Journal of Experimental Psychology:
Learning, Memory, and Cognition
2010, Vol. 36, No. 3, 750–764

© 2010 American Psychological Association
0278-7393/10/$12.00 DOI: 10.1037/a0019178

On the Joint Effects of Stimulus Quality, Regularity, and Lexicality When Reading Aloud: New Challenges

Derek Besner, Shannon O'Malley, and Serje Robidoux
University of Waterloo

A number of computational models have been developed over the last 2 decades that are remarkably successful at explaining the process of translating print into sound. Nevertheless, 2 of the most successful computational accounts on the table fail to simulate the results from factorial experiments reported in this article in which university students read aloud letter strings that varied in terms of spelling–sound regularity and lexicality (regular words vs. exception words vs. nonwords) and stimulus quality (bright vs. dim). Skilled readers yielded additive effects of regularity and stimulus quality and additive effects of lexicality and stimulus quality on both RT and errors when nonwords were mixed with words. When only words appeared in the list, there was an interaction in which exception words were less affected by low stimulus quality than regular words were; no existing account anticipates or explains these results. We advance a hypothesis that assumes a novel module that accommodates these data and provide an existence proof in the form of a simulation.

Keywords: visual word recognition, spelling–sound translation, effects of stimulus quality on spelling–sound translation, joint effects of stimulus quality and lexicality on spelling–sound translation

Skilled readers know thousands of words, and they can read many of them aloud quickly and accurately. However, there is a class of words in English (and a number of other orthographies) that cause difficulty for even skilled readers. These are words like *pint* that are exceptions to the typical rules of pronunciation (*_int* is typically pronounced as in *mint/lint/hint*). Reading aloud of such exception words is slower and often less accurate when compared with regular words like *mint*. Attempts to explain this effect (among others) have given rise to a large number of computational accounts of visual word recognition over the last two decades (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Harm & Seidenberg, 1999; Perry, Ziegler, & Zorzi, 2007; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). Arguably, the two most successful accounts are Coltheart et al.'s (2001) dual route cascaded (DRC) model and Perry et al.'s (2007) connectionist dual process (CDP+) model.[1] Both of these models have a dual route architecture consisting of a lexical route and a nonlexical route. Parallel feature analysis across the array activates parallel letter level analysis that in turn activates both lexical and nonlexical routines. The output of both of these routines converges on the phonemic buffer, which is where the final pronunciation is produced.

The Lexical Route

Both the DRC and CDP+ use essentially identical lexical routes, consisting of an orthographic input lexicon with localist representations (lexical entries) for the spellings of each of the monosyllabic words in English and a phonological output lexicon with localist representations of the phonology for each of these words. Presentation of a word yields activation that cascades from features to letters through to the orthographic input lexicon and then the phonological output lexicon. There is also interactive activation between the letter level and the orthographic input lexicon, between the orthographic input lexicon and the phonological output lexicon, and between the phonemic buffer and the phonological output lexicon. This route correctly pronounces all words that have representations in both of these lexicons, but it is unable to pronounce letter strings aloud that do not have representations in both of these lexicons (e.g., *frane* and *frilp*).

The Nonlexical Route

Following feature and letter activation, the nonlexical route converts spelling to sound sublexically by converting orthographic units into phonological units serially, from left to right (in DRC; in CDP+, letters enter the graphemic buffer serially and left to right). This route correctly reads aloud virtually all letter strings that

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

This research was supported by Grant A0998 from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Derek Besner and NSERC doctoral fellowships to Shannon O'Malley and Serje Robidoux. We thank Claudio Mulatti for the total letter confusability values and Dave Balota, Melvin Yap, Marco Zorzi, and particularly Max Coltheart for comments and discussion (but this does not imply their agreement with any of the ideas expressed here).

Correspondence concerning this article should be addressed to Derek Besner, Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3GI. E-mail: dbesner@uwaterloo.ca

[1] We have not considered any of the parallel distributed processing class of models of reading aloud here because we have been unable to obtain any of these models from their authors.

could be words in terms of their orthography but happen not to be (e.g., *frane* and *frilp*) and all words that are regular in terms of their spelling–sound correspondences (e.g., *gave/save/rave/wave* and *lint/mint/hint/dint*). However, it assigns the regular pronunciation to strings like *pint* and *have*, rather than reading them aloud correctly.

The key difference between the models lies in the details of the nonlexical route's operation. Most significantly, the DRC uses a set of specified rules to convert print into phonology, whereas the CDP+ uses a parallel distributed processing network trained to encode the statistical properties of the language. Though these differences are significant with respect to the ability to simulate some phenomena (e.g., consistency), they are not particularly germane to the present study of these models, so we do not discuss them further.

Both routines are always activated by print. Thus, when the intact models read aloud, the lexical route drives the correct pronunciation for an exception word, but this output is slowed in the phonemic buffer (particularly in the case of low-frequency words) because of competition between the different pronunciations produced by lexical and nonlexical routes.

On the Usefulness of Multifactor Experiments

When only a single factor is manipulated in an experiment, the results can often be explained in a variety of ways (e.g., the effect of word frequency in lexical decision and reading aloud has a wide range of explanations; Becker, 1976; Besner & Smith, 1992; Borowsky & Besner, 1993; Forster & Chambers, 1973; Morton, 1969; Murray & Forster, 2004; Norris, 2006). In short, it is often too easy to generate explanations for a main effect, and hence there are many competing explanations. One approach to discriminating between competing accounts is to jointly manipulate the psycholinguistic factor of interest and a second factor (here, stimulus quality). Adding complexity to an experiment is usually something to be avoided, but it is strategic in the current context because it produces more complex data patterns that in turn can help falsify some of the various accounts (e.g., in the case of word frequency, see Besner & O'Malley, 2009; O'Malley & Besner, 2008; Yap, Balota, Tse, & Besner, 2008). Stimulus quality is a useful second manipulation because it can be simulated in most computational models, typically by modifying the connection weights at one or more levels. The models' behavior can then be compared with how humans behave.

Background to the Present Experiments

Previous work with this multifactor approach has examined the joint effects of stimulus quality when factorially combined with (a) letter length when reading nonwords aloud (Besner & Roberts, 2003), (b) lexical density when reading nonwords aloud (Reynolds & Besner, 2004), and (c) word frequency when reading aloud (O'Malley & Besner, 2008; O'Malley, Reynolds, & Besner, 2007; Yap & Balota, 2007). Table 1 provides a summary of the results of experiments reported to date on the presence or absence of an interaction between various factors and stimulus quality as a function of list type. Of particular relevance to the present investigation, O'Malley and Besner (2008) found that low-frequency words were slowed more by a reduction in stimulus quality than were high-frequency words, provided that only words appeared in the list. In contrast, the joint effects of these same factors were additive on the time to read aloud when words and nonwords were mixed in a single block. O'Malley and Besner therefore argued that when nonwords are present, the effects of stimulus quality are constrained to the feature and letter level (achieved by thresholding the letter level) to reduce the probability of lexicalizations in response to nonwords, particularly when they appear in degraded form.

Two New Predictions

One new prediction that follows from the O'Malley and Besner (2008) account (see also Besner & Roberts, 2003; Reynolds &

Table 1
Results of Experiments on the Joint Effects of Various Lexical (and Nonlexical) Factors and Stimulus Quality When Reading Aloud, as a Function of List Type

	Stimulus quality		
Lexical or nonlexical factor	Pure list (words only)	Pure list (nonwords only)	Mixed list (words and nonwords)
1. Word frequency	Interaction[a] (Yap & Balota, 2008; O'Malley & Besner, 2008)	—	No interaction (O'Malley & Besner, 2008)
2. Neighborhood density		No interaction (Reynolds & Besner, 2004)	
3. Letter length		No interaction (Besner & Roberts, 2003)	
4. Repetition	Interaction[b] (Ferguson et al., 2009)		Interaction/no interaction[c] (Blais & Besner, 2007)
5. Regularity	Interaction[d] (present article)/no interaction (Herdman et al., 1999)	—	No interaction (present article)
6. Lexicality	—	—	No interaction (present article)
7. Semantic priming	Interaction[e] (Ferguson et al., 2009)	—	

Note. Empty cells indicate that no experiments exist. Dashes indicate that it is not possible to do this comparison.
[a] The effect of stimulus quality is greater for low frequency words than for high frequency words. [b] The effect of stimulus quality is greater for nonrepeated words than for repeated words. [c] The effect of stimulus quality is greater for nonrepeated words than for repeated words. There is no interaction for nonwords. [d] The effect of stimulus quality is greater for regular words than for exception words. [e] The effect of stimulus quality is greater for unrelated target words than for related target words when relatedness proportion is .5.

