60$, and word frequency, $F_p(1, 27) = 70.38$, $p < .001$, $MSE = 4{,}498.66$, $\eta^2 = .72$; $F_i(1, 198) = 56.21$, $p < .001$, $MSE = 23{,}168.92$, $\eta^2 = .22$. The stimulus quality × word frequency interaction was not significant, either by participants or by items (F_p and $F_i < 1$). Turning to accuracy, the main effect of stimulus quality approached significance, $F_p(1, 27) = 3.84$, $p = .06$, $MSE = .0038$, $\eta^2 = .13$; $F_i(1, 198) = 9.59$, $p = .002$, $MSE = .0055$, $\eta^2 = .05$, whereas the main effect of word frequency was significant, $F_p(1, 27) = 50.70$, $p < .001$, $MSE = .0038$, $\eta^2 = .65$; $F_i(1, 198) = 35.85$, $p < .001$, $MSE = .019$, $\eta^2 = .15$. The interaction between stimulus quality and word frequency was not significant, either by participants or by items (F_p and $F_i < 1$).

Variance and third cumulant. For variance, the main effects of stimulus quality, $F(1, 27) = 16.35$, $p < .001$, $MSE = 5.17\text{E}+09$, $\eta^2 = .38$, and word frequency, $F(1, 27) = 5.21$, $p = .031$, $MSE = 2.39\text{E}+09$, $\eta^2 = .16$, were significant; the variance was larger for degraded words and low-frequency words. The interaction approached significance, $p = .094$, with smaller frequency effects for degraded words. For the third cumulant, both the main effects of stimulus quality ($p = .326$) and word frequency ($F < 1$) were not significant. The interaction showed a trend towards significance, $F(1, 27) = 2.41$, $p = .132$, with smaller frequency effects for degraded words. As we shall see below, both these trends were consistent with additional aspects of the distributional analyses.

Ex-Gaussian analyses. For μ, the main effects of stimulus quality, $F(1, 27) = 16.35$, $p < .001$, $MSE = 5{,}860.69$, $\eta^2 = .65$, and word frequency, $F(1, 27) = 15.01$, $p = .001$, $MSE = 9{,}673.49$, $\eta^2 = .36$, were significant. The stimulus quality × word frequency interaction was significant, $F(1, 27) = 6.55$, $p = .016$, $MSE = 3{,}232.73$, $\eta^2 = .20$, with larger frequency effects for degraded words. Turning to σ, the main effects of stimulus quality, $F(1, 27) = 7.42$, $p = .011$, $MSE = 2{,}006.20$, $\eta^2 = .22$, and word frequency, $F(1, 27) = 10.84$, $p = .003$, $MSE = 5{,}127.78$, $\eta^2 = .29$, were significant. The interaction was significant, $F(1, 27) = 4.16$, $p = .05$, $MSE = 2{,}154.51$, $\eta^2 = .13$, with larger frequency effects in the degraded condition. Turning to τ, the main effect of stimulus quality was significant, $F(1, 27) = 31.20$, $p < .001$, $MSE = 9{,}792.24$, $\eta^2 = .54$, whereas the effect of word frequency approached significance, $F(1, 27) = 3.71$, $p = .065$, $MSE = 10{,}525.34$, $\eta^2 = .12$. The stimulus quality × word frequency interaction also approached significance, $F(1, 27) = 3.93$, $p = .058$, $MSE = 7{,}763.73$, $\eta^2 = .13$, with significant simple effects of frequency in the clear ($p < .001$), but not the degraded ($t < 1$), condition. As in Experiment 2, the stimulus quality × word frequency interaction was significant for μ; but, in addition, the interaction was also borderline significant for σ and τ.

Table 4
Means of Participants' Lexical Decision Response Time Means, Accuracy, Variance, Third Cumulant, and Ex-Gaussian Parameter Estimates as a Function of Stimulus Quality and Word Frequency in Experiment 3, Pseudohomophone Background

Stimulus quality/word frequency	*M*	% Errors	Mu	Sigma	Tau	Var	3rd cum
Clear							
High-frequency words	732	3.1	507	52	227	8.56E + 04	6.61E + 07
Low-frequency Words	844	11.1	552	79	297	1.22E + 05	7.92E + 07
Frequency effect	112	8	45	27	70	3.64E + 04	1.31E + 07
Degraded							
High-frequency words	944	5.1	582	57	364	1.56E + 05	9.64E + 07
Low-frequency Words	1044	13.6	682	120	369	1.62E + 05	7.76E + 07
Frequency effect	100	8.5	100	63	5	5.79E + 03	−1.88E + 07
Interaction	−12	0.5	55	36	−65	−3.06E + 04	−3.19E + 07
Nonwords							
Clear pseudohomophones	970	13.1	676	108	294		
Degraded pseudohomophones	1232	15.4	862	211	372		

Vincentile analyses. The magnitudes of the frequency effect across the different vincentiles are plotted in Figure 3 (bottom panel). As in Experiment 2, there is a larger frequency effect for degraded words in the fast vincentiles, and this trend reverses sharply in the final vincentiles. This pattern is also very consistent with the tradeoffs in the ex-Gaussian parameters, and the trends in the relatively unstable higher order cumulants. These opposing effects again create the overall additive effects at the level of the mean. Because of the theoretical importance of tradeoffs in RT distributional characteristics, we attempted to replicate this pattern at a third university.

Experiment 4

Results

The overall error rate was 11.4%, and data trimming eliminated an additional 3.5% of the trials. The mean response latencies, accuracy, ex-Gaussian parameters, variances, and third cumulants for Experiment 4 are presented in Table 5.

Mean response latencies and accuracy. The ANOVA on mean response latencies yielded significant main effects of stimulus quality, $F_p(1, 31) = 57.30$, $p < .001$, $MSE = 23{,}256.84$, $\eta^2 = .65$; $F_i(1, 198) = 505.94$, $p < .001$, $MSE = 8{,}145.96$, $\eta^2 = .72$, and word frequency, $F_p(1, 31) = 64.09$, $p < .001$, $MSE = 5{,}078.49$, $\eta^2 = .67$; $F_i(1, 198) = 67.14$, $p < .001$, $MSE = 17{,}947.67$, $\eta^2 = .25$. The stimulus quality × word frequency interaction was not significant, either by participants or by items (F_p and $F_i < 1$). Turning to accuracy, the main effects of stimulus quality, $F_p(1, 31) = 72.39$, $p < .001$, $MSE = .0020$, $\eta^2 = .70$; $F_i(1, 198) = 75.59$, $p < .001$, $MSE = .0059$, $\eta^2 = .28$, and word frequency, $F_p(1, 31) = 67.03$, $p < .001$, $MSE = .0036$, $\eta^2 = .68$; $F_i(1, 198) = 40.00$, $p < .001$, $MSE = .019$, $\eta^2 = .17$, were significant. The interaction between stimulus quality and word frequency was borderline significant by participants ($p = .053$) but not by items ($p = .124$).

Variance and third cumulant. For variance, the main effects of stimulus quality, $F(1, 31) = 23.26$, $p < .001$, $MSE = 5.01E + 09$, $\eta^2 = .43$, and word frequency, $F(1, 31) = 9.64$, $p = .004$, $MSE = 1.33E + 09$, $\eta^2 = .24$, were significant; the variance was larger for degraded words and low-frequency words. The interaction was not significant, $F < 1$. For the third cumulant, only the main effect of stimulus quality was significant, $F(1, 31) = 11.17$, $p = .002$, $MSE = 1.99E + 15$, $\eta^2 = .27$; the third cumulant was larger for degraded words. The main effect of frequency ($p = .307$) and the interaction ($F < 1$) were not significant.

