Reading Aloud: Qualitative Differences in the Relation Between Stimulus Quality and Word Frequency as a Function of Context

Shannon O'Malley and Derek Besner
University of Waterloo

Virtually all theories of visual word recognition assume (typically implicitly) that when a pathway is used, processing *within* that pathway always unfolds in the same way. This view is challenged by the observation that simple variations in list composition are associated with qualitative changes in performance. The present experiments demonstrate that when reading aloud, the joint effects of stimulus quality and word frequency on response time are driven by the presence/absence of nonwords in the list. Interacting effects of these factors are seen when only words appear in the experiment, whereas additive effects are seen when words and nonwords are randomly intermixed. One way to explain these and other data appeals to the distinction between cascaded processing (or interactive activation) on the one hand versus a thresholded mode of processing on the other, with contextual factors determining which mode of processing dominates.

Keywords: reading aloud, word frequency, stimulus quality

Supplemental materials: http://dx.doi.org/10.1037/a0013084.supp

Word frequency is probably the most well-studied psycholinguistic factor over the last 40 years or so. All theories of visual word recognition accommodate the effect of this factor in various ways (among others, see Adelman, Brown, & Quesada, 2006; Balota & Chumbley, 1984; Becker, 1976; Besner, 1983; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Forster & Chambers, 1973; Frederiksen & Kroll, 1976; Grainger & Jacobs, 1996; McCann & Besner, 1987; McClelland & Rumelhart, 1981; Morton, 1969; Murray & Forster, 2004; Norris, 2006; Perry, Ziegler, & Zorzi, 2007; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). It is not surprising that there are so many different accounts of how word frequency exerts its effect(s): A main effect does not place strong constraints on theory building. In contrast, the joint effects of multiple factors are much more constraining in this regard. The focus of the current article is on the joint effects of word frequency and stimulus quality (how easily a word is taken up by the processing system) in the context of reading aloud. The results reported here speak to a number of core issues in visible language processing: in particular, (a) the extent to which processing is dynamic or static, (b) thresholded or cascaded (or engaged in interactive activation [IA]), and (c) automatic or context dependent.

Skilled readers are remarkably adept at reading words that have been distorted or rendered difficult to take up in various ways (e.g., by reducing stimulus quality, cAse mIxInG, or masking). Stanners, Jastrzembski, and Westbrook (1975) were the first to report that the joint effects of stimulus quality (reduced contrast generated by covering the screen with a sheet of acetate) and word frequency were additive on response time (RT) in the context of lexical decision. In the ensuing decades, researchers have replicated this observation a number of times using a number of different ways to reduce stimulus quality (Balota & Abrams, 1995; Becker & Killion, 1977; Borowsky & Besner, 1993; Norris, 1984; O'Malley, Reynolds, & Besner, 2007; Plourde & Besner, 1997; Wilding, 1988; Yap & Balota, 2007).[1] Curiously, *computational* accounts of visual word recognition have to date largely ignored this pattern. The singular exception is Plaut and Booth's (2000, 2006) computational parallel distributed processing (PDP) model, which purports to simulate this additive pattern. However, Besner, Wartak, and Robidoux (2008) have demonstrated that Plaut and Booth's model actually fails in this regard because the joint effects of stimulus quality and word frequency yield a nonmonotonic function (underadditivity, additivity, and overadditivity) depending on the size of the stimulus quality effect, whereas skilled readers yield additivity across a wide range of stimulus qualities. Theories of visual word recognition that do not address basic findings, such as the joint effects noted above (obviously), require development (indeed, such effects ought to be among the "benchmark" phenomena for computational models).

Before considering how such theories might be modified to accommodate these results, several other closely related findings merit consideration. In particular, Yap and Balota (2007) and

Shannon O'Malley and Derek Besner, Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada.

Correspondence concerning this article should be addressed to Shannon O'Malley, Department of Psychology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. E-mail: somalley@uwaterloo.ca

---

[1] Wilding (1988) and Norris (1984) both reported experiments in which there was an interaction between stimulus quality and word frequency in lexical decision when there was a very long intertrial interval (more than 3 s). However, both authors reported additive effects of these factors when the intertrial interval was considerably shorter. Wilding argued that the interaction seen with a long intertrial interval says little or nothing about reading per se and more about attention and recovery from long foreperiods.

O'Malley et al. (2007) reported additive effects of stimulus quality and word frequency in lexical decision but an interaction between these factors in reading aloud.[2] Yap and Balota also reported an interaction between stimulus quality and word frequency in the context of a semantic categorization task when the words were not members of the target category. In short, the pattern of joint effects between stimulus quality and word frequency is clearly more complex than generally appreciated to date.

These new findings raise a basic empirical question that should be addressed before reconsidering various theoretical accounts. What role is played by the presence/absence of nonwords in the relation between stimulus quality and word frequency, given that task (lexical decision, reading aloud, and semantic categorization) and the presence/absence of nonwords are confounded? That is, when stimulus quality and word frequency interact (reading aloud and semantic categorization), nonwords are not part of the stimulus set, whereas when stimulus quality and word frequency have additive effects (lexical decision when the nonwords are orthographically legal), nonwords are part of the stimulus set.

Lexical decision, by definition, involves discriminating between letter strings that spell a word and letter strings that do not; the presence of nonwords is intrinsic to the task. Of course, it is possible to add nonwords to the semantic categorization task. However, doing so invites the criticism that this changes the nature of the task in ways that are not well understood. Thus, it might be difficult to convince various theorists that such a manipulation is important in the context of this task (but see Forster & Hector, 2002). In contrast, there is a long history of experiments on reading words aloud in which nonwords are sometimes present and sometimes not (e.g., Andrews, 1992; Forster & Chambers, 1973; Frederiksen & Kroll, 1976; McCann & Besner, 1987; Monsell, Patterson, Graham, Hughes, & Milroy, 1992; Reynolds & Besner, 2005, 2008). One way to investigate the confounding of task and the presence/absence of nonwords is to have participants read aloud in an experiment in which word frequency and stimulus quality are manipulated and nonwords are randomly intermixed, as compared with when only words appear in the experiment. Three such experiments are reported here.[3]

Before reporting these new experiments, we first provide a brief review of a basic distinction in the way that psycholinguists interested in the processing of visible language think about how the special purpose modules that underlie visual word recognition (e.g., feature level, letter level, word level) communicate with each other. In large part, researchers have typically assumed that how such processing unfolds is fixed—for example, that *how* the processing of high versus low frequency words unfolds over time is not affected by the experimental context (in this case the presence vs. absence of nonwords). It is proposed here instead that several different processing modes operate, but *when* and *where* in the processing sequence each does so depends on the context. Following this brief review, we turn to a new contextually based prediction (the lexicalization hypothesis) concerning the effect that the presence of nonwords has on the joint effects of stimulus quality and word frequency when reading aloud.

The Thresholded–Cascaded/IA Distinction

Sternberg (1969) proposed that many mental processes occur in a discrete series, one beginning when another ends. For example, Process B starts only after Process A finishes. If Factor 1 affects Process A but not Process B, and Factor 2 affects Process B but not Process A, then additive effects of these factors on RT should be observed. In this formulation, additive effects of two factors on RT are the signature of distinct processes that occur sequentially.

McClelland (1979; see also Ashby, 1982) proposed a different account in which mental processes are cascaded. In this formulation, processes overlap in time. For example, as soon as Process A starts, it sends activation to Process B, which begins without awaiting the completion of Process A. This idea was extended by McClelland and Rumelhart (1981), such that ongoing activation of Process B feeds back to Process A (IA).

Computational accounts of visible language processing have typically ignored the idea of discrete mental processes. Instead, these models are almost invariably cascaded and often engaged in IA between various levels. One central question here concerns whether IA can produce systematically additive effects of two factors on RT. To date, we are aware of no existence proof to this effect (see Besner, 2006; Besner et al., 2008; Borowsky & Besner, 2006). In contrast, very simple cascade models are, to a first approximation, able to produce additive effects on *mean* RT under certain conditions (Ashby, 1982; McClelland, 1979; Roberts & Sternberg, 1993).