Besner, 2004) is that the joint effects of regularity and stimulus quality on reaction time (RT) should also be additive when words and nonwords are randomly intermixed. This follows because in their account the effect of stimulus quality does not extend beyond the letter level (or perhaps the feature level) when nonwords are mixed with words. Given that regularity does not affect processing until well after the letter level, we expect that the joint effects of these manipulations ought not to interact. This assumption regarding the locus of the regularity effect is explicit in the implementations of the two models discussed here.

A second new prediction is that lexicality (words vs. nonwords) should also be additive with stimulus quality on RT. This again follows from the assumption that the effect of stimulus quality does not extend beyond the feature or letter level when words are read in the context of being mixed with nonwords.

Stimulus Quality × Regularity When Only Words Appear in the List

It is also of interest to examine the joint effects of stimulus quality and regularity on reading aloud when only words appear in the list. If the effect of stimulus quality extends sufficiently deep into the processing system, it might also produce an interaction with regularity. On the other hand, there is some evidence that the effect of stimulus quality does not extend beyond the orthographic input lexicon (Ferguson, Robidoux, & Besner, 2009). Herdman, Chernecki, and Norris (1999) reported that stimulus quality and regularity had additive effects on the time to read aloud when only words appeared in the experiment. However, the trend in this experiment was unusual; low-frequency exception words were less affected by low stimulus quality than were regular words (51 ms vs. 38 ms; see Table 4 in Herdman et al., 1999). Further, in the same experiment, these authors also reported additive effects of stimulus quality and word frequency on the time to read aloud. Given that other experiments have yielded an interaction between stimulus quality and word frequency under this condition, and given that the Herdman et al. experiment had few items per cell (20), it may well be that their failure to detect both interactions (of opposing signs) reflects Type II errors. In short, their report of additive effects of stimulus quality and regularity on RT is on weak footing and bears closer examination. Whatever the outcome, it will serve to constrain theoretical accounts.

We conducted three experiments with skilled readers. In all experiments a single letter string appeared on the screen, and the subject read it aloud. Experiment 1 manipulated stimulus quality, regularity, and lexicality. Experiment 2 repeated Experiment 1, except that nonwords did not appear in the experiment. Experiment 3 replicated the novel pattern observed in Experiment 2. All conditions were randomized in a single block of trials in all experiments. We then carried out simulations with both the DRC[2] and CDP+ models to determine whether either of these models could simulate the human data.

To anticipate the main results: Skilled readers yielded additive effects of regularity, lexicality, and stimulus quality when nonwords were intermixed with the words but yielded an interaction between regularity and stimulus quality when only words appeared in the experiment. This interaction took the form of a smaller regularity effect when the stimuli were dim compared with when they were bright. Neither of the models simulated the joint effects of the factors explored here under any of the experimental conditions, despite the fact that both models yielded robust main effects of regularity, lexicality, and stimulus quality. The general discussion takes up the issue of how these results can be understood and how the models can be modified to correctly simulate them.

Experiment 1

Method

Participants. Twenty-four undergraduate students from the University of Waterloo participated for course credit. All were native English speakers and reported normal or corrected-to-normal vision.

Stimuli. One hundred regular words and 100 exception words were matched for frequency, neighborhood density, and whammies (Rastle & Coltheart, 1998). The words were run through both DRC and CDP+ with the nonlexical route turned off to verify that the two sets of words were matched for lexical characteristics. Three of the words either were not in DRC's lexicon (*peon*) or were outliers (*mould* and *moult* yielded responses greater than three standard deviations from the mean) and were thus removed from all analyses, leaving 197 experimental stimuli. DRC took an average of 77.7 cycles to read the regular words and 77.9 cycles to read the exception words ($t < 1$). For CDP+, *peon*, *mould*, *moult*, and *gauge* were all absent from the lexicon. CDP+ took an average of 126.4 cycles to read regular words and 128.1 cycles to read exception words ($t < 1$). The average ratings on other important lexical factors for each stimulus type can be seen in Table 2.

There was also a set of 200 nonwords matched to the words for letter length. Seminal work on letter confusability by D. Fiset, Arguin, Bub, Humphreys, and Riddoch (2005) and S. Fiset, Arguin, and Fiset (2006) suggests that this measure will be important in future studies of normal readers. We therefore also report these values for our stimulus set: Average total letter confusability was 467.5 for the regular words, 476.5 for the exception words, and 493.0 for the nonwords. The regular and exception words did not differ on this measure ($t < 1$), whereas the nonwords differed significantly from the regular words ($t = 2.5$, $df = 298$, $p < .001$), but not from the exception words ($t = 1.5$, $df = 298$, $p = .13$).

The stimuli were rotated through stimulus quality conditions across participants, who were assigned to a counterbalancing con-

[2] A reviewer of the first submission of the present article questioned why we would do simulations with DRC, given his/her claim that "DRC will never produce additive effects—at least in its current form—because all processing is cascaded." There are a number of misconceptions here: (a) DRC uses different types of activation to connect various levels, not just cascaded processing—a number of levels are engaged in interactive activation, and at least one operation (part of the nonlexical route) is serial and thresholded; (b) McClelland (1979), Ashby (1982), and Roberts and Sternberg (1993) have demonstrated that purely cascaded models can produce additive effects, at least on mean RT, provided certain boundary conditions are respected; and (c) Reynolds and Besner (2004) reported one simulation with DRC in which additive effects of lexical density and stimulus quality are seen (although DRC does not produce systematic additivity in our experience).

Table 2
Average Ratings on Various Lexical Factors for the Regular Words, Exception Words, and Nonwords

Lexical factor	Regular words	Exception words	Nonwords
Number of letters	4.6	4.6	4.6
Frequency	5.2	5.2	
Neighborhood density	3.1	3.2	7.8
Letter confusability	467.5	476.5	493.0
Cycles to criterion (DRC model intact)	77.8	92.6	156.8
Cycles to criterion (DRC nonlexical route lesioned)	77.7	77.9	
Cycles to criterion (CDP+ model intact)	102.7	117.2	155.0
Cycles to criterion (CDP+ nonlexical route lesioned)	126.4	128.1	

Note. Word frequency = mean count per million (from the CELEX database; Baayen et al., 1993); letter confusability = average total letter confusability.

dition on the basis of order of arrival in the laboratory. Words were displayed in lowercase 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. Stimulus type and brightness level were intermixed in a different random order for each participant.

Apparatus. The data were collected on a Pentium 4 computer using DMDX software (Forster & Forster, 2003). RTs and errors were determined using CheckVocal software (Protopapas, 2007).

Procedure. Participants were tested individually and were seated approximately 50 cm from the screen. 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. Each trial consisted of a fixation symbol (+) at the center of the screen for 56 ms, followed by a blank screen for 150 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 procedure. Responses were coded offline as correct, incorrect, or mistrial (e.g., the participant coughed or some other error in the recording occurred) using the CheckVocal software.

Results

RTs. The data for one of the participants was lost during data transfer, and therefore only 23 participants were analyzed. Trials on which there was a mistrial (1.6%) or an incorrect response (9.1%) were removed prior to RT analysis. The remaining RTs for the words were submitted to a recursive data trimming procedure (Van Selst & Jolicoeur, 1994), resulting in the removal of an additional 5.1% of the data. Mean RTs and mean percentage errors from the subject analysis can be seen in Table 3 for all conditions. Following Reynolds and Besner (2004; see also O'Malley & Besner, 2008), the item analysis is based on the z-scored RTs to minimize the contribution from between-subject variance. (Item means and z scores are available at http://artsweb.uwaterloo.ca/~dbesner2/publications.html)

Regularity × Stimulus Quality.