Ex-Gaussian analyses. For μ, the main effects of stimulus quality, $F(1, 31) = 48.77$, $p < .001$, $MSE = 6{,}093.32$, $\eta^2 = .61$, and word frequency, $F(1, 31) = 76.06$, $p < .001$, $MSE = 1{,}494.41$, $\eta^2 = .71$, were significant. The stimulus quality × word frequency interaction was again significant, $F(1, 31) = 5.17$, $p = .030$, $MSE = 2{,}641.78$, $\eta^2 = .14$, with larger frequency effects for the degraded words. Turning to σ, the main effects of stimulus quality, $F(1, 31) = 10.84$, $p = .002$, $MSE = 3{,}512.34$, $\eta^2 = .26$, and word frequency, $F(1, 31) = 9.18$, $p = .005$, $MSE = 1{,}407.75$, $\eta^2 = .23$, were significant. The interaction approached significance, $F(1, 31) = 2.96$, $p = .096$, $MSE = 2{,}402.63$, $\eta^2 = .09$, with larger frequency effects in the degraded condition. Turning to τ, the main effects of stimulus quality, $F(1, 31) = 36.59$, $p < .001$, $MSE = 10{,}696.82$, $\eta^2 = .54$, and word frequency, $F(1, 31) = 14.15$, $p = .001$, $MSE = 3{,}827.70$, $\eta^2 = .31$, were significant. The stimulus quality × word frequency interaction showed a trend towards significance, $F(1, 31) = 2.52$, $p = .123$, with significant simple effects of frequency in the clear ($p < .001$), but not the degraded ($p = .098$), condition. In summary, the stimulus quality × word frequency interaction was significant for μ, and approached significance for σ and τ.

Vincentile analyses. The magnitudes of the frequency effect across the different vincentiles are plotted in Figure 4 (bottom panel). As in Experiments 2 and 3, there is a larger frequency effect for degraded words in the fast vincentiles, and this trend reverses in the final vincentiles, creating additivity at the level of the mean. In general, Experiment 4 replicates the empirical pattern observed in Experiments 2 and 3, supporting the robustness of the tradeoffs.

Composite Analyses

Because participants in Experiments 2, 3, and 4 were presented with the same set of items, we also conducted composite analyses that combined the data of the 116 participants across the three experiments, with university as a between-participant variable and stimulus quality and frequency as within-participant variables.

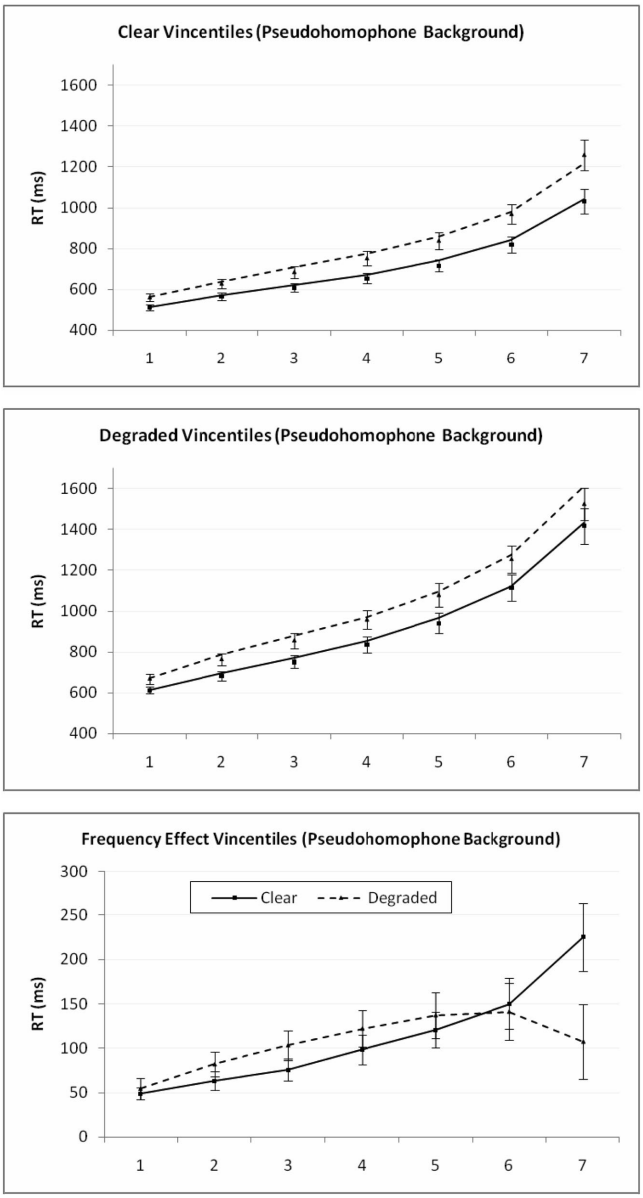

Figure 3. Lexical decision performance from Experiment 3 (pseudohomophone context) as a function of word frequency and vincentiles in the clear (top panel) and degraded (middle panel) conditions, along with frequency effects as a function of vincentiles (bottom panel). In the top and middle panels, participants' mean vincentiles (■ = high frequency, ▲ = low frequency) are represented by data points and standard error bars. Best-fitting ex-Gaussian vincentiles are represented by lines (solid line = high frequency, dashed line = low frequency).

This was done in order to maximize power and to establish the reliability of the observed distributional effects across three sets of participants who are associated with different lexical decision profiles. Table 6 presents the mean response latencies, accuracy, and ex-Gaussian parameters for these analyses, collapsing across Experiments 2 to 4.

Mean response latencies and accuracy. The ANOVA on mean response latencies yielded significant main effects of stimulus quality, $F(1, 113) = 278.39$, $p < .001$, $MSE = 12{,}279.72$, $\eta^2 = .71$, and word frequency, $F(1, 113) = 220.60$, $p < .001$, $MSE = 4{,}037.88$, $\eta^2 = .66$. The stimulus quality × word frequency interaction was not significant, $F < 1$. University interacted significantly with stimulus quality ($p = .002$) and frequency ($p = .012$). Compared to the participants in Washington University (Experiment 2), participants in University at Albany (Experiment 3) and the University of Waterloo (Experiment 4) produced larger main effects of stimulus quality and word frequency in means. Importantly, the three-way interaction between university, stimulus quality, and word frequency was not significant ($F < 1$), indicating that additive effects of stimulus quality and word frequency were replicated across sites.

Turning to accuracy, the main effects of stimulus quality, $F_p(1, 113) = 72.45$, $p < .001$, $MSE = .0021$, $\eta^2 = .39$, and word frequency, $F_p(1, 113) = 207.78$, $p < .001$, $MSE = .0029$, $\eta^2 = .65$, were significant. The interaction between stimulus quality and word frequency was borderline significant ($p = .053$), with larger frequency effects for degraded words. University interacted significantly with stimulus quality ($p < .001$) and frequency ($p = .019$). For stimulus quality, participants in Experiment 4 produced a larger main effect than participants in Experiment 2 and Experiment 3; for frequency, participants in Experiment 3 and Experiment 4 produced a larger main effect than participants in Experiment 2. The three-way interaction between university, stimulus quality, and word frequency was not significant ($F < 1$).

Variance and third cumulant. For variance, the main effects of stimulus quality, $F(1, 113) = 75.97$, $p < .001$, $MSE = 3.68\text{E} + 09$, $\eta^2 = .40$, and word frequency, $F(1, 113) = 25.44$, $p < .001$, $MSE = 1.26\text{E} + 09$, $\eta^2 = .18$, were significant; the variance was larger for degraded words and low-frequency words. The interaction approached significance ($p = .106$), with smaller frequency effects for degraded words. For the third cumulant, the main effect of stimulus quality was significant, $F(1, 113) = 16.16$, $p < .001$, $MSE = 3.28\text{E} + 15$, $\eta^2 = .13$; the third cumulant was larger for degraded words. Although not significant, $F(1, 113) = 1.98$, $p = .163$, the interaction in the third cumulant also indicated smaller frequency effects for degraded than clear words. Again, as expected, the high variability in these measures does not allow strong inferences to be made. University did not interact significantly with any other variable in both cumulants.