Despite widespread acceptance of the idea that processing in visual word recognition is cascaded, there are circumstances in which such an account is not easy to reconcile with the data produced by skilled readers. For example, Besner and Roberts (2003) reported that when reading nonwords aloud, RT increased as letter length increased, a reduction in stimulus quality also increased RT, and the joint effect of these two factors was additive. In contrast, simulations with the most successful computational model at that time (Coltheart et al.'s, 2001, dual route cascaded [DRC] model) yielded an interaction in which longer letter strings were *less* affected by low stimulus quality than were shorter letter strings. To fix this problem, Besner and Roberts proposed that the DRC model be modified such that the letter level is thresholded rather than allowing it to cascade. This way, the effect of reduced stimulus quality would not affect the model beyond the letter level, and given that the letter length effect arises from subsequent serial left to right assignment of phonemes to letters, the joint effects of stimulus quality and letter length would be additive on RT. Unpublished simulation work in our laboratory confirms that changing the model in this way is successful in that it now produces additive effects of letter length and stimulus quality when reading nonwords aloud.

Another computational account of reading aloud is Perry et al.'s (2007) connectionist dual process (CDP+) account. The lexical

---

[2] To manipulate stimulus quality, Yap and Balota (2007) rapidly alternated a mask and the letter string, whereas O'Malley et al. (2007) used contrast reduction. Both manipulations yielded an interaction between stimulus quality and word frequency in reading aloud and additive effects of stimulus quality and word frequency in lexical decision.

[3] Alexander Pollatsek (personal communication, March 2008) raised an important direction for future research. How would the present claims play out in the context of eye movement studies? One approach might be to have participants read prose and use a high proportion of very low frequency words that are unknown to the participants so as to mimic the presence of nonwords.

route in this model is taken from the DRC model, whereas the nonlexical route starts with the graphemic buffer and then uses a two layer assembly network. For present purposes, the important characteristic of this model is that the connection from the letter level to the nonlexical route is functionally thresholded, which should result in additive effects of stimulus quality and letter length when reading nonwords aloud (see Perry et al., 2007, p. 283). Conceptually, this suggests that Sternberg's (1969) notion of discrete processing stages that has been ignored by psycholinguists for almost 3 decades has been rediscovered by some computational modelers.

A second example in which cascaded processing appears to be problematic concerns the joint effects of neighborhood density and stimulus quality. Neighborhood density refers to the number of words (N) that can be generated by changing one letter at a time in a letter string (Coltheart, Davelaar, Jonasson, & Besner, 1977; but see Mulatti, Reynolds, & Besner, 2006, who report data suggesting that N reflects phonemes rather than letters). As N increases, the time to read both words and nonwords aloud decreases (e.g., Andrews, 1992; McCann & Besner, 1987; among others). Reynolds and Besner (2004) reported that when reading nonwords aloud, the joint effects of stimulus quality and N were additive on RT, whereas simulations with the DRC model produced an interaction in which low stimulus quality slowed low N nonwords more than high N ones. Reynolds and Besner suggested that the same modification to DRC as proposed earlier: threshold the letter level rather than allowing it to cascade.

By way of summary then, thresholding the letter level rather than allowing it to cascade provides a simple way to allow two otherwise very successful computational models of visual word recognition to simulate the performance of skilled readers with respect to the joint effects of stimulus quality and letter length, and stimulus quality and N, both when reading nonwords aloud.

There is, however, a problem with thresholding the letter level in DRC and CDP+. Doing so would produce additive effects of stimulus quality and word frequency when reading aloud, but as we noted earlier, these two factors interact in this context (O'Malley et al., 2007; Yap & Balota, 2007), and Reynolds and Besner (2004) showed that the DRC model also produces an interaction between stimulus quality and word frequency when reading aloud. At first blush then, there is a contradiction across the three findings we have discussed so far. Thresholding the letter level so that the effect of stimulus quality does not affect processing beyond that level is sufficient when considering the joint effects of stimulus quality and letter length, and stimulus quality and N, but it fails in the case of the joint effects of stimulus quality and word frequency, all in the context of reading aloud.

However, this contradiction may be more illusory than real. When the joint effects of stimulus quality and a second factor (letter length; N) were additive, participants were reading *nonwords* aloud. When the joint effects of stimulus quality and word frequency interacted, only *words* appeared in the experiment. The proposal advanced here is that when reading nonwords aloud the letter level is thresholded, but when reading only words aloud the letter level cascades through to the word level where performance is affected by word frequency and, hence, yields an interaction (as reported both in the data from skilled readers; O'Malley et al., 2007; Yap & Balota, 2007; and in the DRC model as reported by Reynolds & Besner, 2004).

The obvious theoretical question concerns why early processes would be so flexible in terms of their configuration. What benefit does this confer on the process of reading aloud? The hypothesis advanced here is that cascaded processing risks lexical capture when stimulus quality is low; a nonword may activate a word sufficiently strongly that the reader mistakenly reads it as a word instead of the nonword. To reduce this problem, participants can threshold the letter level. Rather than attempt to alter the parameter settings on the fly, it is easier to set them (unconsciously) for a block of trials so that processing is either in cascaded mode or in thresholded mode.

At this juncture a reader might reasonably object that the account offered here is consistent with the data, but post hoc. What new predictions, if any, does this account make? If, as suggested above, including nonwords in the experiment serves to modulate the way that processing unfolds, a straightforward prediction is that the joint effects of stimulus quality and word frequency will be additive on RT rather than interact when nonwords are randomly mixed together with the words. This is because the letter level will now be thresholded rather than cascaded so as to avoid the problem of lexical capture in response to nonwords when stimulus quality is low. Two experiments are reported here that test this prediction. A third experiment directly compares the condition in which words and nonwords were mixed together with one in which only words were presented.

Experiment 1

Method

Participants. Thirty-two undergraduate students from the University of Waterloo (Waterloo, Ontario, Canada) were each paid $4.00 for their participation. All were native English speakers and reported normal or corrected-to-normal vision.

Stimuli. The stimulus set consisted of 200 words and 200 nonwords and is available, along with the item means, online (see the supplemental materials). The 100 high frequency words (mean count per million = 411.6) and 100 low frequency words (mean count per million = 17.3) were taken from Yap and Balota (2007). The mean number of letters in the words was 4.8 (range = 3–7). The mean orthographic neighborhood size (N values; see Coltheart et al., 1977) for the high frequency words was 4.8, and the mean summed bigram frequency was 6370. For low frequency words, the mean orthographic neighborhood size was also 4.8, and the mean summed bigram frequency was 6149. The nonwords, taken from O'Malley et al. (2007), were matched to the words for length (M = 4.8, range = 3–7), and the mean orthographic neighborhood size was 8.9. We split the nonwords into four lists, and we rotated them through conditions using a partial Latin square such that each nonword list was presented with each word list equally often across participants, resulting in eight lists.

The stimuli were rotated through stimulus quality conditions across participants, who were assigned to a counterbalancing condition on the basis of order of arrival in the laboratory, with words and nonwords randomly intermixed. The letter strings were displayed in 16-point Times New Roman font on a black background (Red, Green, Blue [RGB] 0, 0, 0). In the bright condition, the letter strings appeared in RGB (120, 120, 120); in the dim condition, they appeared in RGB (36, 36, 36). The lighting in the room was

dim (a measure of luminance at the level of the screen would have been preferable, but this laboratory lacks this expensive piece of equipment).

Apparatus. The data were collected on a Pentium 4 computer running E-Prime 1.1 (Schneider, Eschman, & Zuccolotto, 2001). Stimuli were displayed on two 17-in. (43.18-cm) monitors: One monitor presented stimuli to the participants. The other monitor allowed the experimenter to observe what letter string was presented without disturbing the participant. We collected vocal responses using a Plantronics (Santa Cruz, CA) LS1 microphone headset and a voice key assembly.