RTs. Words presented brightly were read aloud faster than those in the dim condition, $F_1(1, 22) = 148.7$, $MSE = 495.7$, $p < .001$, $F_2(1, 195) = 601.5$, $MSE = 0.07$, $p < .001$. Regular words were read aloud faster than exception words, $F_1(1, 22) = 79.1$, $MSE = 294.5$, $p < .001$, $F_2(1, 195) = 35.8$, $MSE = 0.49$, $p < .001$. Most centrally, there was no interaction between stimulus quality and regularity, $F_1 < 1$, $F_2(1, 195) = 1.6$, $MSE = 0.07$, $p = .2$. The 95% confidence interval around the interaction (calculated using the Masson & Loftus, 2003, within-subjects method) is ±6 ms.

Errors. There was no main effect of stimulus quality, $F_1(1, 22) = 0.3$, $MSE = 20.9$, $p > .05$, $F_2(1, 195) = 0.5$, $MSE = 72.8$, $p > .05$. More errors were made to exception words than to regular words, $F_1(1, 22) = 50.3$, $MSE = 33.4$, $p < .001$, $F_2(1, 195) = 27.1$, $MSE = 254.3$, $p < .001$. There was no interaction between stimulus quality and regularity (Fs < 1).

Lexicality × Stimulus Quality. Here we report a pair of 2 × 2 analyses of variance (ANOVAs) to facilitate later comparison with the computational models. One ANOVA compared the effect of levels of stimulus quality on regular words versus nonwords, and the other compared the effect of levels of stimulus quality on exception words versus nonwords.

RTs (regular words vs. nonwords). There was a main effect of lexicality, $F_1(1, 22) = 62.6$, $MSE = 419.8$, $p < .001$, $F_2(1, 298) = 50.2$, $MSE = 0.46$, $p < .001$; a main effect of stimulus quality, $F_1(1, 22) = 118.5$, $MSE = 579.7$, $p < .001$, $F_2(1, 298) = 736.8$, $MSE = 0.08$, $p < .001$; and no interaction (Fs < 1).

Errors (regular words vs. nonwords). There was a main effect of lexicality, $F_1(1, 22) = 26.4$, $MSE = 39.3$, $p < .001$, $F_2(1, 298) = 42.7$, $MSE = 140.4$, $p < .001$; no main effect of stimulus quality (Fs < 1), and no interaction, (Fs < 1).

RTs (exception words vs. nonwords). There was no main effect of lexicality (Fs < 1). There was a main effect of stimulus

Table 3
Mean RTs (ms) and Mean Percentage Errors (%E) in Experiments 1, 2 and 3 When Reading Aloud as a Function of Regularity, Lexicality, and Stimulus Quality

	Bright		Dim	
Stimulus type	RT	%E	RT	%E
Experiment 1				
Exception	517	13.8	574	13.0
Regular	486	5.0	542	4.7
Difference	**31**	**8.8**	**32**	**8.3**
Nonwords	521	11.1	574	12.0
Experiment 2				
Exception	514	10.7	614	12.8
Regular	473	1.3	588	4.6
Difference	**41**	**9.4**	**26**	**8.2**
Experiment 3				
Exception	570	12.2	646	14.3
Regular	520	2.4	609	4.7
Difference	**50**	**9.8**	**37**	**9.6**

Note. RT = reaction time.

quality, $F_1(1, 22) = 105.6$, $MSE = 667.3$, $p < .001$, $F_2(1, 295) = 545.7$, $MSE = 0.10$, $p < .001$, but no interaction (Fs < 1).

Errors (exception words vs. nonwords). There was a main effect of lexicality in the subject analysis, $F_1(1, 22) = 8.0$, $MSE = 9.6$, $p = .01$, but not in the item analysis ($F_2 < 1.5$); no main effect of stimulus quality (Fs < 1); and no interaction (Fs < 1.4).

Vincentiles. Given that Experiment 1 produced additive effects of stimulus quality and regularity on mean RT, we also investigated whether this additivity was true of the RT distribution in general or whether the distribution reveals an interaction not observed in the mean RT (see Yap et al., 2008, for an example in which such an analysis reveals a four-way interaction between stimulus quality, word frequency, foil type, and the RT distribution). 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) and are plotted as a function of word type and stimulus quality in Figure 1. The difference scores for the words only (exception words − regular words) for clear and degraded items are plotted in Figure 2. The regularity effect increased across vincentiles for both clear and degraded items, consistent with the additivity observed in the means.

Discussion

The results from Experiment 1 show that both regularity and lexicality have additive effects with stimulus quality on both mean RT and errors and that this additivity is observed throughout the RT distribution (except for the slowest vincentile; this may represent more reprocessing because of uncertainty, particularly in the case of dim stimuli). These observations are consistent with the hypothesis that the effect of stimulus quality is restricted to a process common to words and nonwords, at least when words and nonwords are intermixed. That is, stimulus quality at most affects feature and letter level processing but does not extend further (O'Malley & Besner, 2008; see also Besner & Roberts, 2003; Reynolds & Besner, 2004). We take this issue up further in the General Discussion.

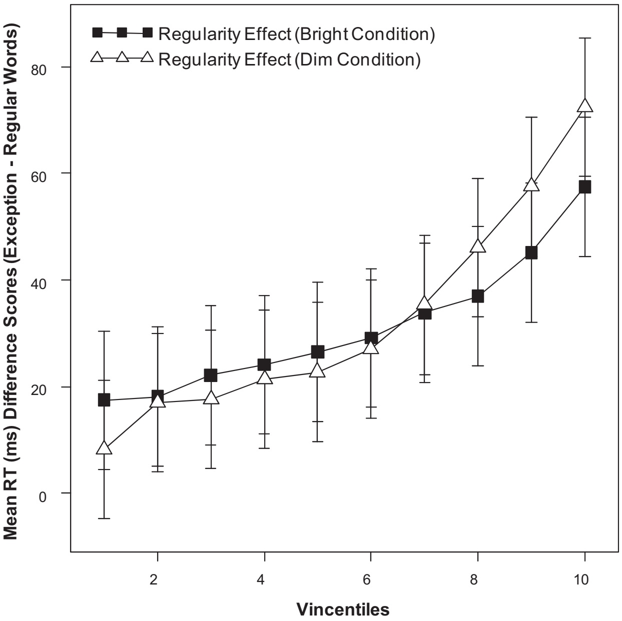

Figure 2. Experiment 1: The difference in the vincentile means for exception versus regular words for participants' reading aloud times. The bars represent 95% confidence intervals. RT = reaction time.

Experiment 2

Experiment 2 investigated the joint effects of stimulus quality and regularity when only words appear in the experiment. The empirical issue here is whether additive effects of these two factors are observed or whether there is an interaction in which exception

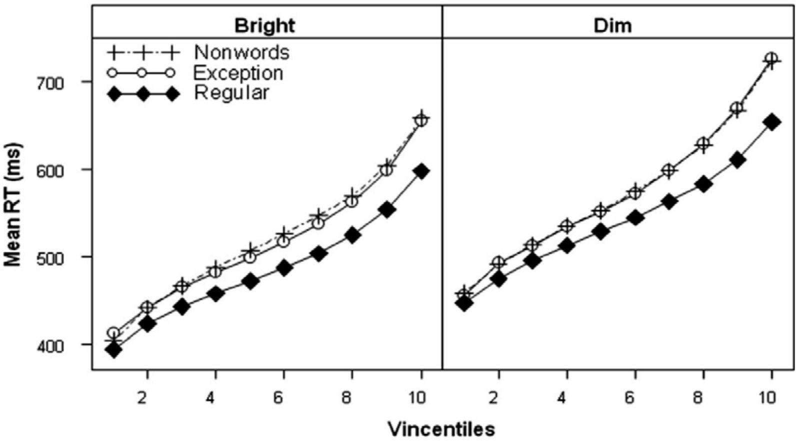

Figure 1. Experiment 1: Vincentile means for participants' reading aloud times as a function of stimulus type and stimulus quality. RT = reaction time.

words are less affected by low stimulus quality than are regular words (as is the trend in the data from Herdman et al., 1999).