Ex-Gaussian analyses. For μ, the main effects of stimulus quality, $F(1, 113) = 188.06$, $p < .001$, $MSE = 3{,}918.34$, $\eta^2 = .63$, and word frequency, $F(1, 113) = 101.82$, $p < .001$, $MSE = 3{,}296.89$, $\eta^2 = .47$, were significant. The stimulus quality × word frequency interaction was significant, $F(1, 113) = 20.22$, $p < .001$, $MSE = 2{,}035.66$, $\eta^2 = .15$, with larger frequency effects for the degraded words. University interacted significantly with stimulus quality ($p < .001$) and frequency ($p = .024$). Compared to the participants in Experiment 2, participants in Experiment 3 and Experiment 4 produced larger main effects of stimulus quality and word frequency in μ. The three-way interaction between university, stimulus quality, and word frequency was not significant ($p = .26$), indicating that the same interaction between stimulus quality and word frequency was present across sites.

Turning to σ, the main effects of stimulus quality, $F(1, 113) = 22.69$, $p < .001$, $MSE = 1{,}942.11$, $\eta^2 = .17$, and word frequency, $F(1, 113) = 35.61$, $p < .001$, $MSE = 1{,}877.97$, $\eta^2 = .24$, were significant. The stimulus quality × word frequency interaction was significant, $F(1, 113) = 12.16$, $p = .001$, $MSE = 1{,}587.42$, $\eta^2 =$

Table 5
Means of Participants' Lexical Decision Response Time Means, Accuracy, Variance, Third Cumulant, and Ex-Gaussian Parameter Estimates as a Function of Stimulus Quality and Word Frequency in Experiment 4, Pseudohomophone Background

Stimulus quality/word frequency	*M*	% Errors	Mu	Sigma	Tau	Var	3rd cum
Clear							
High-frequency words	658	2.5	503	42	156	4.08E + 04	2.49E + 07
Low-frequency Words	753	10	542	47	212	5.84E + 04	2.67E + 07
Frequency effect	95	7.5	39	5	56	1.77E + 04	1.75E + 06
Degraded							
High-frequency words	856	8	579	62	282	9.88E + 04	4.80E + 07
Low-frequency Words	963	17.9	659	97	308	1.21E + 05	5.63E + 07
Frequency effect	107	9.9	80	35	26	2.24E + 04	8.27E + 06
Interaction	12	2.4	41	30	−30	4.71E + 03	6.52E + 06
Nonwords							
Clear pseudohomophones	824	10	621	60	205		
Degraded pseudohomophones	1163	17.9	869	172	297		

.10, with larger frequency effects in the degraded condition. University interacted significantly with stimulus quality ($p = .006$) and frequency ($p = .004$). Compared to the participants in Experiment 2, participants in Experiment 3 and Experiment 4 produced larger main effects of stimulus quality and word frequency in σ. The three-way interaction was not significant, $F < 1$.

Turning to τ, the main effects of stimulus quality, $F(1, 113) = 130.04$, $p < .001$, $MSE = 7{,}724.74$, $\eta^2 = .54$, and word frequency, $F(1, 113) = 29.70$, $p < .001$, $MSE = 4{,}812.90$, $\eta^2 = .21$, were significant. The stimulus quality × word frequency interaction was significant, $F(1, 113) = 9.78$, $p = .002$, $MSE = 3{,}816.32$, $\eta^2 = .08$, with larger frequency effects in the clear condition. University did not interact with either stimulus quality ($p = .166$) or frequency ($p = .82$). The three-way interaction was not significant, $p = .220$.

In summary, stimulus quality and word frequency were additive in means but interactive in all three ex-Gaussian parameters. Compared to clear words, degraded words produced larger frequency effects in μ and σ, which reflect the modal part of the distribution, but smaller frequency effects in τ, which reflects the tail of the distribution; this is consistent with the results of the individual experiments. The three-way interaction between university, stimulus quality, and frequency was not significant in any of the parameters, indicating that the same distributional tradeoffs were observed at all three sites. This is all the more impressive given the variability in lexical decision performance across the different participant populations. Specifically, compared to participants in Experiment 2, participants in Experiment 3 and Experiment 4 were producing larger main effects of stimulus quality and word frequency in means, accuracy, μ, and σ.

Vincentile analyses. The magnitudes of the frequency effect across the different vincentiles are plotted in Figure 5 (bottom panel), collapsed across all 116 participants. In line with the previous analyses, there is a larger frequency effect for degraded words in the fast vincentiles, and this trend reverses in the slowest vincentiles. More importantly, these opposing effects produce additivity in the analysis of means. We explored these effects by conducting an ANOVA with stimulus quality, word frequency, and vincentiles as within-participant variables. Interestingly, even with a Greenhouse-Geisser correction to compensate for potential violation of sphericity, the three-way interaction between stimulus quality, word frequency, and vincentiles was significant, $F(1.79, 205.73) = 6.26$, $p = .003$, $MSE = 7{,}602.17$, $\eta^2 = .05$, confirming that the shape of the frequency effect in the clear condition is markedly different from its counterpart in the degraded condition. To probe the three-way interaction, we separately examined the effects of stimulus quality and word frequency in the three fastest vincentiles versus the slowest vincentile. In the analysis based on the three fastest vincentiles, the interaction between stimulus quality and word frequency was significant, $F(1, 115) = 7.08$, $p = .009$, $MSE = 2{,}717.69$, $\eta^2 = .06$, with larger frequency effects for degraded words. In the analysis using the slowest vincentile, the interaction between stimulus quality and word frequency was also significant but in the opposite direction, $F(1, 115) = 4.15$, $p = .044$, $MSE = 13{,}242.66$, $\eta^2 = .04$; frequency effects were now larger for clear words. It is indeed noteworthy that ex-Gaussian and vincentile analyses, two very different approaches for examining RT distributions, converge on such similar conclusions.

General Discussion

This is the first study to examine the joint effects of stimulus quality and word frequency as a function of nonword type, using both conventional and distributional analyses, and the main findings can be summarized succinctly. In Experiment 1, using an LDT with legal nonword distracters (e.g., FLIRP), the joint effects of stimulus quality and word frequency were additive in means, accuracy, and distributional characteristics. These findings are compatible with the extant literature (Plourde & Besner, 1997; Yap & Balota, 2007). More intriguingly, when pseudohomophones (e.g., BRANE) were used as foils (Experiments 2, 3, and 4), opposing interactive effects in the underlying distributions were observed, producing apparent additivity at the level of the mean. This same basic pattern was reproduced in three experiments across three different universities, attesting both to the robustness of the effect and to the stability of RT distributional measures. Specifically, the composite analyses revealed that stimulus quality and word frequency interacted significantly in μ and σ, with larger frequency effects for degraded words. This interactive trend in μ (reflecting modal RTs) was offset by an opposing pattern in τ (reflecting slowest RTs), where frequency effects were larger for clear words. As the mean is the algebraic sum of μ and τ, these