Procedure. Participants were tested individually and were seated approximately 50 cm from the screen. At this distance, three-letter words subtended approximately 1.2° of visual angle and seven-letter words subtended approximately 3.1° of visual angle. Participants were instructed that when a letter string appeared on the screen, their task was to pronounce it as quickly and as accurately as possible. Responses were coded as correct, incorrect, or mistrial (e.g., voice key error) by the experimenter. Each trial consisted of a fixation symbol (+) at the center of the screen for 250 ms followed by a blank screen for 56 ms, after which the word was presented at fixation until a vocal response was detected. A set of 20 practice trials (10 words and 10 nonwords) served to familiarize the participant with the task and allowed the experimenter to adjust the microphone sensitivity to minimize spoiled trials (i.e., trials in which either the microphone failed to respond or it responded prematurely).

Results

RTs and errors were analyzed across participants and items, with both stimulus quality and word frequency as within-subject factors in the subject analysis. In the item analysis, stimulus quality was a within-item factor and word frequency was a between-items factor. To remove individual subject variance, we z-scored the item data prior to the analysis (e.g., see Reynolds & Besner, 2004). The subject data can be seen in Table 1.

Analysis of only the mean RTs is potentially misleading. For example, Yap, Balota, Tse, and Besner (2008)—using a lexical decision task—found opposing interacting effects in a distributional analysis, leading to additivity of two factors in the means. Here, if early processing is thresholded, we would also expect that the joint effects of word frequency and stimulus quality would be additive through much of the distribution. Specifically, the size of the word frequency effect should be the same for bright and dim words across the distribution. We therefore report vincentile plots for the joint effects of stimulus quality and word frequency.

RTs. Trials on which there was a voice key error (1.4%) or an incorrect response (2.9%) were removed prior to RT data analysis. The remaining RTs were submitted to a recursive data trimming procedure in which the criterion for outlier removal was established on the basis of the sample size in that cell (Van Selst & Jolicoeur, 1994), resulting in an additional 1.9% of the data being removed. Mean RTs and errors can be seen in Table 1; item means can be seen online in the supplemental materials. Words presented brightly were read aloud faster than dim ones, $F_1(1, 31) = 66.5$, $MSE = 2,384$, $p < .001$; $F_2(1, 198) = 834.2$, $MSE = 0.032$, $p < .001$. High frequency words were read aloud faster than low frequency words, $F_1(1, 31) = 48.9$, $MSE = 405.4$, $p < .001$; $F_2(1,$

Table 1

Mean Response Times (RTs in Milliseconds), 95% Confidence Intervals (CIs), and Mean Percentage Errors (%Es) in Reading Aloud as a Function of Word Frequency and Stimulus Quality in Experiments 1, 2, and 3

Experiment	Clear			Degraded		
	RT	CI	%E	RT	CI	%E
Experiment 1						
Low frequency	543		1.3	615		2.9
High frequency	520		0.7	589		0.8
Difference	**23**	**±4**	**0.6**	**26**	**±4**	**2.1**
Nonwords	575		5.3	648		6.4
Experiment 2						
Low frequency	509		2.1	568		4.0
High frequency	491		0.8	550		1.9
Difference	**18**	**±4**	**1.3**	**18**	**±4**	**2.1**
Nonwords	551		7.3	612		10.9
Experiment 3						
Words only						
Low frequency	481		1.7	602		5.1
High frequency	470		0.8	576		3.9
Difference	**11**	**±5**	**0.9**	**26**	**±5**	**1.2**
With nonwords						
Low frequency	513		1.8	624		5.3
High frequency	506		1.4	614		3.3
Difference	**7**	**±4**	**0.4**	**10**	**±4**	**2.0**
Nonwords	562		7.7	677		12.9

Note. The 95% CIs for the difference scores were calculated with Masson and Loftus's (2003) within-subjects procedure. The variance for stimulus quality was greater than for word frequency; the CIs therefore were calculated using the mean standard errors for the interaction.

198) = 36.6, $MSE = 0.126$, $p < .001$. Critically, there was no interaction between the effects of stimulus quality and word frequency ($Fs < 1$).

Errors. There was a main effect of stimulus quality, $F_1(1, 31) = 7.1$, $MSE = 3.5$, $p < .05$; $F_2(1, 198) = 7.8$, $MSE = 56.3$, $p < .01$. More errors were made to low frequency words than to high frequency words, $F_1(1, 31) = 15.2$, $MSE = 3.9$, $p < .001$; $F_2(1, 198) = 53.1$, $MSE = 15.6$, $p < .001$. There was a 1.5% interaction between stimulus quality and word frequency in which low stimulus quality affected low frequency words more than high frequency ones, $F_1(1, 31) = 6.2$, $MSE = 2.9$, $p < .05$; $F_2(1, 198) = 5.7$, $MSE = 9.8$, $p < 05$.

Vincentile analysis. A vincentizing procedure was used in which the RT distributions for individual participants were averaged across participants to produce the RT distribution (Vincent, 1912). Ten vincentiles (the mean of observations within a given percentile range) were first computed for each participant. The individual vincentiles were then averaged across participants and the mean vincentiles plotted. The vincentile plots reported here were computed in R (R Development Core Team, 2004).

The mean vincentiles are plotted as a function of word frequency and stimulus quality in Figure 1. The difference scores (high frequency − low frequency) for clear and degraded items are plotted in Figure 2. The frequency effect increased across vincentiles for both clear and degraded items. For present purposes, the key result is that the overlap in the size of the frequency effects for
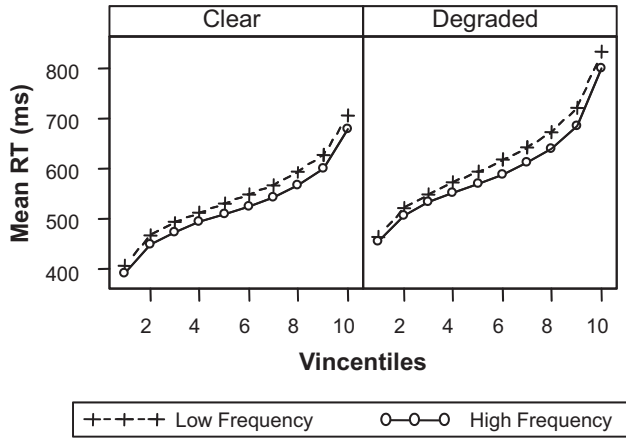
Figure 1. Experiment 1: Vincentile means for participants' reading aloud times as a function of word frequency and stimulus quality. RT = response time.

clear and degraded items is consistent with additivity between stimulus quality and word frequency throughout the distribution.

Discussion

Experiment 1 yielded clear additivity in the RT data and in the vincentiles (but there was a small [1.5%] interaction in the error data). This result in the RTs contrasts with the results reported by both Yap and Balota (2007) and O'Malley et al. (2007), who used the same word set. Both sets of investigators found that stimulus quality and word frequency interacted on RT in the context of reading aloud. The primary difference between those experiments and the present one is the absence of nonwords in the prior experiments, and their presence here. We discuss these results further after reporting Experiments 2 and 3.

Experiment 2

Given that the results of Experiment 1 are novel and surprising to many colleagues, we report a replication. In Experiment 2, we used the DMDX software (Forster & Forster, 2003), which has the advantage of recording the vocal responses. Using this software in conjunction with CheckVocal (Protopapas, 2007) allows one to determine RTs using the waveform and hence serves to reduce measurement error associated with voice key timing (Rastle & Davis, 2002) as well as possible experimenter bias associated with determining errors online. In using the DMDX software, other small adjustments were necessary in the method and procedure to accommodate the new software, specifically (a) the brightness of the stimuli and (b) the timing of presentation of stimuli.

Method

Participants.   A new set of 32 undergraduate students from the University of Waterloo were each paid $4.00 for their participation. All were native English speakers and reported normal or corrected-to-normal vision.