Method

Participants. Twenty-four undergraduate students from the University of Waterloo participated for \$5 or for course credit. All were native English speakers and reported normal or corrected-to-normal vision.

Stimuli. The same 100 regular words and 97 exception words from Experiment 1 were used in Experiment 2. Nonwords did not appear in the experiment. 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. Words were displayed in lowercase 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. All conditions were randomly intermixed in a single block of trials.

Apparatus. The data were collected on a Pentium 4 computer using DMDX software (Forster & Forster, 2003). RTs and errors were determined using CheckVocal software (Protopapas, 2007).

Procedure. The same procedure as in Experiment 1 was used. Responses were coded offline as correct, incorrect, or no response using the CheckVocal software.

Results

Trials on which there was a mistrial (1.7%) or an incorrect response (7.3%) were removed prior to RT analysis. The remaining RTs for the words were again submitted to the recursive data trimming procedure, which resulted in the removal of an additional 1.5% of the data. Mean RTs and mean percentage errors from the subject analysis can be seen in Table 3 for all conditions (item means and *z* scores can be downloaded at http://artsweb.uwaterloo.ca/~dbesner2/publications.html). We report and discuss the vincentile plots after we report Experiment 3.

RTs. Words presented brightly were read aloud faster than those in the dim condition, $F_1(1, 23) = 207$, $MSE = 1{,}336$, $p < .001$, $F_2(1, 195) = 1{,}571$, $MSE = 0.071$, $p < .001$. Regular words were read aloud faster than exception words, $F_1(1, 23) = 94$, $MSE = 285$, $p < .001$, $F_2(1, 195) = 29.3$, $MSE = 0.46$, $p < .001$. There was an interaction between stimulus quality and regularity, $F_1(1, 23) = 5.2$, $MSE = 259$, $p < .05$, though this was not significant in the item analysis, $F_2(1, 195) = 0.8$, $MSE = 0.071$, $p = .4$. The 95% confidence interval around the interaction was ± 7 ms. The regularity effect was smaller under the dim condition compared with the bright condition.

Errors. There was a main effect of stimulus quality, $F_1(1, 23) = 21.4$, $MSE = 8.1$, $p < .001$, $F_2(1, 196) = 14.4$, $MSE = 45.2$, $p < .001$. More errors were made to exception words than to regular words, $F_1(1, 23) = 71.4$, $MSE = 25.8$, $p < .001$, $F_2(1, 196) = 26.3$, $MSE = 259$, $p < .001$. There was no interaction between stimulus quality and regularity (*F*s < 1).

Discussion

The results of Experiment 2 are clear in one respect. Exception words are not *more* affected by low stimulus quality than are regular words; they are *less* affected. This effect in which the slower of two conditions is less affected by the action of a second factor that also slows RT is highly unusual in the context of standard reading-aloud experiments (though not in the context of the psychological refractory period paradigm; e.g., see Besner, Reynolds, & O'Malley, 2009; Reynolds & Besner, 2006).[3] Experiment 3 was therefore conducted to determine whether the interaction observed in Experiment 2 is replicable.

Experiment 3

Method

Participants. Thirty-six undergraduate students from the University of Waterloo participated for \$5 or for course credit. All were native English speakers and reported normal or corrected-to-normal vision.

Procedure. The stimuli, apparatus, and procedure were identical to that of Experiment 2.

Results

One participant was dropped due to poor performance (less than 75% correct). For the remaining 35 participants, trials on which there was a mistrial (2.8%) or an incorrect response (8.4%) were removed prior to RT analysis. The remaining RTs were submitted to the recursive data trimming procedure, which resulted in the removal of an additional 1.5% of the data. Mean RTs and mean percentage errors from the subject analysis can be seen in Table 3 for all conditions. (Item means and z-scores are available at http://artsweb.uwaterloo.ca/~dbesner2/regxsqexp.html)

RTs. Words presented brightly were read aloud faster than those in the dim condition, $F_1(1, 34) = 34.5$, $MSE = 6{,}899$, $p < .001$, $F_2(1, 195) = 750$, $MSE = 0.052$, $p < .001$. Regular words were read aloud faster than exception words, $F_1(1, 34) = 144$, $MSE = 449$, $p < .001$, $F_2(1, 195) = 45.7$, $MSE = 0.39$, $p < .001$. Most centrally, there was an interaction between the effects of stimulus quality and regularity, $F_1(1, 34) = 5.6$, $MSE = 280$, $p < .05$, though it was not significant in the item analysis, $F_2(1, 195) = 1.9$, $MSE = 0.052$, $p = .17$. The 95% confidence interval for the interaction was ± 6 ms.

Errors. There was a main effect of stimulus quality, $F_1(1, 34) = 8.1$, $MSE = 21.1$, $p < .01$, $F_2(1, 195) = 12.1$, $MSE = 34.8$, $p < .05$. More errors were made to exception words than to regular words, $F_1(1, 34) = 76$, $MSE = 43.4$, $p < .001$, $F_2(1, 195) = 38.1$, $MSE = 239$, $p < .001$. There was no interaction between stimulus quality and regularity (*F*s < 1).

Combined Analysis of Experiments 2 and 3

Because Experiments 2 and 3 were essentially identical, we also computed an analysis in which the data were combined. Namely, an ANOVA was carried out in which experiment was not a factor. Critically, this analysis yielded an interaction between stimulus quality and regularity for both subjects and items, $F_1(1, 58) =$

[3] Though it never achieved significance, we note that Herdman et al.'s (1999) experiments showed a small trend toward underadditivity between case mixing and regularity (for low-frequency words).

11.2, $MSE = 272$, $p < .01$, and items, $F_2(1, 195) = 3.9$, $MSE = 0.032$, $p < .05$. There was no significant interaction in the errors, $F_1(1, 58) = 0.2$, $MSE = 20.1$, $p > .1$; $F_2(1, 195) = 0.7$, $MSE = 18.8$, $p > .1$.

Vincentiles. The mean vincentiles for the data from Experiments 2 and 3 are plotted as a function of regularity and stimulus quality in Figure 3. The difference scores (exception words − regular words) for bright and dim items can be seen in Figure 4 (note the scale differences between Figures 3 and 4). Figure 4 shows that the regularity effect is larger under the bright condition than under the dim condition, particularly so for the slower vincentiles.

Cross-Experiment Comparison: Experiment 1 Versus Experiments 2 and 3

We also undertook a cross-experiment analysis in which we compared the results of Experiment 1 for the words with those of Experiments 2 and 3 (treated as a single experiment). Critically, the three-way interaction of Experiments × Stimulus Quality × Regularity was significant by subjects, $F_1(1, 80) = 4.1$, $MSE = 250$, $p < .05$, although it was not significant by items ($F_2 < 1$). The lack of an interaction in the overall item analysis is not surprising, given that inspection of the RT distribution and the associated confidence intervals shows that the interaction in Experiments 2 and 3 is specific to items in the slow end of the distribution.

In short, experiments with small numbers of items (as are too often seen in psycholinguistic experiments) are unlikely to have the power to detect such interactions. The present analyses support the view that the different patterns observed in Experiment 1 versus those seen in Experiments 2 and 3 are genuine.

Discussion

The separate results of Experiments 2 and 3 reveal a novel pattern in which the slowing induced by low stimulus quality affected the faster condition (regular words) more than the slower one (exception words); the pattern can been seen even more powerfully in the combined analysis, where the interaction was also significant in the item analysis. This is a result unanticipated by any theoretical discussion of visual word recognition of which we are aware.