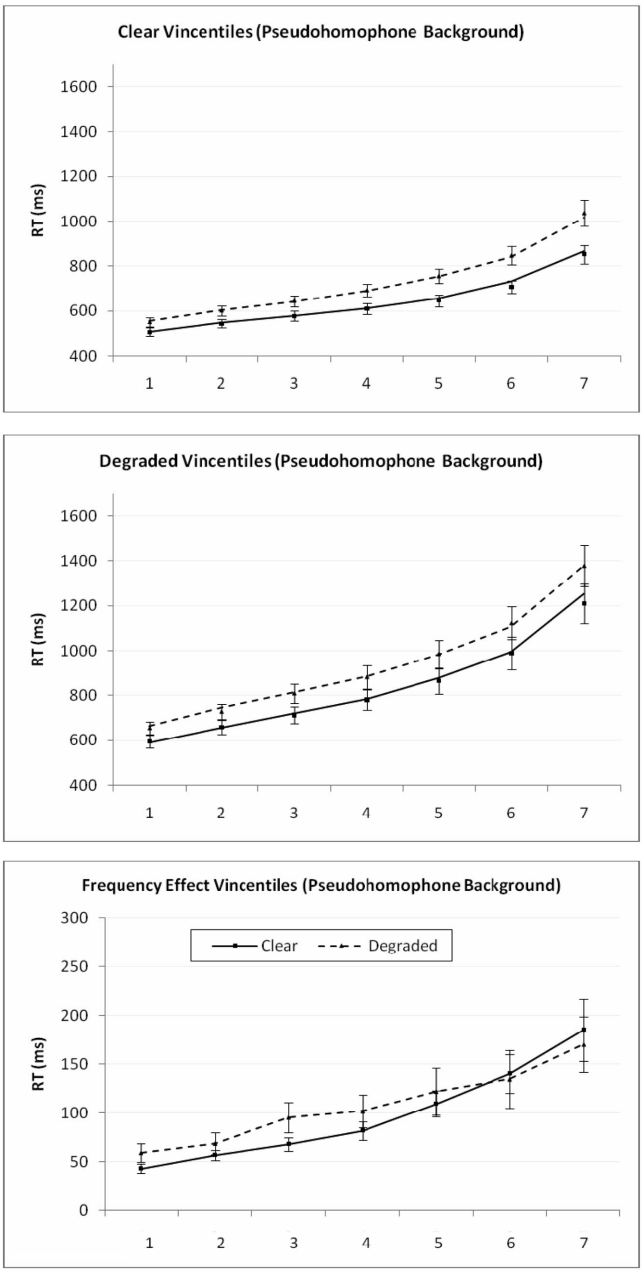

Figure 4. Lexical decision performance from Experiment 4 (pseudohomophone context) as a function of word frequency and vincentiles in the clear (top panel) and degraded (middle panel) conditions, along with frequency effects as a function of vincentiles (bottom panel). In the top and middle panels, participants' mean vincentiles (■ = high frequency, ▲= low frequency) are represented by data points and standard error bars. Best-fitting ex-Gaussian vincentiles are represented by lines (solid line = high frequency, dashed line = low frequency).

effects offset each other to produce additive effects in means. It is noteworthy that the same distributional tradeoffs were produced by distinct participant populations that varied along a number of dimensions, including overall response latency and accuracy (see Tables 3–5), Shipley vocabulary scores (see Table 1), and the magnitude of stimulus quality and frequency effects.

These data indicate that the joint effects of factors in means can be dissociated from joint effects in distributional characteristics. The trade-off between μ and τ is strikingly similar to the Stroop color-naming performance data reported by Heathcote et al. (1991) and Spieler et al. (1996). As discussed at the beginning of this article, they observed no difference between the congruent and baseline conditions in mean RTs, but ex-Gaussian analyses revealed that this null effect was spuriously produced by opposing effects in μ and τ. Likewise, in Experiments 2 to 4, the apparent additivity in means was driven by countervailing interactive effects in μ and τ, trends which are also reflected by the vincentile plots (see Figures 2 to 5). This is markedly different from Experiment 1, where the additivity in means was mirrored identically in ex-Gaussian parameters and vincentiles.

For ease of exposition, our discussion has, thus far, focused on how effects in means can be partitioned into μ and τ effects. This is, of course, an oversimplification. For example, one could argue that in ex-Gaussian analyses, μ and τ are not truly independent parameters. As mentioned in the Are Additive Effects Task-Specific section, the mean is the algebraic sum of μ and τ, suggesting that any effect on μ will necessarily also have an effect on τ. This could be responsible for the trade-off between μ and τ. Specifically, increased effects in μ could be accompanied by decreased effects in τ, which is the pattern observed in Experiments 2 to 4. We do not believe this to be a problem in the present analyses. First, the trade-off was only demonstrated when pseudohomophones, but not legal nonwords, were used. A fitting artifact should manifest itself in both nonword type conditions. Second, when we examined the correlations between μ and τ estimates across Experiments 2 to 4, the correlations were always positive, which is inconsistent with a trade-off.

In addition, it is important to note that ex-Gaussian fits yield a σ parameter that reflects variability of the modal RTs. The composite analyses indicate that when pseudohomophones were present, stimulus quality and frequency interacted significantly in σ, with larger frequency effects in degraded words. This interaction qualifies the interpretation that additive effects in means are due entirely to a trade-off between μ and τ. To see why this is the case, consider Table 6, which indicates a 24 ms interaction in σ. As can be seen, the σ frequency effect is much larger in degraded words (34 ms) than in clear words (10 ms), and this is driven primarily by a large σ for degraded low-frequency words. Because the estimate for σ (the standard deviation of the modal RTs) is disproportionately high for degraded low-frequency words, the estimate for τ will be correspondingly lowered, because there is effectively less skew to model. This will attenuate the τ frequency effect for degraded words.

There is always the possibility that the results from an RT distributional analytic procedure can be influenced by the particular model that is being fitted or the fitting process (Luce, 1986; Van Zandt, 2000, 2002). Although the σ interaction is certainly responsible for part of the trade-off between μ and τ, it is also very clear that the opposing effects cannot be fully explained away as a consequence of the fitting procedure. Here is where the vincentile analyses are useful. In particular, the vincentile analyses reveal precisely the same empirical pattern (see Figure 5); and, as mentioned, vincentizing is a nonparametric technique that plots raw data directly without making any assumptions about the true underlying RT distribution. Hence, we are confident about the observed trade-off in the present results, although we do recom-

Table 6
Means of Participants' Lexical Decision Response Time Means, Accuracy, Variance, Third Cumulant, and Ex-Gaussian Parameter Estimates as a Function of Stimulus Quality and Word Frequency Collapsed Across Experiments 2, 3, and 4, Pseudohomophone Background

Stimulus quality/word frequency	*M*	% Errors	Mu	Sigma	Tau	Var	3rd cum
Clear							
High-frequency words	654	2.5	491	45	164	4.64E + 04	3.22E + 07
Low-frequency Words	739	9	525	55	216	6.68E + 04	3.83E + 07
Frequency effect	85	6.5	34	10	52	2.04E + 04	6.12E + 06
Degraded							
High-frequency words	821	5.5	549	50	273	9.96E + 04	6.12E + 07
Low-frequency Words	908	13.3	619	84	293	1.11E + 05	5.62E + 07
Frequency effect	87	7.8	70	34	20	1.14E + 04	−4.96E + 06
Interaction	2	1.3	36	24	−32	−9.03E + 03	−1.11E + 07
Nonwords							
Clear pseudohomophones	823	9.7	590	59	233		
Degraded pseudohomophones	1059	13.7	723	112	336		

mend that all three parameters should be considered when interpreting the results of ex-Gaussian analyses and that vincentile plots should be examined for evidence of convergence or divergence (see Balota et al., in press, for further discussion of advantages and possible constraints in RT distributional analyses).