Stimuli.   Experiment 2 used the same items as in Experiment 1. The stimuli were rotated through stimulus quality conditions

across participants, who were assigned to a counterbalancing condition based on order of arrival in the laboratory. Words were again displayed in 16-point Times New Roman font on a black background (writing color 000, 000, 000). In the bright condition, the letter strings appeared in writing color (255, 255, 255); in the dim condition, they appeared in writing color (075, 075, 075). These values differ from Experiment 1 because e-prime and DMDX software have different parameters for RGB settings (the color white is set as 120, 120, 120 in e-prime and 255, 255, 255 in DMDX), making it difficult to set the exact same brightness across experiments. However, this difference adds to the strength of a replication of the findings in the sense of generalizing across more than one brightness level.

Apparatus.   We collected the data using the DMDX software, and we determined RTs and errors using CheckVocal software.

Procedure.   The procedure was identical to that of Experiment 1 except for two small changes: (1) Responses were coded offline as correct, incorrect, or mistrial (e.g., voice key error) by the experimenter using the CheckVocal software; (2) the fixation symbol (+) appeared for 56 ms, followed by a blank screen for 150 ms, after which the stimulus was presented at fixation until a response was detected.

Results

RTs.   Trials on which there was a mistrial (1.1%) or an incorrect response (4.5%) were removed prior to RT analysis. The remaining RTs were submitted to the same recursive data trimming procedure as in Experiment 1, resulting in an additional 1.6% of the data being removed. These data can be seen in the middle part of Table 1. Words presented brightly were read aloud faster than those in the dim condition, $F_1(1, 31) = 275.3$, $MSE = 402.7$,
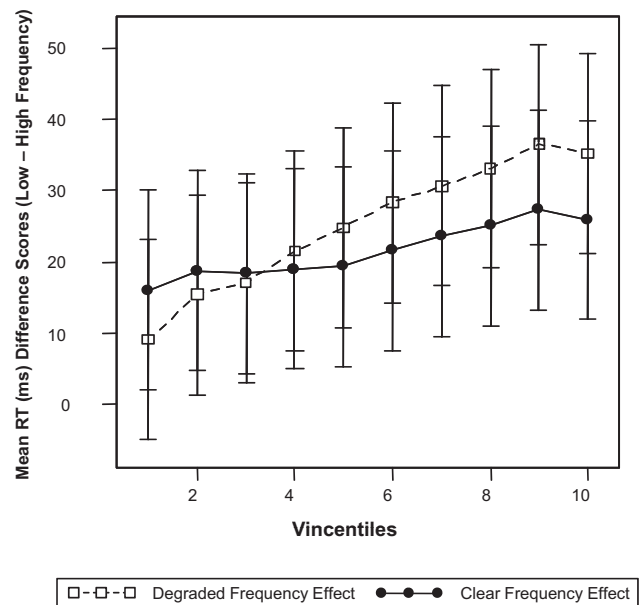

Figure 2. Experiment 1: The difference in the vincentile means for low versus high frequency items for participants' reading aloud times and 95% confidence intervals as a function of stimulus quality. Error bars indicate the 95% confidence intervals. RT = response time.

$p < .001$; $F_2(1, 198) = 1,030$, $MSE = 0.013$, $p < .001$. High frequency words were read aloud faster than low frequency words, $F_1(1, 31) = 78.1$, $MSE = 137.2$, $p < .001$; $F_2(1, 198) = 17.3$, $MSE = 0.226$, $p < .001$. There was no interaction between the effects of stimulus quality and word frequency ($Fs < 1$).

Errors. There was a main effect of stimulus quality, $F_1(1, 31) = 14.4$, $MSE = 4.9$, $p < .01$; $F_2(1, 198) = 18.3$, $MSE = 12.5$ $p < .001$. More errors were made to low frequency words than to high frequency words, $F_1(1, 31) = 22.3$, $MSE = 4.4$, $p < .001$; $F_2(1, 198) = 15.1$, $MSE = 18.9$, $p < .001$. There was no interaction between stimulus quality and word frequency, $F_1(1, 31) = 1.58$, $MSE = 3.9$, $p = .22$; $F_2(1, 198) = 1.4$, $MSE = 12.5$, $p = .22$.

Vincentile analysis. The mean vincentiles were again plotted as a function of word frequency and stimulus quality and appear in Figure 3. The difference scores (low frequency – high frequency) for clear and degraded items are plotted in Figure 4. It is clear that the frequency effect increases across vincentiles for both clear and degraded items. The overlap in the size of the frequency effects for clear and degraded items is consistent with additivity between stimulus quality and word frequency throughout the distribution.

Discussion

Experiment 2 replicated the RT results observed in Experiment 1. When nonwords are randomly intermixed with words, and the task is to read items aloud, stimulus quality and word frequency have additive effects on both mean RT and throughout the distribution. Experiment 2 also produced additive effects of these factors on the errors (thus failing to replicate the small interaction in the error data observed in Experiment 1).

Experiment 3

The results of Experiments 1 and 2 differ from the results reported by O'Malley et al. (2007), in which the task was to read *words* aloud using the same word set and the same kind of stimulus quality manipulation as used here. Nonwords did *not* appear in O'Malley et al.'s experiment, and an interaction between stimulus quality and word frequency was observed (see also Yap & Balota,

Figure 4. Experiment 2: The difference in the vincentile means for low versus high frequency items for participants' reading aloud times and 95% confidence intervals as a function of stimulus quality. Error bars indicate the 95% confidence intervals. RT = response time.

2007). Given the importance of this change in the relation between stimulus quality and word frequency as a function of the presence/absence of nonwords, we sought to further strengthen the case by replicating the previous experiments using a new word set (at the request of the editor) and by including a condition in which only words appear. We expected this experiment to produce a three-way interaction in which an interaction between stimulus quality and word frequency is observed for participants who are only presented with words, whereas participants who are presented with words and nonwords mixed together yield additive effects of stimulus quality and word frequency.

Method

Participants. Fifty-six undergraduate students from the University of Waterloo were each paid $4.00 for their participation. Thirty-two of them read words and nonwords aloud, the other 24 students read only words aloud. All were native English speakers and reported normal or corrected-to-normal vision.

Stimuli. In Experiment 3, we used a new set of 200 words and 200 nonwords, which is available, along with the item means, online (see the supplemental materials). The 100 high frequency words (mean count per million = 666.2) and 100 low frequency words (mean count per million = 16.1) both had a mean of 4.8 letters (range = 3–7). The mean orthographic neighborhood size for the high frequency words was 4.9, and the mean summed bigram frequency was 7093. The mean orthographic neighborhood size was also 4.9 for the low frequency words, and the mean summed bigram frequency was 5954. The nonwords matched the words in length (M = 4.8, range = 3–7), and the mean orthographic neighborhood size was 5.9. We rotated the items through
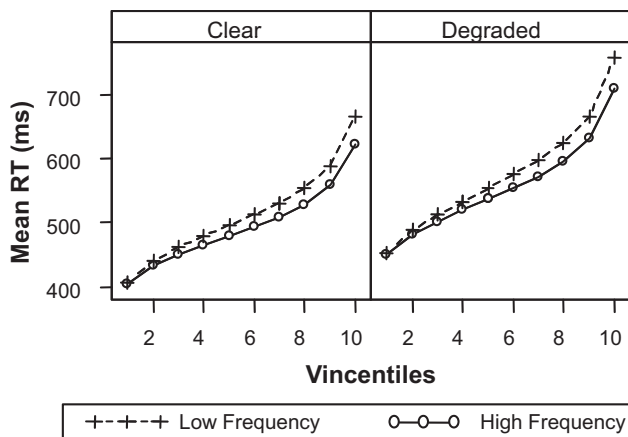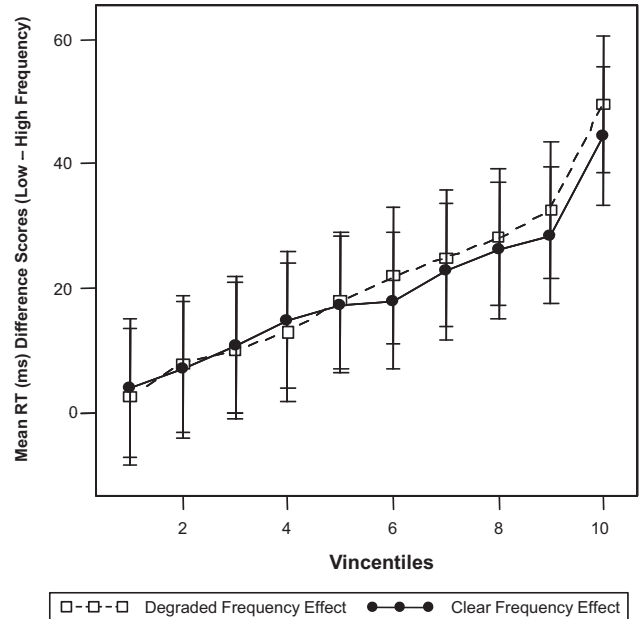