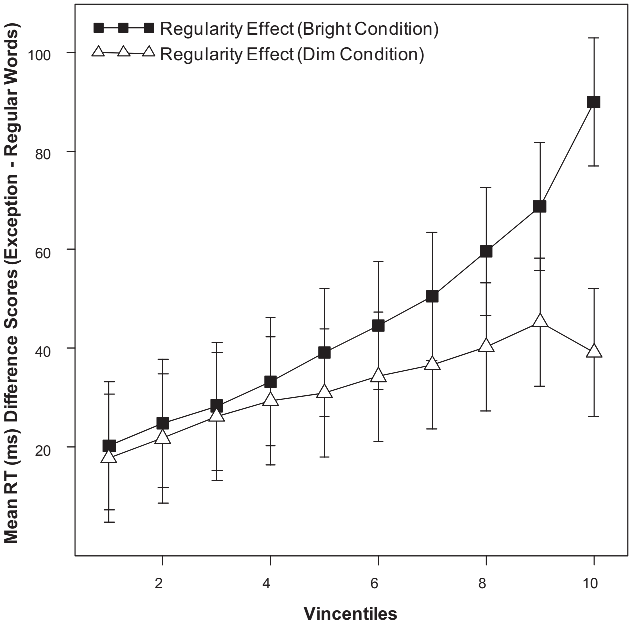

Figure 4. Experiment 2 and 3 combined: The difference in the vincentile means for exception versus regular words for participants' reading aloud times. The bars represent 95% confidence intervals. RT = reaction time.

Simulations. We turn now to a consideration of the two most successful computational accounts of reading aloud. Here we report the results of a series of simulations with the DRC and CDP+ models to determine whether they are able to simulate the results of the experiments reported here. To anticipate the findings:

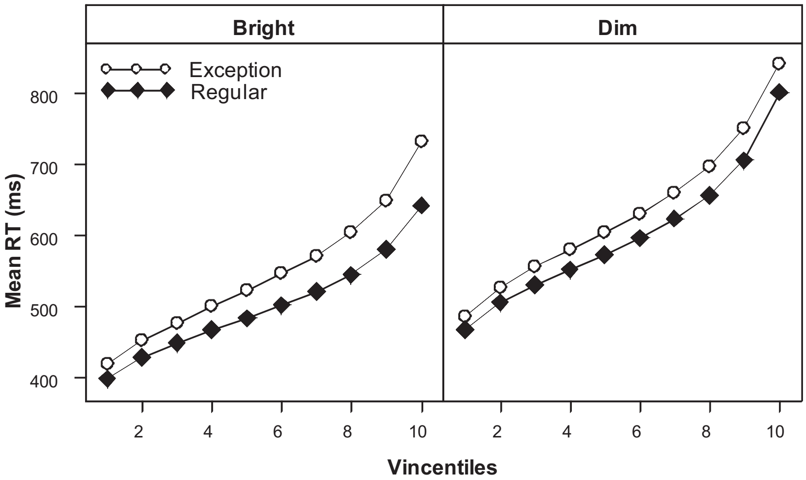

Figure 3. Experiment 2 and 3 combined: Vincentile means for participants' reading aloud times as a function of regularity and stimulus quality. RT = reaction time.

Neither model correctly simulates the joint effects of any of the factors considered here, despite yielding robust main effects of stimulus quality, regularity, and lexicality.

In both DRC and CDP+ we initially simulated low stimulus quality with the method used in the context of the DRC model by Besner and Roberts (2003) and Reynolds and Besner (2004), in which feature to letter level activation is reduced (from .005 to .004) and the feature to letter inhibition is reduced (from –.15 to –.12). These data can be seen in Table 4. (Item means from all of the simulations can be downloaded at http://artsweb.uwaterloo.ca/~dbesner2/publications.html)

Regularity × Stimulus Quality.

DRC. The model made no errors, thus all 197 items were included in the analysis. Words in the bright condition were read in fewer cycles than those in the dim condition, $F(1, 195) = 2{,}677$, $MSE = 0.588$, $p < .001$. Regular words were read in fewer cycles than exception words, $F(1, 195) = 347$, $MSE = 72.5$, $p < .001$. Most critically, there was a interaction in which exception words were more affected by low stimulus quality than regular words, $F(1, 195) = 240$, $MSE = 0.588$, $p < .001$, a result not seen in any of the experiments reported here.

CDP+. The model made seven errors to bright exception words and nine errors to dim exception words. If an error was made to an item in one condition, it was also removed from the other condition. This left 187 items in the analysis. Words in the bright condition were read in fewer cycles than those in the dim condition, $F(1, 185) = 3{,}687$, $MSE = 2.7$, $p < .001$. Regular words were read in fewer cycles than were exception words, $F(1, 185) = 64$, $MSE = 356$, $p < .001$. Most critically, there was an interaction in which exception words were more affected by low stimulus quality than regular words, $F(1, 185) = 48$, $MSE = 2.7$, $p < .001$, again inconsistent with all the results of the experiments reported here.

Table 4
Mean Cycles to Criterion and Mean Percentage Errors (%E) for DRC and CDP+ as a Function of Lexicality, Regularity, and Stimulus Quality

	Bright		Dim		Interaction	
Stimulus type	Cycles	%E	Cycles	%E	Cycles	%E
DRC						
Exception	92.6	0.0	97.7	0.0		
Regular	77.8	0.0	80.6	0.0		
Nonwords	156.8	0.0	160.5	3.0		
Differences						
Exception − Regular	14.8	0.0	17.1	0.0	2.3^{**}	0.0
Nonwords − Regular	79.0	0.0	79.9	3.0	0.9^{*}	3.0
Nonwords − Exception	64.2	0.0	62.8	3.0	$-1.4^{\dagger,**}$	3.0
CDP+						
Exception	117.2	7.3	128.8	9.3		
Regular	102.7	0.0	112.0	0.0		
Nonwords	155.0	1.0	159.0	4.0		
Differences						
Exception − Regular	14.5	7.3	16.8	9.3	2.3^{**}	2.0
Nonwords − Regular	52.3	1.0	47.0	4.0	$-5.3^{\dagger,**}$	3.0
Nonwords − Exception	37.8	−6.3	30.2	−5.3	$-7.6^{\dagger,**}$	1.0

Note. Low stimulus quality was simulated by reducing feature to letter level activation from .005 to .004, and feature to letter level inhibition was reduced from −.15 to −.12.
† A negative interaction value indicates that the effect of stimulus quality is greater for words than for nonwords.
$^{*}p < .05$. $^{**}p < .001$.

Ziegler, Zorzi, and Perry (2009) reported simulations using the CDP+ model aimed at producing additive effects of stimulus quality and word frequency, given that this is the pattern reported by O'Malley and Besner (2008) in three experiments with skilled readers. Ziegler et al. claimed that CDP+ is able to simulate this pattern.

There are two key differences between the simulations reported above and those reported by Ziegler et al. (2009). First, both activation and inhibition from features to letters were reduced here to simulate the effect of stimulus quality, whereas Ziegler et al. only reduced activation. To us, modifying only the activation from feature to letter level lacks a strong theoretical rationale. For humans, stimulus quality is generally conceived of as a very early visual manipulation. It therefore seems likely that its effects are best thought of as reducing input to the network. We cannot directly manipulate input strength in the implemented models; reducing the strength of the output from the feature level will achieve the same end—but only if all output (both excitatory and inhibitory) is weakened. Ziegler et al. argued that changing only one parameter is more parsimonious, which is true only within the confines of the model. When placed in the context of a broader visual processing system, it is difficult to see why the effect would be selective for only one type of output, and if it is, why it is selective for the excitatory connections rather than the inhibitory connections.