Foil Type Modulates the Relation Between Stimulus Quality and Word Frequency

The additive effects of stimulus quality and word frequency have traditionally been considered one of the most robust benchmarks in the word recognition literature. However, the generality of this finding is an issue. As noted, Yap and Balota (2007) have demonstrated that although additive effects are clearly obtained in lexical decision performance, one instead observes interactive effects in other word identification paradigms such as speeded pronunciation (replicated by O'Malley et al., 2007) and semantic classification. Moreover, the nature of the interaction, larger frequency effects for degraded words, is exactly the pattern predicted by leading computational models of word recognition, such as a parallel distributed processing model (Plaut et al., 1996), the dual route cascaded model (Coltheart et al., 2001), and the multiple read-out model (Grainger & Jacobs, 1996). This suggests that discrete stages, the traditional interpretation of additive effects, may not be a fixed property of word recognition processing dynamics but may instead reflect processes that are specific to the LDT. Specifically, lexical decision, unlike other word recognition tasks, places a premium on familiarity-based information for discriminating between words and nonwords (Balota & Chumbley, 1984; Balota & Spieler, 1999; Besner, 1983). Stimulus degradation undermines this familiarity, making it necessary for degraded words to undergo an additional normalization procedure prior to word identification. In speeded pronunciation, where familiarity is not useful for driving responses, one would not expect additive effects, which is indeed what Yap and Balota and O'Malley et al. reported. Furthermore, Yap and Balota also reported an interaction in semantic classification, where familiarity discrimination is also unlikely to play a role.

The departure point of the present experiments was the examination of lexical decision performance when pseudohomophones are used as foils. As discussed earlier, all previous studies investigating the joint effects of stimulus quality and word frequency have used legal nonwords, which allows familiarity to function as a viable dimension for word–nonword discrimination. In the context of pseudohomophones, which are orthographically legal and phonologically very similar to real words, the utility of familiarity for driving lexical decision responses is undermined. One expects the lexical decision mechanism to rely less on global lexical activity (a proxy for familiarity) and more on the activation levels of individual representations, making the LDT functionally more similar to speeded pronunciation and semantic classification. Hence, when familiarity is deemphasized in lexical decision, the early normalization stage and its attendant additive effects may no longer be mandatory. Instead, stimulus quality and frequency should interact, just as it does in other word recognition tasks (Yap & Balota, 2007).

Our results indicate that the type of nonword used in an LDT does modulate the joint effects of stimulus quality and frequency. When pseudohomophones are used, additive effects are seen in means, but this additivity is obscuring other effects at the level of the underlying RT distributions. In fact, the ex-Gaussian analyses on the composite data from Experiments 2–4 indicate that stimulus quality and word frequency interact significantly in μ and σ (see Table 6), with larger frequency effects for degraded words. This is precisely the predicted theoretical pattern from most computational models of lexical decision performance. Importantly, the corresponding vincentile plot also shows significantly larger frequency effects for degraded words in the early and modal vincentiles (see Figure 5). Indeed, follow-up tests based on data from the three fastest vincentiles (i.e., approximately the fastest 40% of the responses) indicate a significant interaction between stimulus quality and word frequency. This pattern, obtained across three universities, is particularly noteworthy because when legal nonwords (e.g., FLIRP) are used, the two factors are unequivocally additive in μ. One sees this not only in Experiment 1 but also in other studies featuring distributional analysis (Plourde & Besner, 1997; Yap & Balota, 2007). These results suggest that when familiarity-based information is less useful for driving lexical decisions, interactive effects of stimulus quality and word frequency can be detected in the modal portion of the RT distribution.

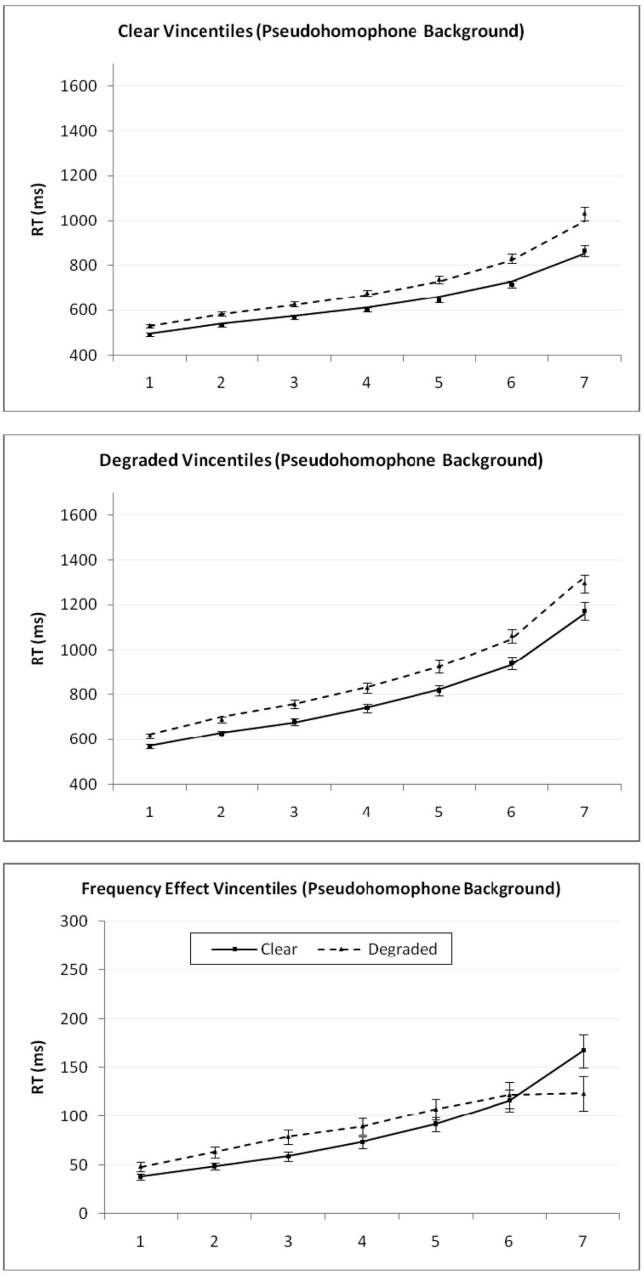

Figure 5. Lexical decision performance collapsed across Experiments 2–4 (pseudohomophone context) as a function of word frequency and vincentiles in the clear (top panel) and degraded (middle panel) conditions, along with frequency effects as a function of vincentiles (bottom panel). In the top and middle panels, participants' mean vincentiles (■ = high frequency, ▲ = low frequency) are represented by data points and standard error bars. Best-fitting ex-Gaussian vincentiles are represented by lines (solid line = high frequency, dashed line = low frequency).

Why Are Frequency Effects Smaller for Degraded Items in the Later Vincentiles?

One of the most interesting aspects of this study was the observation that although degraded words show larger frequency effects for the faster and modal RTs, this pattern abruptly reverses itself in the slowest RTs. The larger frequency effects for clear items in the slowest vincentiles offset the larger frequency effects for degraded items in the earlier vincentiles, effectively producing additive effects at the level of the mean. There are two ways one might accommodate this counterintuitive finding. First, these effects may reflect additional task-specific checking processes, which in this context refer to postlexical attention-demanding analytic processes (e.g., spelling verification or retrieval of semantic referents) that increase RTs (Balota & Spieler, 1999). Specifically, it is difficult to discriminate between low-frequency words and pseudohomophones, and so, whether low-frequency words are clearly presented or visually degraded, it is likely that they will undergo additional checking. In contrast, this may not apply to high-frequency words. Whereas degraded high-frequency words might undergo checking, clearly presented high-frequency words are sufficiently familiar that they may not require additional checking. Checking processes are more likely to be reflected in the τ component or the slowest RTs (Balota & Spieler, 1999), and this account predicts larger frequency effects in the slow RTs for clear words but more modest frequency effects for degraded words. This is because clear low-frequency words engage additional checking processes to a greater extent than clear high-frequency words, but all degraded items undergo checking, independent of frequency.