Figure 3. Experiment 2: Vincentile means for participants' reading aloud times as a function of word frequency and stimulus quality. RT = response time.

conditions using a partial Latin square such that each nonword list was presented with each word list equally often across participants in the words and nonwords condition, resulting in eight lists. The stimuli were rotated through stimulus quality conditions across participants in both the words only and the words and nonwords conditions. These were assigned to a counterbalancing condition based on order of arrival in the laboratory, with the nonword condition alternating between participants.

Procedure. The procedure was identical to that of Experiment 2.

Results

As in the previous analyses, RTs and errors were analyzed across participants and items, with both stimulus quality and word frequency as within-subject factors in the subject analysis. In the item analysis, stimulus quality was a within-item factor and word frequency was a between-items factor. The item data were z-scored prior to the analysis to reduce the impact of individual subject variance; the z-scores were calculated by collapsing across all conditions. The subject data can be seen in Table 1. Trials on which there was a voice key error (0.9%) or an incorrect response (4.5%) were removed prior to RT data analysis. The remaining RTs were submitted to the same recursive data trimming procedure used in Experiments 1 and 2, resulting in an additional 1.8% of the data being removed.

Three-way interaction. The critical three-way interaction between stimulus quality, word frequency, and experimental condition (presence/absence of nonwords) was significant in the RT analysis, $F_1(1, 54) = 4.5$, $MSE = 125$, $p = .04$; $F_2(1, 198) = 5.9$, $MSE = 0.038$, $p = .02$, but not in the error analysis, $F_1(1, 54) = 0.9$, $MSE = 6.5$, $p = .34$; $F_2(1, 198) = 1.1$, $MSE = 20.3$, $p = .30$. Further analysis treated the data for the list conditions (presence/absence of nonwords) separately.

Words only. Brightly presented words were read aloud faster than dimly presented ones, $F_1(1, 23) = 138.2$, $MSE = 2,222$, $p < .001$; $F_2(1, 198) = 2,850$, $MSE = 0.054$, $p < .001$. High frequency words were read aloud faster than low frequency words, $F_1(1, 23) = 22.7$, $MSE = 365$, $p < .001$; $F_2(1, 198) = 16.2$, $MSE = 0.213$, $p < .001$. The interaction between the effects of stimulus quality and word frequency was significant, $F_1(1, 23) = 11.3$, $MSE = 123$, $p < .01$; $F_2(1, 198) = 10.1$, $MSE = 0.024$, $p < .01$.

Errors. There was a main effect of stimulus quality, $F_1(1, 23) = 16.3$, $MSE = 15.9$, $p < .01$; $F_2(1, 198) = 424.4$, $MSE = 25.6$, $p < .001$. There was no main effect of frequency, $F_1(1, 23) = 2.9$, $MSE = 8.9$, $p = .10$; $F_2(1, 198) = 3.4$, $MSE = 32.3$, $p = .07$, and there was no interaction between stimulus quality and word frequency ($Fs < 1$).

Words when mixed with nonwords. Brightly presented words were read aloud faster than dimly presented ones, $F_1(1, 31) = 327$, $MSE = 1,178$, $p < .001$; $F_2(1, 198) = 2,698$, $MSE = 0.039$, $p < .001$. High frequency words were read aloud faster than low frequency words, $F_1(1, 31) = 14.1$, $MSE = 168$, $p = .001$; $F_2(1, 198) = 3.2$, $MSE = 0.158$, $p = .07$. There was no interaction between the effects of stimulus quality and word frequency ($Fs < 1$).[4]

Errors. There was a main effect of stimulus quality, $F_1(1, 31) = 30.2$, $MSE = 7.8$, $p < .001$; $F_2(1, 198) = 44.1$, $MSE = 16.8$, $p < .001$. More errors were made to low frequency words than to high frequency words, $F_1(1, 31) = 7.1$, $MSE = 6.7$, $p < .05$; $F_2(1,$

$198) = 5.8$, $MSE = 25.7$, $p < .05$. The interaction between stimulus quality and word frequency was marginal, $F_1(1, 31) = 2.9$, $MSE = 6.7$, $p = .09$; $F_2(1, 198) = 3.6$, $MSE = 16.8$, $p = .06$.

Vincentile analysis. The mean vincentiles are plotted as a function of word frequency and stimulus quality for the condition in which only words appeared (see Figure 5) and for the condition in which both words and nonwords were mixed together (see Figure 6). The difference scores (low frequency – high frequency) for clear and degraded items are plotted in Figures 7 and 8 for the condition in which only words were presented, and when words mixed with nonwords, respectively. It is clear that the frequency effect increases across vincentiles for both clear and degraded items. In Figure 7 there is a clear divergence in the size of the frequency effect such that as reaction times increased, the size of the frequency effect for degraded items increased more than for the clear items. In contrast, inspection of Figure 8 reveals that when words and nonwords were mixed together, the size of the frequency effect was approximately the same throughout the distribution for bright and dim words.

Discussion

Experiment 3, with a new stimulus set, provides a replication of the interaction between word frequency and stimulus quality when reading only words aloud, as in Yap and Balota (2007) and O'Malley et al. (2007), totaling three such demonstrations. It also provides a second replication of the null interaction between these two factors on RT when nonwords are present, totaling three demonstrations.

Combined Analysis

The three experiments reported here yielded no significant interaction between word frequency and stimulus quality in either the mean RTs or the vincentiles when both words and nonwords were mixed in the same block and read aloud. In the first experiment, a small interaction was observed in the error data but not in the following two experiments, although there was a trend toward an interaction in the third experiment. To further explore these findings and increase power, we combined

---

[4] When nonwords were included, the main effect of frequency was only marginal in the item analysis. However standard item analyses are generally associated with low power. To investigate this further, we fit the word data (in the nonword condition) to a linear mixed-effects model as outlined by Baayen (in press), with subjects and items as crossed random effects. There was a main effect stimulus quality, $\hat{\beta} = 115.2$, $t(6601) = 15.5$, $p < .001$, and, critically, a main effect of word frequency, $\hat{\beta} = 8.5$, $t(6601) = 2.02$, $p < .05$. The interaction did not approach significance. We complete the specification of the model by reporting the standard deviation of the random effects. The standard deviation of the random effect of Word was estimated at 26.9. In this model there were two random effects associated with Subjects. First, the standard deviation of the by-subject adjustments was estimated at 61.7. Additionally, subjects were differentially sensitive to stimulus quality (log-likelihood ratio = 37,896, $p < .0001$), the standard deviation for the by-subject adjustments to the quality coefficient was 42.5, and the correlation of the by-subject adjustment to intercept and Family size was .007. The residual standard deviation was 71.9.
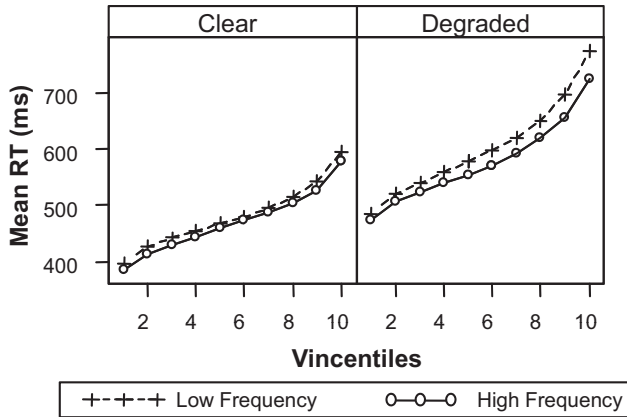
Figure 5. Experiment 3: Vincentile means for participants' reading aloud times as a function of word frequency and stimulus quality when only words are read aloud. RT = response time.

the three experiments in one analysis. These data can be seen in Table 2.