Second, Ziegler et al. (2009) also reduced the activation weighting on the letter to orthographic lexical level, arguing that it is central to allowing the nonlexical route to play more of a role and therefore also central to producing additive effects of stimulus quality and word frequency. A reviewer of the present work argued that our failure to implement both of these parameter changes in CDP+ is the reason we do not see additivity between stimulus quality and regularity. Our rationale for not reducing the weights on the letter to orthographic input lexicon rests on the well-known fact that the balance between lexical and nonlexical routines in terms of their respective strength is a central issue in both CDP+ and in DRC. Thus, when CDP+ reads exception words, decreasing the influence of the lexical route is problematic because this renders the nonlexical route too dominant; therefore, CDP+ is more likely to read exception words as though they are regular in terms of their spelling sound correspondences. Indeed, Besner and O'Malley (2009) showed that adopting Ziegler et al.'s parameter set to simulate additive effects of stimulus quality and word frequency yielded a large interaction in the error data, such that low-frequency words were more impaired by low stimulus quality than high-frequency words. Further inspection showed that this interaction was almost entirely due to the items in those experiments that were low-frequency exception words, which is exactly what is to be expected on the basis of the theoretical analysis offered above, as well as by Besner and O'Malley (2009).

Nonetheless, we provide further evidence of this problem by presenting CDP+ with the stimulus set from the present experiment, this time using Zeigler et al.'s (2009) "mixed list" parameter set, in which the letter level to orthographic input lexicon activation value is reduced from .075 to .0598. Also following Ziegler et al., we simulate the low stimulus quality condition by reducing the feature to letter activation from .005 to .001 but leaving the inhibition parameter value untouched. These data can be seen in Table 5.

The first thing to note is that with these parameter values, the model makes a very large number of errors to exception words (over 50%) when stimulus quality is reduced. Given that skilled readers do not make anything close to this amount of errors, it is evident that this version of the model does not capture what skilled readers do. Nonetheless, for completeness we report an ANOVA on the remaining stimuli. There was a main effect of stimulus quality, $F(1, 134) = 2{,}605$, $MSE = 272$, $p < .001$; a main effect of regularity, $F(1, 134) = 55$, $MSE = 675$, $p < .001$; and an interaction in which exception words are again more impaired than regular words by low stimulus quality, $F(1, 134) = 40$, $MSE = 2$, $p < .001$. Thus, even when the model produces the correct phonological code, CDP+ does not, with this parameter set, simulate the pattern produced by skilled readers.

Finally, we did further simulations with CDP+ using a smaller reduction of feature to letter level activation to reduce the number of errors in the exception word condition. We were able to bring the error rate down in this condition to acceptable levels (around 10%); however, the interaction between stimulus quality and regularity in the cycles to criterion measure remains significant in the wrong direction.

Lexicality × Stimulus Quality. The next set of simulations addresses the issue of whether DRC and CDP+ can simulate the joint effects of lexicality (words vs. nonwords) and stimulus quality. As we noted earlier, the human data yielded additive effects for both RTs and errors (see Table 3). The data from DRC and CDP+ are reported in Table 4.

Regular words versus nonwords. The parameters used to simulate low stimulus quality as above were again used (feature to letter activation = .004; feature to letter inhibition is set to –.12).

DRC. A 2 × 2 ANOVA on the cycles to criterion for correct responses yielded a main effect of lexicality, $F(1, 295) = 966$, $MSE = 867$, $p < .001$; a main effect of stimulus quality, $F(1, 295) = 211$, $MSE = 6.6$, $p < .001$; and a small interaction in which nonwords are more affected by low stimulus quality than the regular words, $F(1, 295) = 4.0$, $MSE = 6.6$, $p < .05$.

CDP+. A 2 × 2 ANOVA on the cycles to criterion for correct responses yielded a main effect of lexicality, $F(1, 298) = 272.8$, $MSE = 1{,}320$, $p < .001$; a main effect of quality, $F(1, 298) = 11{,}808$, $MSE = 0.50$, $p < .001$, and an interaction in which nonwords are less affected by low stimulus quality than are regular words, $F(1, 298) = 1840$, $MSE = 0.50$, $p < .001$.

Table 5
Mean Cycles to Criterion and Mean Percentage Errors (%E) for CDP+ as a Function of Regularity and Stimulus Quality

Stimulus type	Bright		Dim	
	Cycles	%E	Cycles	%E
Exception	125.4	12.5	250.6	57.2
Regular	113.7	0.0	211.2	5.0
Difference	11.7	12.5	39.4	52.2

Note. Low stimulus quality was simulated by reducing feature to letter level activation from .005 to .001, and the contribution of the lexical route was reduced by changing letter to orthographic input lexicon activation from .075 to .0598 for both bright and dim conditions, following Ziegler et al. (2009).

Exception words versus nonwords.

DRC. A 2 × 2 ANOVA on the cycles to criterion for correct responses yielded a main effect of lexicality, $F(1, 292) = 570$, $MSE = 919$, $p < .001$; a main effect of stimulus quality, $F(1, 292) = 370$, $MSE = 6.9$, $p < .001$; and an interaction in which nonwords are less affected by low stimulus quality than are exception words, $F(1, 292) = 10.5$, $MSE = 6.9$, $p = .001$. Clearly, the pattern seen with DRC does not mimic what is seen with the skilled readers in either of these two cases.

CDP+. A 2 × 2 ANOVA on the cycles to criterion for correct responses yielded a main effect of lexicality, $F(1, 285) = 107$, $MSE = 1{,}492$, $p < .001$; a main effect of stimulus quality, $F(1, 285) = 5{,}811$, $MSE = 1.2$, $p < .001$; and an interaction in which nonwords were again less affected by low stimulus quality than the exception words, $F(1, 285) = 1{,}376$, $MSE = 1.2$, $p < .001$.

Summary of the Simulation Results

Aside from the main effects of regularity, lexicality,[4] and stimulus quality, neither of these computational models simulated additive effects of stimulus quality and regularity as seen in Experiment 1, additive effects of stimulus quality and lexicality in Experiment 1, or the correct form of the interaction between stimulus quality and regularity in Experiments 2 and 3.

General Discussion

The results of the experiments with skilled readers are clear. First, Experiment 1 shows that when words and nonwords are randomly intermixed there are robust main effects of stimulus quality and regularity and stimulus quality and lexicality, but no hint of any interactions in either RT or errors or in the general distribution of RTs. None of these joint effects are simulated by either the DRC or the CDP+ models.

Second, Experiments 2 and 3 show that when only words appear in the experiment, there is an interaction in which low stimulus quality affects exception words less than regular words. Neither of the computational models simulated this result either. Instead, they always produced an interaction in which exception words are more affected by low stimulus quality than are regular words.

The discrepancies reported here between the models' performance and skilled readers' performance thus call for some form of

[4] Presented with our stimulus set, both models produce significant differences between the exception words and the nonwords in both cycle times and errors. Experiment 1 yielded no difference between these item types in RT and only a small one in the errors. The RT data suggests that the nonlexical route in these models is too slow relative to the lexical route. However, speeding the nonlexical route would likely increase error rates to exception words, potentially exacerbating the discrepancy in the error data. It is worth noting that the nonwords have a higher neighborhood density (see Table 2), which has been shown to speed RTs in reading aloud (e.g., Andrews, 1992; Reynolds & Besner, 2004). The models may not be as sensitive to this factor as human subjects. Whatever the case, the models are clearly not properly capturing this effect.

modification to these models. What modification would make it possible for these models to simulate not only these data but others as well (e.g., those summarized in Table 1), and how would this change our current understanding of the underlying processes?

Accounting for Additivity

O'Malley and Besner (2008; see also Besner & Roberts, 2003; Reynolds & Besner, 2004) proposed that the effect of stimulus quality is restricted to a point shared by both lexical and nonlexical routines when words and nonwords are mixed together and read aloud (or nonwords alone are read aloud). The consequence is that a reduction in stimulus quality yields the same amount of delay to both lexical and nonlexical routes. Given that the only front-end processes common to both routes are feature and letter levels, and given the assumption that low stimulus quality affects both levels, a sufficient assumption is that the output of the letter level is the same under bright and dim conditions, although the time to generate such an output is delayed by low stimulus quality. On this analysis, if letter level processing does not pass activation on until letters have been identified, this would produce additive effects of stimulus quality and word frequency, as reported by O'Malley and Besner (2008), and would also produce additive effects of stimulus quality and regularity when reading aloud, as observed here in Experiment 1. This assumption leads to the prediction that low stimulus quality will affect nonwords and words to the same extent, which is also what was observed with skilled readers in Experiment 1 here. Phrased differently, this solution assumes that at least some processes are serially organized and discrete (Sternberg, 1969, 1998).