Of course, by invoking an additional checking process that operates after familiarity-based discrimination, the checking account increases the theoretical complexity of the machinery underlying lexical decision. Indeed, after examining the joint effects of nonword type (pseudohomophones vs. legal nonwords vs. illegal nonwords) and word frequency using distributional analyses, Yap et al. (2006) argued that the overall pattern of results was more consistent with a single-process random-walk model than a two-stage model that incorporated familiarity-based and checking processes (Balota & Spieler, 1999). We have three responses to this. First, although Yap et al. implemented a particular instantiation of a two-process model, their results obviously do not rule out two-process models in general (see their discussion on p. 1341). Second, the primary data that distinguished the two-process account from the random-walk perspective in Yap et al. (2006) came from the illegal nonword condition (e.g., XNIV). It is possible that lexical decision performance in that condition is qualitatively outside the scope of the standard LDT, because participants may be making orthographic legality decisions. Third, and most importantly, the constellation of findings in the present study presents an important challenge to any one-process model. We are not aware of any single-process model that has explicitly simulated opposing effects in different portions of the RT distribution. For example, consider the diffusion model, which suggests that lexical decision reflects the unitary accumulation of noisy evidence over time. Based on the geometry of the diffusion process, one would expect any effect observed in fast RTs to stay either the same size or become larger in slow RTs. Hence, going from fast to slow RTs, it should not be possible for an effect to become smaller in magnitude, much less reverse in direction. Yet, this is the precise empirical pattern we have observed across three experiments. Although the diffusion model has been remarkably successful in accounting for aspects of lexical decision performance (Ratcliff et al., 2004; Yap et al., 2006), we would argue that there are instances where two (or perhaps more) processes need to be invoked. For

example, as discussed in the Additive Effects and Models of Word Recognition section, the diffusion model had difficulty handling the additive effects of stimulus quality and word frequency in the presence of legal nonwords (O'Malley et al., 2007; Yap & Balota, 2007), a limitation that applies to other models of lexical decision performance. Likewise, it is unclear how the diffusion model can accommodate the consistent opposing interactive effects of stimulus quality and frequency in fast and slow RTs without incorporating a secondary process.

A second possible explanation for the crossover at the slowest vincentiles is that there is a temporal deadline or functional ceiling in lexical decision performance, which the degraded low-frequency words are up against. When participants are responding to such items, there is sufficient partial lexical activity to drive a "word" response when the deadline is reached. This will cause the slowest responses in the degraded low-frequency condition to be truncated, reducing distributional skew and attenuating the frequency effect in the slow RTs for the degraded words. Of course, one might ask why this functional ceiling exists when pseudohomophones but not legal nonwords are used. Nonword type affects word latencies in lexical decision; the more similar nonword foils are to words, the longer word latencies are (Stone & Van Orden, 1993; Yap et al., 2006). Hence, words in the pseudohomophone condition, compared to the legal nonword condition, will have longer latencies and are more likely to encounter the above-mentioned functional ceiling. One problem with the simple deadline account is that it predicts very little (or no) variance for the slowest RTs of degraded low-frequency words. This is inconsistent with the present results where, compared to the slowest RTs in the other three conditions, the slowest RTs of degraded low-frequency words were associated with the greatest variability. Permitting variability in the deadline (à la Coltheart et al., 2001) might address this, but these explanations are ultimately post hoc and require independent verification (see Wagenmakers, Ratcliff, Gomez, & McKoon, 2008, for further discussion of deadline vs. nondeadline models of lexical decision performance).

Implications for Models of Visual Word Recognition

How do the findings reported in this study inform the computational dual route cascaded (Coltheart et al., 2001) and multiple read-out (Grainger & Jacobs, 1996) models, which are able to carry out lexical decisions? As discussed at the beginning of this article, these models produce word lexical decisions when either local or global lexical activity exceeds their respective thresholds. In general, we are skeptical that these models, which are extensions of the logogen approach, can accommodate the present set of lexical decision data. In speeded pronunciation at least, simulations of the dual route cascaded model (and by inference, the architecturally similar multiple read-out model) yield interactive, not additive, effects of stimulus quality and word frequency (Reynolds & Besner, 2004). In order to produce additive effects in lexical decision, O'Malley et al. (2007) have suggested modifying the dual route cascaded model such that the output of the letter units level is thresholded during lexical decision. That is, activation does not reach the orthographic input lexicon until activity in the letter level has reached threshold. Thresholding at the letter level is of course functionally similar to an early perceptual normalization stage. As we have argued, thresholded processing (or perceptual normalization) in lexical decision is adaptive because it allows familiarity-based information to be recovered prior to word–nonword discrimination. Furthermore, the implication here is that the dual route cascaded model is sensitive to task demands and is able to appropriately engage cascaded and thresholded letter-level processing during speeded pronunciation and lexical decision, respectively. The preceding account handles additive effects of stimulus quality and frequency in lexical decision, in both means and distributional characteristics, which is the pattern observed when legal nonwords are used as distracters.

This account is complicated by the results of the present study, which demonstrated that the presence of pseudohomophone foils produces an interaction between stimulus quality and word frequency in modal RTs, with larger frequency effects for degraded words. How might the model accommodate this? When word-like nonwords (i.e., pseudohomophones) are used as foils, global lexical activity is less diagnostic for word–nonword discrimination. We have suggested that thresholded processing is useful for recovering familiarity-based information. In lexical decision contexts where the utility of this information is undermined, the model should revert to cascaded processing, which yields interactive effects. However, note that this only accounts for interactive effects in the modal portion of the RT distribution. It is unclear how one can handle the trade-off in the tail of the distribution without positing additional checking or deadline-based mechanisms (see the previous section). Finally, we need to emphasize that the discussion in this section is post hoc and, in some ways, counterintuitive. A priori, there is no principled reason why any computational model of visual word recognition would selectively engage thresholded processing in lexical decision and cascaded processing in pronunciation. At the very least, these results support the adaptive nature of subtle characteristics of the lexical processing system to task constraints (Balota & Yap, 2006).

Of course, the account above rests on the assumption that additive effects of stimulus quality and word frequency are more consistent with thresholded than cascaded processing. So far, we have argued that additivity is easier to reconcile with a stage framework, given that cascade models make incorrect predictions about the relationship between means and variances (Roberts & Sternberg, 1993). However, it is important to note that the experimental tasks Roberts and Sternberg based their simulations on did not include the LDT, and it is unclear if the superiority of the stage account holds when one examines additivity in lexical decision performance. We explored this by conducting a supplementary analysis on the data in Experiment 1, where additive effects in means and ex-Gaussian characteristics were observed. Specifically, we computed the variance-change statistic proposed by Roberts and Sternberg for adjudicating between the stage and cascade models; this statistic reflects the ratio of variance effect size to mean effect size. In the simulations described by Roberts and Sternberg, cascade models in general yielded variance-change statistics larger than those observed behaviorally in their datasets. Interestingly, based on the variance-change statistic for our dataset, it turns out that the results in Experiment 1 can also be accommodated by cascade models. In other words, the clear-cut additivity in Experiment 1 may reflect either thresholded processing (the position we have been taking) or cascaded processing. Cascade models can produce additive effects in means and variances, but it is necessary to make additional assumptions. For example, assume we start out with a two-process cascade model and process A is selectively influenced by stimulus

quality whereas process B is selectively influenced by word frequency. For the model to produce stage-like additivity, each process has to be joined to a fixed slow process that is unaffected by either stimulus quality or word frequency; this effectively becomes a three-process model. Specifically, one might have a fast process that is affected by stimulus quality, followed by a slow process that is influenced by neither factor, followed by a fast process that is affected by word frequency.