For RTs, there was a main effect of stimulus quality, $F_1(1, 95) = 344$, $MSE = 131.3$, $p < .001$; $F_2(1, 398) = 2,817$, $MSE = 0.05$, $p < .001$, a main effect of word frequency, $F_1(1, 95) = 105$, $MSE = 280$, $p < .001$; $F_2(1, 398) = 26$, $MSE = 0.156$, $p < .001$, but no interaction ($Fs < 1$).

For errors, there were main effects of both stimulus quality, $F_1(1, 95) = 48.7$, $MSE = 5.8$, $p < .001$; $F_2(1, 398) = 65.2$, $MSE = 11.7$, $p < .001$, and word frequency, $F_1(1, 95) = 37.9$, $MSE = 5.0$, $p < .001$; $F_2(1, 398) = 18.9$, $MSE = 20.75$, $p < .001$. There was a significant interaction in the errors, $F_1(1, 95) = 8.7$, $MSE = 4.4$, $p < .01$; $F_2(1, 398) = 7.9$, $MSE = 11.7$, $p < .01$.

The fact that there is a reliable interaction in the errors might be seen as undermining the additivity observed in the RT data, but this is not the only interpretation. Plourde and Besner (1997) suggested that when early processing is thresholded, some participants may not always have completely finished the cleanup op-
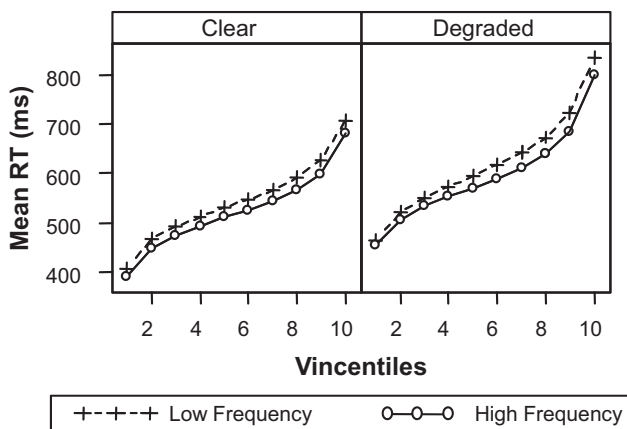

Figure 7. Experiment 3: The difference in the vincentile means for low versus high frequency items for participants' reading aloud times and 95% confidence intervals as a function of stimulus quality when only words are read aloud. Error bars indicate the 95% confidence intervals. RT = response time.

eration (i.e., activation is passed on before the effect of degradation is fully resolved), resulting in a tendency to produce an interaction in the error data. Participants who make many errors are arguably those who unduly emphasize speed and are thus likely to terminate

Figure 6. Experiment 3: Vincentile means for participants' reading aloud times as a function of word frequency and stimulus quality when words and nonwords are read aloud. RT = response time.
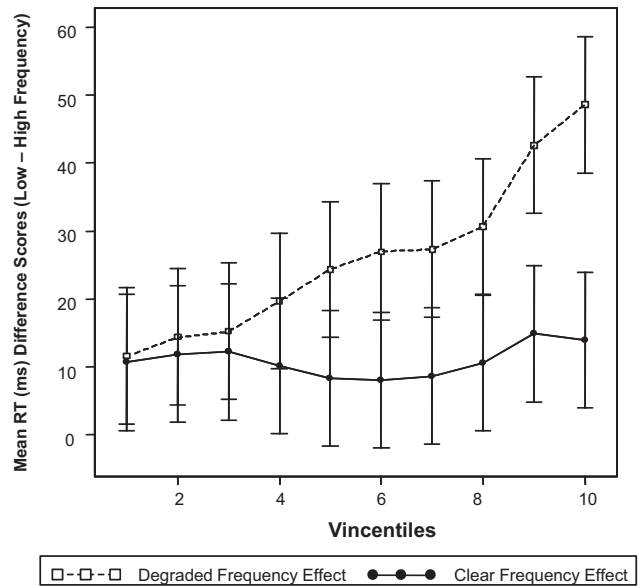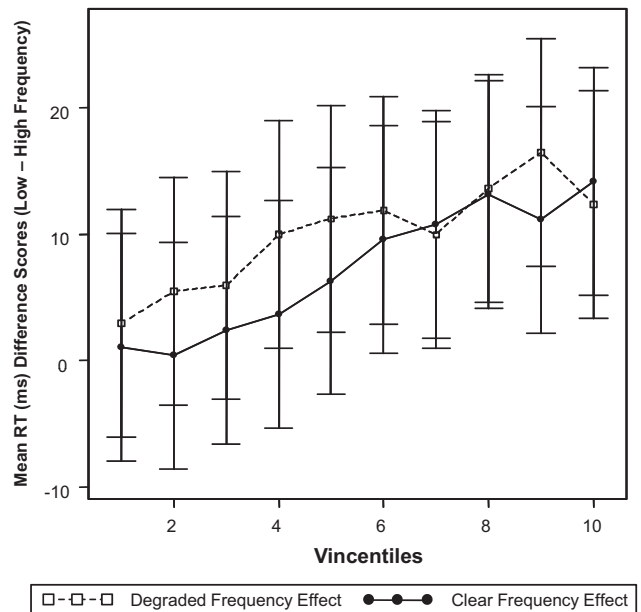

Figure 8. Experiment 3: The difference in the vincentile means for low versus high frequency items for participants' reading aloud times and 95% confidence intervals as a function of stimulus quality when words and nonwords are read aloud. Error bars indicate the 95% confidence intervals. RT = response time.

Table 2

Mean Response Times (RTs in Milliseconds) and Mean Percentage Errors (%Es) When Reading Words Aloud for the Combined Analysis of Experiments 1–3 Along With a Median Split Based on the Average Errors

Group	Clear		Degraded	
	RT	%E	RT	%E
Combined analysis (N = 96)				
Low frequency	522	1.7	604	4.1
High frequency	506	0.9	585	2.0
Difference	**16**	**0.8**	**19**	**2.1**
Nonwords	562	6.8	646	10.1
Few errors group (n = 48)				
Low frequency	523	0.8	599	1.5
High frequency	509	0.4	580	0.9
Difference	**14**	**0.4**	**19**	**0.6**
High errors group (n = 48)				
Low frequency	522	2.7	609	6.7
High frequency	503	1.5	590	3.2
Difference	**19**	**1.2**	**19**	**3.5**

the cleanup process early on some proportion of the trials. To investigate this issue, we did a median split here on the participant data on the basis of the average percentage of errors across conditions.

For participants with few errors (average = 0.87%), there was a main effect of stimulus quality on RTs, $F_1(1, 47) = 138$, $MSE = 1,894$, $p < .001$, and of word frequency, $F_1(1, 47) = 49.8$, $MSE = 253$, $p < .001$, but no interaction, $F_1(1, 47) = 1.6$, $MSE = 122.9$, $p = .21$. In the error analysis, there were main effects of stimulus quality, $F_1(1, 47) = 12.6$, $MSE = 1.4$, $p < .01$, and of word frequency, $F_1(1, 47) = 10.8$, $MSE = 1.0$, $p < .01$, but no interaction, $F_1(1, 47) = 0.26$, $MSE = 2.0$, $p = .61$.