A different account is more in keeping with one of the processing assumptions that has been made in implementing these models. Although it is not widely appreciated, under certain conditions it is possible for purely cascaded processes to produce additive effects of two factors on mean RT (e.g., when the factors influence the rates of two different, relatively fast processes; see McClelland, 1979; Roberts & Sternberg, 1993). Nonetheless, our preference is for the discrete processing (thresholded) account, because (a) it is simpler, (b) the evidence for discrete processes in various RT tasks is both broad and deep (see Roberts & Sternberg, 1993), and (c) none of the current computational models of visual word recognition to date are purely cascaded (they include components engaged in interactive activation).

Accounting for the Interaction: Stimulus Quality × Word Frequency

O'Malley and Besner (2008) proposed that when stimulus quality and word frequency interact (i.e., when only words appear in the experiment), this reflects processing along the lexical route that has reverted to cascaded processing or interactive activation. Consistent with this proposal, Reynolds and Besner (2004) reported simulations with the DRC model that produced an interaction between stimulus quality and word frequency when the lexical route was either purely cascaded or engaged in interactive activation.

Accounting for the Interaction: Stimulus Quality × Regularity

We emphasize that though both regularity and word frequency interact with stimulus quality in reading aloud when only words appear in the experiment, the nature of the interactions is qualitatively different. Stimulus quality and regularity yield an interaction in which the slower condition (exception words) yields a smaller increase in RT relative to the faster condition (regular words) when stimulus quality decreases. When stimulus quality and word frequency interact, the slower condition (low-frequency words) is more affected by reductions in stimulus quality than is the faster condition (high-frequency words). We know of no account on the table that can accommodate these results, and certainly the two models examined here do not produce them with the parameters we have explored thus far. Below we offer one plausible hypothesis that can explain the interaction between stimulus quality and regularity.

An Independent Influence of Stimulus Quality on the Nonlexical Route When Only Words Appear

Stated baldly, stimulus quality may have an influence on the nonlexical route that is independent of the general slowing of the processing system (which affects both routes equally) when only words appear in the experiment. The regularity effect arises due to competition between the two routes in the phonemic buffer when naming exception words. If the nonlexical route is more strongly influenced by low stimulus quality in this context, then on dim trials there will be less opportunity for the nonlexical route's regularized pronunciation to compete with the lexical route's correct pronunciation, and thus the regularity effect will be smaller. Because the nonlexical route does not contribute to the word frequency effect in the same way, low-frequency words do not benefit from this differential influence of stimulus quality.

A Simulation Proof

Conceptually, this account is straightforward, and thus it is not surprising that our attempts to simulate the pattern were successful. For an existence proof we turned to the most successful manipulation of stimulus quality (reducing all output from the feature level, as in the simulations in Table 4), but in addition, we further slowed the nonlexical route by reducing the strength of the excitatory connections from the nonlexical processes (the grapheme–phoneme conversion in the DRC and the two-layer assembly network in CDP+) to the phonemic buffer.[5] The idea here is that the reduction in feature level output represents a general slowing of the system that is not route specific, whereas the reduction in the output of the nonlexical route represents an additional slowing of this route relative to the lexical route. As is clear from Table 6, both models now successfully simulate the RT pattern found with

[5] We are agnostic about where exactly this additional slowing takes place. It may be in the connections from the letters to the nonlexical processes, within the processes themselves, or in the connections from the processes to the phonemic buffer. In the implemented models, we do not have access to parameters that would allow us to simulate the first two, so we implemented the third.

Table 6

Mean Cycles to Criterion and Mean Percentage Errors (%E) for DRC and CDP+ as a Function of Stimulus Quality and Regularity and Stimulus Quality and Word Frequency, With a Greater Slowing of the Nonlexical Route

	DRC				CDP+			
	Bright		Dim		Bright		Dim	
Stimulus type	Cycles	%E	Cycles	%E	Cycles	%E	Cycles	%E
1. Stimulus quality by regularity								
Exception	92.4	1.0	96.3	0.0	117.7	7.3	133.1	3.1
Regular	77.8	0.0	84.9	0.0	102.7	0.0	120.3	0.0
Difference	14.6	1.0	11.4	0.0	15.0	7.3	12.8	3.1
2. Stimulus quality by word frequency								
Low frequency	78.3	0.0	85.2	0.0	100.7	0.0	116.6	0.0
High frequency	73.6	0.0	79.0	0.0	83.7	0.0	93.4	0.0
Difference	4.7	0.0	6.1	0.0	17.0	0.0	23.2	0.0

Note. In DRC, the low stimulus quality condition was simulated by reducing feature to letter level activation from .005 to .003, feature to letter level inhibition from −.15 to −.09, and the output from the nonlexical route from .055 to .04. In CDP+, low stimulus quality was simulated by reducing feature to letter level activation from .005 to .004, feature to letter level inhibition from −.15 to −.12, and the output from the nonlexical route from .085 to .05.

humans in Experiments 2 and 3 while preserving the pattern observed between stimulus quality and word frequency. In the simulations of the present experiments (see Table 6, top panel), the main effects of stimulus quality and regularity and the interaction between them is significant for both models. In simulating the O’Malley and Besner (2008) data when only words are present, both models show main effects of stimulus quality and word frequency and a significant interaction between them that takes the correct form.[6]

How Does the Differential Influence of Stimulus Quality on the Two Routes Arise?

Descriptively and computationally, our account successfully produces the pattern of RT data found in Experiments 2 and 3. This leaves the question of why stimulus quality would influence the nonlexical route more than the lexical route when only words appear in the experiment. We can envision only two possibilities: Either it is simply an emergent property of the system, or it is a result of dynamic changes to the system. In either case, we do not think that subjects are anticipating a dim trial; rather, when they encounter a dim trial, they dampen the nonlexical route. That is, there exists a process that monitors the activation over time at the feature and/or letter level. When this activation is “slow” in the context of an experiment such as ours, this leads to a dampening of the nonlexical route (for related examples of online parameter adjustment, see Bub, Masson, & Lalonde, 2006; Coltheart, Davelaar, Jonasson, & Besner, 1977; Reynolds & Besner, 2005, 2008; Ridderinkhof, 2002). We discuss these two possibilities (emergent property vs. dynamic control) below.

Emergent Property

It may be that the system is organized in such a way that a reduction in stimulus quality affects the nonlexical route more than the lexical route (i.e., there is an independent effect of stimulus quality on the nonlexical route). If this is the case, then stimulus quality is not simply an input strength manipulation, as computational modelers have implicitly assumed in their simulation work (and most theorists have assumed in their theorizing), but has a much more complex relationship with the processes involved in reading aloud. Such a change in thinking about the effects of stimulus quality requires reconsideration of a body of work. To begin, any simulation result that has involved a manipulation of stimulus quality needs to be reexamined using a route-selective manipulation like the one described here (though note that we have already demonstrated that the interaction between word frequency and stimulus quality is preserved when only words appear in the list). It also raises the question of where else in the system stimulus quality is having additional influences and what those influences look like. In short, if the present data reflect an emergent property of how stimulus quality manipulations affect processing, then those engaged in the study of visual word recognition processes have overlooked this influence, with potentially serious ramifications for theories and models.

Dynamic Control

Another possibility is that the greater slowing of the nonlexical route by low stimulus quality when only words appear in the list reflects the reading system trying to achieve its goal: to read aloud both rapidly and accurately. If reducing the quality of the stimulus has a greater influence on the lexical route than on the nonlexical route, then on dim trials there will be more opportunity for the nonlexical route to influence pronunciation. For exception words, this increases the likelihood of regularization. It may be, then, that on detecting a dim stimulus the system dampens the nonlexical route to prevent these regularizations. This account raises two questions.