The important point here is that the additive effects of word frequency and stimulus quality (with pronounceable nonwords) should not be taken as evidence that eliminates all possible cascade models. However, we believe that such accounts of the present results become quite cumbersome and ad hoc. To produce additive effects in Experiment 1, one requires a cascade model (e.g., the three-process model described above) where the two processes of interest are mated to a slow process that is uninfluenced by both stimulus quality and frequency. When pseudohomophones are used, the slow process stops operating, and the model reverts back to what are functionally two processes, yielding interactive effects in the modal portion of the RT distribution. As noted, we prefer the stage model account for the additive effects of word frequency and degradation for the following reasons: (a) It is relatively simple and parsimonious, (b) it also accommodates performance in another familiarity-based task (e.g., the additive effects of stimulus quality and set size observed in memory scanning), and (c) it is motivated by existing accounts of lexical decision performance. Most important, when familiarity-based information is disabled by including pseudohomophones as the nonwords, thereby forcing individuation of the lexical representations, the system reverts back to the interactive cascadic processes that are typically engaged in most lexical processing tasks. For the cascaded account to be equally viable in accounting for the full set of data, a theorist needs to explain why and how processes in the cascade model are modulated by nonword type. For example, why would the presence of legal nonwords engage additional slow processes that are not modulated by experimental factors? How do the processes reconfigure themselves in a pseudohomophone context, and what might be the reason for this? A modeler trying to simulate these processes faces the additional hurdles of having to explicitly specify which factors influence which processes and to consider other parameters like the shape, location, and spread of the response criterion's distribution, as well as the specific activation function mediating input and output (Roberts & Sternberg, 1993).

Mapping Distributional Characteristics Onto Cognitive Processes

Based on the foregoing discussion, it may be tempting to conclude that lexical retrieval processes are reflected by the μ component and the fast and modal RTs, whereas postlexical checking processes are reflected by the τ component and the slowest RTs. This view is seductive, but clearly oversimplistic. Although some early theorists attempted to map specific processes onto different distributional parameters (Hohle, 1965; McGill & Gibbon, 1965), we agree with Heathcote et al. (1991), Schwarz (2001), and Van Zandt (2002) that the ex-Gaussian model is probably too simple for such substantive cognitive attributions to be made. In that light, it is necessary to emphasize that our "mapping" between distributional characteristics and different classes of cognitive operations should be construed as metaphorical at this point. It is only appropriate to ascribe specific processes to specific distributional parameters when one has a computational model of the targeted task. The model needs to be explicit and well-specified enough for the researcher to predict a priori how experimental manipulations might modulate different RT parameters (Balota & Spieler, 1999). In general, the present distributional analyses should be viewed as a descriptive tool that partitions factor effects into distributional shifting and skewing. These finer-grained effects (especially the types of tradeoffs observed in the current results) can then be used to impose finer constraints on extant models and processes.

Conclusions

The results from the present experiments further underscore the utility of distributional analyses in better understanding the influence of multiple variables. As Heathcote et al. (1991) have pointed out, relying on analyses of means alone can potentially lead to inadequate or misleading conclusions. In this instance, distributional analyses and analyses of means yielded markedly divergent outcomes. At the level of the mean, additive effects of stimulus quality and word frequency were observed in lexical decision, regardless of nonword type. At the level of the RT distribution, however, interactive effects in μ were produced only when pseudohomophones, but not legal nonwords, were used, and this pattern was replicated in three different populations of subjects. Collectively, these findings indicate that distributional analyses not only resolve data at a finer level of granularity but may also provide more faithful insights into underlying processes.

References

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

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Learning, memory, and thinking* (Vol. 1, pp. 242–293). San Francisco: Freeman.

Balota, D. A., & Abrams, R. A. (1995). Mental chronometry: Beyond onset latencies in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1289–1302.

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.

Balota, D. A., & Yap, M. J. (2006). Attentional control and flexible lexical processing: Explorations of the magic moment of word recognition. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing* (pp. 229–258). Hove, England. Psychology Press.

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (in press). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language*.

Becker, C. A., & Killion, T. H. (1977). Interaction of visual and cognitive effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 389–401.

Besner, D. (1983). Basic decoding components in reading: Two dissociable feature extraction processes. *Canadian Journal of Psychology, 37,* 429–438.

Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 813–840.

Coltheart, M., Davelaar, E., Jonasson, J., & 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, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256.

Cousineau, D., Brown, S. D., & Heathcote, A. (2004). Fitting distributions using maximum likelihood: Methods and packages. *Behavior Research Methods, Instruments, & Computers, 36,* 742–756.

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

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

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

Heathcote, A., & Brown, S. D. (2004). Reply to Speckman and Rouder: A theoretical basis for QML. *Psychonomic Bulletin & Review, 11,* 577–578.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin, 109,* 340–347.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology, 69,* 382–386.

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

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers, 28,* 203–208.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330.

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

McGill, W. J., & Gibbon, J. (1965). The general gamma distribution and reaction times. *Journal of Mathematical Psychology, 2,* 1–18.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76,* 165–178.

O'Malley, S., Reynolds, M., & Besner, D. (2007). Qualitative differences between the joint effects of stimulus quality and word frequency in lexical decision and reading aloud: Extensions to Yap and Balota. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 451–458.

Paap, K. R., Newsome, S. L., McDonald, J. E., & Schvaneveldt, R. W. (1982). An activation-verification model for letter and word recognition: The word superiority effect. *Psychological Review, 89,* 573–594.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114,* 273–315.

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

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.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1988). *Numerical recipes: The art of scientific computing.* Cambridge, England: Cambridge University Press.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Review, 86,* 446–461.

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

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356.

Reynolds, M., & Besner, D. (2004). Neighborhood density, word frequency, and spelling-sound regularity effects in naming: Similarities and differences between skilled readers and the dual route cascaded computational model. *Canadian Journal of Experimental Psychology, 58,* 13–31.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer and S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 611–653). Cambridge, MA: MIT Press.

Schneider, W., Eschman, A., & Zuccolotto, A. (2001). *E-prime user's guide.* Pittsburgh: Psychology Software Tools, Inc.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers, 33,* 457–469.

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *Journal of Psychology, 9,* 371–377.

Shulman, H. G., Hornak, R., & Sanders, E. (1978). The effects of graphemic, phonetic, and semantic relationships on access to lexical structures. *Memory and Cognition, 6,* 115–123.

Speckman, P. L., & Rouder, J. N. (2004). A comment on Heathcote, Brown, and Mewhort's QMLE method for response time distributions. *Psychonomic Bulletin & Review, 11,* 574–576.

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in normal older adults and individuals with senile dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 461–479.

Stanners, R. F., Jastrzembski, J. E., & Westbrook, A. (1975). Frequency and visual quality in a word-nonword classification task. *Journal of Verbal Learning and Verbal Behavior, 14,* 259–264.

Sternberg, S. (1967). Two operations in character recognition: Some evidence from reaction-time measurements. *Perception & Psychophysics, 2,* 45–53.

Sternberg, S. (1969a). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30,* 276–315.

Sternberg, S. (1969b). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist, 57,* 421–457.

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

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review, 7,* 424–465.

Van Zandt, T. (2002). Analysis of response time distributions. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 461–516). Hoboken, NJ: John Wiley & Sons.

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 & Language, 58,* 140–159.

Yap, M. J., & Balota, D. A. (2007). Additive and interactive effects on response time distributions in visual word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 274–295.