For participants with more errors (average = 3.5%), there was a main effect of stimulus quality on RTs, $F_1(1, 47) = 218.4$, $MSE = 1,662$, $p < .001$, and of word frequency, $F_1(1, 47) = 55.1$, $MSE = 310$, $p < .001$, but no interaction, $F_1(1, 47) = 0.02$, $MSE = 140.8$, $p = .89$. However, there was an interaction in the errors, $F_1(1, 47) = 10.2$, $MSE = 6.3$, $p = .002$. The results of this analysis are therefore consistent with the suggestion that the interaction in the error data is driven by those participants who generate more errors and are likely to have prematurely terminated the cleanup process on some proportion of trials.

General Discussion

The results of the present three experiments can be summarized as follows. Stimulus quality and word frequency have additive effects on reading aloud RTs when words and nonwords are randomly intermixed (Experiments 1, 2, and 3). However, when only words appear in the experiment, stimulus quality and word frequency interact such that high frequency words are less affected by low stimulus quality than are low frequency words (Experiment 3; see also O'Malley et al., 2007; Yap & Balota, 2007).

Three issues merit discussion here. First, *how* do the same factors (stimulus quality and word frequency) produce both additive and interacting effects on RT as a function of the presence/ absence of nonwords in the list? Second, *why* does this happen? And third, *what* general implications, if any, do these results have for our understanding of visual word recognition processes?

The How of Additive Effects of Stimulus Quality and Word Frequency

One way to understand how additive effects of stimulus quality and word frequency arise is in terms of Sternberg's (1969) proposal that additive effects of two factors on RT reflect serially organized processes in which some process only starts after the prior process has finished, and each of two manipulated factors affects a separate process. There is both broad and deep support for this seemingly implausible proposal (see Roberts & Sternberg, 1993; Sternberg, 1998). This account has been suggested before in the context of exactly these factors (albeit in the context of lexical decision). Stimulus quality affects feature and letter processing but not subsequent orthographic lexical activation, whereas word frequency affects lexical activation but not feature and letter level processing. (e.g., Borowsky & Besner, 1993; O'Malley et al., 2007; Plourde & Besner, 1997; Yap & Balota, 2007; Yap et al., 2008).

It is also known that cascaded processing, provided certain constraints are satisfied, can produce additive effects of two factors on mean RT (Ashby, 1982; McClelland, 1979; Roberts & Sternberg, 1993). For example, in a localist model, if stimulus quality affects the feature level and not the letter level, and word frequency affects the lexical level, additivity could be observed in a cascaded model provided the feature and lexical levels are relatively fast, whereas the letter level is relatively slow (see Roberts & Sternberg, 1993). Additivity could presumably also be generated by a PDP model if there is a relatively slow hidden unit level between relatively fast feature and letter levels with the feature level being affected by stimulus quality and the letter level by word frequency.

That said, we are aware of no implemented model of visual word recognition in its current form that produces additivity of stimulus quality and word frequency in reading aloud. Indeed, exploration of this issue by Reynolds and Besner (2004) failed to yield additivity of these factors in the context of the DRC model (even when feedback was eliminated by zeroing out the connections between levels). Simulating additive effects of these factors (and others) in the context of such models may be less easy to accomplish than implied by prior work described by McClelland (1979, 1987) and by Roberts and Sternberg (1993), given that the parameter constraints that need to be satisfied may not be so easily reconciled with the architecture and processing dynamics currently implemented that play a critical role in simulating other phenomena. In particular, and as noted earlier, cascaded processing, at least in the context of the DRC model in which it feeds a serial process in the nonlexical route, leads to an unusual outcome in which a factor that slows processing (letter length) when combined with another factor that also slows processing (stimulus quality) yields an interaction in which the effect of stimulus quality *decreases* as letter length *increases* (Besner & Roberts, 2003). The human data do not yield this pattern.

Interactive activation between various levels is also a central assumption in many computational accounts of visual word recognition, reading aloud, and perceptual identification (e.g., Coltheart et al., 2001; McClelland, 1987; McClelland & Rumelhart, 1981; Perry et al., 2007). We are aware of no demonstration to date

that any IA model can produce all the effects currently considered benchmarks and also produce systematically additive effects of word frequency and stimulus quality. Proponents of such models might therefore take the additivity of stimulus quality and word frequency (along with a number of other examples; see Besner, 2006) as an issue that merits attention. To be absolutely clear, we are not claiming that these models are incapable of producing additivity of factor effects but rather that they do not do so in their current form.[5]

The How of an Interaction Between Stimulus Quality and Word Frequency

The how of the interaction between stimulus quality and word frequency is likely to be uncontroversial given that IA (as in the DRC model) produces an interaction between stimulus quality and word frequency when reading aloud (Reynolds & Besner, 2004). As well, cascaded processing (as in the DRC model when IA is prevented by lesioning feedback) also produces an interaction between stimulus quality and reading aloud (Reynolds & Besner, 2004). We do not expect that producing such an interaction in the context of a PDP model would be difficult either, but that of course remains to be demonstrated.

It is not immediately obvious to us how serially organized processes as in Sternberg's (1969) proposal can produce an interaction between stimulus quality and word frequency, given that these same factors are additive when nonwords are intermixed with the words. One might suppose that feature and letter processing fail to completely clean up the internal representation of the stimulus before passing it for lexical processing when only words appear in the list (relatedly, see Sternberg, 1967, Session 1 vs. Session 2). Or, there might be reasons (unidentified to date) why stimulus quality affects both feature/letter processing *and* lexical processing under these conditions. Until there is some plausible proposal as to why this might be the case, we are inclined to the view that discrete processes are problematic when only words appear in the list.

In summary, it is easy to produce *additive* effects of stimulus quality and word frequency when the processes affected by these factors are serially arranged, discrete, and doubly dissociated in the sense that Factor A affects the first process and not the second, and Factor B affects the second process but not the first. Cascading processes (feed-forward) and cascading processes combined with feed-back face rather more difficulties. This situation is reversed when considering the *interaction* between these same factors of stimulus quality and word frequency when the background context changes such that nonwords are no longer present in the list. Now it is difficult to see how a Sternbergian arrangement of processes can produce the observed outcome, whereas cascaded processing and/or IA produces the observed pattern with ease (at least in the context of DRC; see Reynolds & Besner, 2004).

One resolution to this conundrum, as proposed in the introduction, is that serially arranged and discrete processes are in play when additive effects are observed, whereas cascaded processing and/or IA are in play when the interaction is observed. This proposal is simply not a post hoc account generated to explain these data; rather, it is a hypothesis generated to explain other data, also discussed in the introduction, and it predicted the outcome of the experiments reported here. That said, few psycholinguists are

likely to find the explanation offered here appealing. Be that as it may, the empirical pattern of data appears clear.

The Lexicalization Hypothesis: Looking Forward

The lexicalization account is not without its own issues. First, in the current experiments, the neighborhood density of the nonwords was relatively high level. One implication of the lexicalization account as expressed here is that when nonword N is low (i.e., the nonwords have few neighbors), the probability of lexical capture (pronouncing a nonword as a word) would decrease. This should therefore reduce the probability of participants using a thresholded mode of processing. This line of reasoning leads to the expectation of an interaction between word frequency and stimulus quality when the nonwords mixed with the words are low N.

Second, Blais and Besner (2007) reported a three-way interaction between repetition, lexicality (words vs. nonwords), and stimulus quality when reading aloud (the lag between repetitions was 16 items). Repetition and stimulus quality interacted for the words but had additive effects for the nonwords. Blais and Besner's results are not surprising on their own, but the interaction between repetition and stimulus quality in the nonword context is unexpected given the account suggested here. If the letter level is thresholded when nonwords are present (as the lexicalization hypothesis assumes), then repetition should also have been additive with stimulus quality. It remains to be seen what insights can be gleaned from an experiment that replicates Blais and Besner's experiment and that also manipulates word frequency.

An Alternative Account

Can CDP+ simulate these data by emphasizing the nonlexical route, which Perry et al. (2007) claimed is functionally thresholded? Johannes C. Ziegler (personal communication, March 16, 2008) suggested the following:

> CDP+ can produce both an interaction between stimulus quality and frequency as well as an additive effect of these two variables. Whether one or the other is obtained does seem to depend on the strength of the nonlexical route. If these arguments are correct, it should be the case that the size of the frequency effect is reduced in the mixed list compared with the pure list. The size of the frequency effect would therefore provide an important marker for the strategic shift from lexical to nonlexical processing. This information would provide a crucial constraint for further simulations of these effects.