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

(*Appendix follows*)

Appendix

Table A1
High-frequency words

AIR	FILM	LIFE	SORT
ARTIST	FINAL	LIKE	SOUND
BALL	FIRE	LONG	SOUTH
BEHIND	FOOD	LOSS	STAGE
BOTTOM	FORCE	MACHINE	START
CARRY	FREE	MAJOR	STATION
CASE	GAS	MONEY	STORY
CAUSE	GIRL	MOTOR	STUDENT
CHANCE	GOAL	MUSIC	STUDY
CHIEF	GUN	NAME	STYLE
CHILD	HAPPEN	NOVEL	
CHURCH	HIGH	PAID	THERE
CLAIM	HOME	PARTY	THING
CLOSE	HOTEL	PICTURE	TODAY
COLD	HOUSE	PIECE	TOP
COLOR	HUMAN	PLACE	TOTAL
CUT	INDEX	PLANE	VIEW
DESK	JOB	PRETTY	VOICE
DINNER	KITCHEN	ROAD	WAIT
DOCTOR	LABOR	SEA	WELL
DRIVE	LAND	SHIP	WISH
EIGHT	LARGE	SIDE	WOMAN
FATHER	LATER	SIGHT	WORLD
FEEL	LESS	SIGN	YELLOW
FEET	LETTER	SIX	YOU

Table A2
Low-frequency words

ADEPT	EMBARK	MERRY	SOCK
ANVIL	EXIT	MESH	SPICE
APRON	FARE	MINT	SPIN
ARID	FLOAT	MULE	SPOIL
AWE	FLU	MUNCH	SPOON
BANJO	FLUTE	OUNCE	SPY
BEAN	GAZE	PAIL	STACK
BEGGAR	GORGE	PEACH	STINK
BOOM	GRAPE	PLUMP	STOOL
CANON	GRAVEL	POLAR	THORN
CARVE	HASTE	QUEEN	THRILL
CAVERN	HICK	RIM	TORCH
CHEER	HOBBY	ROAST	TRAITOR
CHORE	HOWL	RUDE	TRAMP
COMIC	JARGON	RULER	TROUT
CONCEDE	JOLT	RUMOR	UNCLE
CORAL	LASS	SCOOP	VALVE
COWARD	LOFT	SCRATCH	VEIL
CRATE	LORD	SCRUB	VILE
DENSE	LUST	SEAM	WEAVE
DUAL	MAGNET	SERVANT	WELD
DUMMY	MARCH	SHRUG	WITCH
DUNE	MASK	SKULL	WIZARD
DUSK	MAYOR	SLAB	WRECK
EDIT	MENTOR	SMASH	ZOO

Table A3
Legal-nonwords

AMANG	DROMB	LARCH	SHOIR
BAW	DUT	LARN	SHOV
BEIRN	EAP	LEMM	SINDOL
BELONS	FANE	LOAP	SINDWHA
BIHEAV	FARN	LODDER	SKORT
BIRGAN	FEARTY	LOICE	SKULP
BOAK	FERAN	LOUST	SLOOK
BOIM	FIMUS	MANT	SLOTE
BOITT	FLOD	MATTUL	SOID
BONSE	FLOOCE	MEARNIN	SOIN
BORST	FLOOM	MINY	SOOWOI
BOUD	FLORT	MOIR	SORCH
BOUP	FOLT	MOTCH	SORVE
BOUR	FRAKLE	MOUDEN	SOUT
BROME	FRANT	MUTH	SOUZE
BRULLAN	FREAM	NACE	SPART
BRUTH	FRIM	NARSE	SPRUD
CAMFAR	FROLE	NAZE	STOMPAD
CAMMEN	FROTE	NEIZE	STOOM
CAMPLOO	FROUDA	NELD	STRALE
CANSAR	FURSE	NOOTY	SUKE
CARSE	FUTHOR	NOUCE	SWARL
CATTEN	GEAM	NOUTE	TADE
CEROOR	GEMBOL	OUZY	TARKOY
CHALLO	GEWS	PANSOL	TAZE
CHOIK	GIRDAN	PEIP	THOY
CHOOK	GLAE	PHANNE	THRA
CIRBAN	GOACE	PHON	THRAT
CLORK	GREPE	PILLOT	THRID
COL	GRIE	POPLE	TINCE
CONSAR	GRINE	PORPLE	TOILE
CONSUN	GUTE	PRESOI	TOIM
COOGE	HANNY	PRIVE	TONSHA
COOM	HEAKE	RAF	TOPE
COUM	HOARO	RATHAM	TRAFF
CRADET	HOD	REAB	TRUBE
CRAT	HOF	REAZE	TRULE
CUFFAY	HOIZE	RESURS	UKE
DAFF	HOOVE	ROIDE	UTE
DALT	ILLWOY	ROIP	VALLEN
DEACE	JARM	ROULY	VOCAME
DEAT	JILE	RUNKLE	VULT
DEBBLE	JONTLE	SAMENT	WAIRD
DECE	JORK	SANTRO	WARTH
DESCARS	JUNDER	SARTEN	WEAGE
DITOLE	KEF	SATCH	WOUT
DOULE	KIPY	SAUK	WRONT
DOUM	KNOIT	SERPRAZ	WUTE
DRAS	KONDY	SHAT	YEAK
DREM	LADDLE	SHOAP	YOULD

Table A4
Pseudohomophones

AIP	KOPY	AKE	LAMM
ALLWAYS	LEECE	AMUNG	LEEP
BAID	LEEST	BALANSE	LERN
BEAP	MAYDEN	BEED	MENT
BEEK	MENY	BEEM	MITH
BERN	MOAR	BETT	MUTCH
BIRST	NAWTY	BLAIM	NEACE
BOR	NEAD	BOTE	NERSE
BOUNSE	NOIZE	BRAIT	NOOZE
BRETH	NOZE	BRUME	OTE
CAIGE	PAIJE	CAIM	PEAP
CARBUN	PEEPLE	CAREAR	PENSIL
CAUST	PERPLE	CEL	PHAN
CHEAK	PHIRM	CHELLO	PHUNNEL
CHOAK	PIRCH	CLIRK	PLEED
CLOO	PROSEED	COAM	PRUVE
COFFEY	RADE	COMFERT	RATHFOO
COMMEN	REELY	CONSERT	RESORSE
CORSE	RINKLE	COTTEN	RITHUM
CRAIT	ROAB	CROOD	ROAP
CUNSENT	ROZE	DAIT	RUF
DEFF	SAIN	DELT	SEAK
DET	SED	DETALE	SEET
DILE	SEEZE	DOAM	SEMENT
DREEM	SENTRAL	DRES	SERTON
DRUMB	SEZ	DUBBLE	SHEAP
EAZY	SHEAT	FALT	SHURE
FAMUS	SHUV	FETHER	SIRCH
FIRN	SKALP	FLAIM	SKERT
FLEACE	SLEAK	FLUD	SLITE
FLURT	SOAL	FOME	SOKE
FONE	SPERT	FORSE	SPRED
FOURTY	STAIL	FRAIM	STROLE
FRALE	STUK	FREKLE	SURVE
FRITE	SUTCH	FRUM	SWERL
FRUNT	TAYP	GAIM	TENCE
GAMBEL	TERKEY	GARDUN	THAY
GLOE	THEFE	GOOCE	THRED
GOTE	THRET	GRAE	TIPE
GRONE	TODE	GRUPE	TOLE
HAF	TOOB	HED	TRALE
HEERO	TROFF	HEEVE	TROLE
HOAP	TRUPE	HOKES	TUME
HOZE	TYDE	HUNNY	VACUME
JALE	VALT	JENDER	VILLEN
JENTLE	WAIGE	JERM	WEIT
JIRK	WEERD	JOAK	WEET
KAF	WIRTH	KANDY	WITE
KEAP	WRENT	KNET	WURK
KOME	YEELD	KOMPLIT	YOAK

Received January 23, 2007
Revision received October 25, 2007
Accepted November 12, 2007 ■