We are unable to follow exactly how Ziegler's account could work, but then we are not modelers. Our comprehension failure here may only mean that there are subtleties associated with this computational model that we do not yet appreciate. A demonstration that CDP+ can simulate the present results would therefore be important, but a critical aspect of the computational modeling enterprise surely involves understanding *why* the models behave the way they do.

Our second point is that Ziegler is very clear that the size of the word frequency effect across the presence/absence of nonwords must differ according to his route shift emphasis account. We note,

---

[5] We note with interest that the newest computational version of DRC (Version 1.1.4, which is available online at http://www.maccs.mq.edu.au/~ssaunder/DRC/) includes an option for thresholding various modules.

however, that the data do not support this claim, given that in Experiment 3 the magnitude of the word frequency effect in the *bright* condition was the same size when nonwords were present as when they were absent ($F_1 < 1$).

The Why Question

Why do the processing dynamics appear to vary so dramatically across the present contexts? The proposal advanced here is that cascaded processing increases the probability of lexicalizing the pronunciation of a nonword when stimulus quality is low, something that participants should wish to avoid given the typical emphasis on accuracy in these kinds of experiments. Hence, they engage in discrete processing at an early level throughout the experiment, leading to additive effects of stimulus quality and word frequency in the presence of nonwords. When only words are present in the reading aloud task, then cascaded processing is adopted because in this context lexicalization is not a potential problem.

The What Question

What general implications, if any, do the results discussed here have for understanding visual word recognition? One major account of visual word recognition is that many of the subprocesses are "automatic" in one way or another (e.g., T. L. Brown, Gore, & Carr, 2002; see also the long list of investigators noted in Reynolds & Besner, 2006). A strong view of such automaticity is that it is context independent. This claim is problematic given the present results (and many others). The typical response to this point is to claim that automaticity is context dependent. However, assuming that processing is automatic but context *dependent* in an unspecified way is unappealing to us because it is too theoretically vacuous and potentially circular at present. A more profitable direction is to look for additional examples in which the joint effects of a pair of factors change depending on the level of a third factor (see M. Brown, Stolz & Besner, 2006; Ferguson, Robidoux, & Besner, in press; Stolz & Neely, 1995). At the very least, this will serve to broaden the empirical base that will need a theoretical perspective. More generally, we currently lack any broad theory of context effects that will help guide such a search.

Conclusions

The lexicalization account proposed here should be viewed as tentative; it clearly needs to be explored further. Whatever one's theoretical predilections, the central implication of the results discussed here is that the processing underlying aspects of visual word recognition are rather more dynamic than widely assumed. This conclusion is neither particularly welcome nor especially appealing to the extent that it makes theorizing about mental performance more difficult. Nonetheless, it reflects a direction (e.g., see also Yap et al., 2008) that the field at large will need to take into account when attempting to explain skilled "reading" in particular and mental performance more generally.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17,* 814–823.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 234–254.

Ashby, G. F. (1982). Deriving exact predictions from the cascade model. *Psychological Review, 89,* 599–607.

Baayen, R. H. (in press). *Analyzing linguistic data: A practical introduction to statistics.* Cambridge, MA: Cambridge University Press.

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,* 289–302.

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.

Becker, C. A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 556–566.

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,* 492–538.

Besner, D. (2006). *Visual language processing and additive effects of multiple factors on timed performance: A challenge for the interactive activation framework?* Available from PsyCrit Web site, http://psycrit.com/

Besner, D., & Roberts, M. A. (2003). Reading nonwords aloud: Results requiring change in the dual route cascaded model. *Psychonomic Bulletin & Review, 10,* 398–404.

Besner, D., Wartak, S., & Robidoux, S. (2008). Constraints on computational models of basic processes in reading. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 242–250.

Blais, C., & Besner, D. (2007). Reading aloud: When the effect of stimulus quality distinguishes between cascaded and thresholded components. *Experimental Psychology, 54,* 215–224.

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

Borowsky, R., & Besner, D. (2006). Parallel distributed processing and lexical-semantic effects in visual word recognition: Are a few stages necessary? *Psychological Review, 113,* 181–195.

Brown, M., Stolz, J. A., & Besner, D. (2006). Dissociative effects of stimulus quality on semantic and morphological contexts in visual word recognition. *Canadian Journal of Experimental Psychology, 60,* 190–199.

Brown, T. L., Gore, C. L., & Carr, T. H. (2002). Visual attention and word recognition in Stroop color naming: Is word recognition automatic? *Journal of Experimental Psychology: General, 131,* 220–240.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

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

Ferguson, R., Robidoux, S., & Besner, D. (in press). Reading aloud: Evidence for contextual control over lexical activation. *Journal of Experimental Psychology: Human Perception and Performance.*

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12,* 627–635.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers, 35,* 116–124.

Forster, K. I., & Hector, J. (2002). Cascaded versus noncascaded models of lexical and semantic processing: The turple effect. *Memory & Cognition, 30,* 1106–1117.

Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches

to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 361–379.

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

Masson, M., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.

McCann, R. S., & Besner, D. (1987). Reading pseudohomophones: Implications for models of pronunciation assembly and the locus of word-frequency effects in naming. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 14–24.

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. (1987). The case for interactionism in language processing. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 3–36). Hillsdale, NJ: Erlbaum.

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.

Monsell, S., Patterson, K. E., Graham, A., Hughes, C. H., & Milroy, R. (1992). Lexical and sub-lexical translation of spelling to sound: Strategic anticipation of lexical status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 452–467.

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

Mulatti, C., Reynolds, M. G., & Besner, D. (2006). Neighborhood effect in reading aloud: New findings and new challenges for computational models. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 799–810.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111,* 721–756.

Norris, D. (1984). The effects of frequency, repetition and stimulus quality in visual word recognition. *Quarterly Journal of Experimental Psychology, 36*A, 507–518.

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

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

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., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107,* 786–823.

Plaut, D. C., & Booth, J. R. (2006). More modeling but still no stages: Reply to Borowsky and Besner. *Psychological Review, 113,* 196–200.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (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.

Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behaviour Research Methods, 39,* 859–862.

Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 307–314.

R Development Core Team. (2004). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Reynolds, M. G., & 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.

Reynolds, M. G., & Besner, D. (2005). Basic processes in reading: A critical review of pseudohomophone effects in reading aloud and a new computational account. *Psychonomic Bulletin & Review, 12,* 622–646.

Reynolds, M. G., & Besner, D. (2006). Reading aloud is not automatic: Processing capacity is required to generate a phonological code from print. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 1303–1323.

Reynolds, M. G., & Besner, D. (2008). Contextual effects on reading aloud: Evidence for pathway control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 50–64.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer & 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, PA: Psychology Software Tools.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

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. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30,* 267–315.

Sternberg, S. (1998). Discovering mental processing stages: The method of additive factors. In D. Scarborough & S. Sternberg (Eds.), *Methods, models, and conceptual issues: An invitation to cognitive science* (pp. 703–863). Cambridge, MA: MIT Press.

Stolz, J. A., & Neely, J. H. (1995). When target degradation does and does not enhance semantic context effects in word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 596–611.

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *Quarterly Journal of Experimental Psychology, 47*A, 631–650.

Vincent, S. B. (1912). The function of vibrissae in the behavior of the white rat. *Behavioral Monographs, 1*(Whole No. 5).

Wilding, J. M. (1988). The interaction of word frequency and stimulus quality in the lexical decision task: Now you see it, now you don't. *Quarterly Journal of Experimental Psychology, 40*A, 757–770.

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,* 274–296.

Yap, M. J., Balota, D. A., Tse, C., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency on lexical decision: Evidence for opposing interactive influences revealed by RT distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 495–513.