[6] We also considered the hypothesis that the system might increase feedback along the lexical route on dim trials to more strongly support the lexical contribution to the pronunciation. Our attempts to simulate this approach were unsuccessful; even increasing the feedback by a factor of 15 had little effect on cycle times in the CDP+ model.

First, given that only words appear in Experiments 2 and 3, why not completely suppress the nonlexical route? One possible reason is structural; there may be restrictions on the extent to which the route can be dampened. For example, it may not be possible to turn the route off entirely. This view is held by many theorists; phonological processes are automatic in the sense of being ballistic and impossible to prevent. Another possibility is that the reader still needs the nonlexical route on some trials. All of the words used here are low frequency. Enough items may be unknown to subjects that they need the nonlexical route to decipher them correctly (this would work for the regular words, though of course this would result in errors to any unknown exception words).

The second question raised by the dynamic control account is that if dampening the nonlexical route reduces regularizations, why not dampen it during bright trials as well? One answer is that although the nonlexical route does slow reading of exception words and can result in regularizations, it is not widely appreciated that it has the opposite effect for regular words. Both routes are converging on the same pronunciation for regular words; the nonlexical route is thus facilitating rapid responding on half of the trials (indeed, simulations with the CDP+ model reveal that responses to regular words when both routes are operating are much faster than when only the lexical route is operating, and this benefit is larger than the cost of the difference between regular and exception words when both routes are operating). Slowing the nonlexical route across all trials would thus incur a speed penalty on bright trials where the system is already reading at a high rate of accuracy. On dim trials, the system accepts the reduction in speed as a fair trade for improvements in accuracy.

Regularizations and the Nonlexical-Route-Specific Influence of Stimulus Quality

Our data and the present simulations do not offer a way of distinguishing between the emergent property and dynamic control accounts, and thus we are agnostic at present as to which account should be preferred. What we can confidently conclude from our data is that something about regularity differs from any other linguistic factor studied to date (at least of those that have been jointly manipulated with stimulus quality). Simulation results support the hypothesis that the difference arises from a specific influence of low stimulus quality on the nonlexical route. However, this route-specific influence makes a prediction that we can address by examining the types of errors that both the models and our subjects produce. Specifically, if the nonlexical route is dampened on dim trials, then regularizations should be less common. The simulation results support this prediction. The error data for CDP+ in Table 6 clearly show a pattern where the number of errors to exception words when the nonlexical route is dampened in the dim condition is smaller than in the bright condition when the nonlexical route is not dampened (in the DRC model there is a weak trend in the same direction, but it is less prominent due to the very small number of errors in general). This is at least consistent with our intuition that the present account predicts reduced regularizations under dim conditions. Do human readers show a similar pattern?

To examine whether or not humans also showed a tendency toward fewer regularizations under dim conditions (consistent with the idea that the nonlexical route is playing a lesser role), we classified errors to exception words according to whether they were regularizations or not. These data are summarized in Table 7.

Selecting a Baseline

The question of whether or not there is a reduction in the number of regularizations on dim trials necessitates a discussion of the appropriate comparison condition. The ideal comparison is between dim trials with and without the extra slowing of the nonlexical route. However, that comparison cannot exist in human subjects; the nonlexical route either is or is not locally attenuated on dim trials.

One possibility is to compare the error pattern in Experiment 1 (nonwords present) to those of Experiments 2 and 3 (words only). There are problems with this comparison. First, it assumes that the additional attenuation of the nonlexical route does not apply when nonwords are present. If attenuation is present in both conditions, then the critical factor is not varied. The difficulty is that our theory is agnostic with respect to whether route-specific attenuation is operating in Experiment 1 or not: If the attenuation is an emergent property, then we assume that it would be operative; otherwise, it may or may not be operative. Second, using this baseline ignores the influence of thresholding on regularizations. If we assume for the moment that the route-specific attenuation is not operating in Experiment 1, then reducing stimulus quality should increase the number of regularizations. However, thresholding the output should counteract this effect and reduce the number of regularizations. Potentially confounding these two theoretical mechanisms makes interpretation of the results difficult.

The other possibility is to compare regularizations under dim conditions to those under bright conditions. This comparison has the advantage that the theory explicitly assumes that the attenuation of the nonlexical route is absent in the baseline (bright) and present in the comparison (dim) condition. Thresholding is also held constant across the two conditions (present for the Experiment 1 comparison, absent for the Experiment 2 and 3 comparisons, by hypothesis). The difficulty with this baseline is that it confounds the general dimming of the stimuli (which ought to increase the number of regularizations—and errors in general) with the nonlexical-route-specific slowing (which ought to reduce the number of regularizations). Because these factors influence

Table 7
Regularizations and Other Errors in Response to Exception Words in Experiments 1, 2, and 3

	Bright		Dim	
Error types	# Errors	% Errors	# Errors	% Errors
Experiment 1				
Regularizations	82	52.9	69	45.5
Other	72	47.1	83	54.6
Experiment 2				
Regularizations	77	57.5	84	52.5
Other	57	42.5	76	47.5
Experiment 3				
Regularizations	124	53.9	110	42.5
Other	106	46.1	149	57.5

regularizations in opposite directions, the results will underestimate the real influence of route-specific attenuation.

Both baselines have their problems. Use of the bright condition as a baseline biases against finding an effect by pitting the general slowing of the system against the route-specific attenuation of the nonlexical route. However, we are of the opinion that this is preferable to using the Experiment 1 dim condition as a baseline, as it is unclear whether or not we would be varying the factor of interest: local attenuation of the nonlexical route.

Words only. The results from Experiment 3 are consistent with our account. There is a lower proportion of regularizations under dim than bright conditions, $\chi^2(1, N = 489) = 6.391, p < .05$. Experiment 2 is identical to Experiment 3, and the trend is toward a similar pattern in the error data; however, it is not significant, $\chi^2(1, N = 294) < 1$. When we collapse the two experiments together (as they are effectively the same experiment), the effect remains significant, $\chi^2(1, N = 783) = 6.199, p < .05$. These results suggest that humans are behaving in a way that is consistent with the predictions of a low-stimulus-quality effect specific to the nonlexical route.[7]

Nonwords and words. The error data from Experiment 1 show a similar trend toward a reduction in the number of regularizations. However, this trend is also not significant, $\chi^2(1, N = 306) = 1.887, p > .16$, though as argued earlier, it is less clear what is going on in Experiment 1 with respect to nonlexical-route-specific attenuation.

We stress that error data were not a primary dependent variable in our experiments. Nonetheless, in all three experiments the relative number of regularizations is reduced in the dim condition. This reduction is significant for Experiment 3 and for the two words-only experiments combined, though not for Experiments 1 or 2 independently. In our view this provides some support for the theory that stimulus quality is having a route-specific effect that is not currently considered by theories of visual word recognition. That being said, the present data are clearly not strong enough to provide a final word on the subject.

Some readers may be uncomfortable with the idea of a dynamically controlled process rather than an emergent process. Though our simulation results and data cannot differentiate the two accounts, this is not the first time that context-dependent dynamic processing has been proposed (e.g., see Bub et al., 2006; Coltheart et al., 1977; Reynolds & Besner, 2005, 2008; Ridderinkhof, 2002). There are considerable extant data that appear to require that processing vary across contexts in ways that have not been widely entertained. Empirically, when words and nonwords are intermixed in the context of reading aloud (or when only nonwords are presented), the effect of stimulus quality is to simply add a constant to RT. This is true for the joint effects reported here, as well as for the joint effects of stimulus quality and letter length when reading nonwords aloud (Besner & Roberts, 2003), stimulus quality and word frequency (O'Malley & Besner, 2008), and stimulus quality and *N* when reading nonwords aloud (Reynolds & Besner, 2004). Thus, if other ways of modifying processing are to be proposed, they must accommodate these facts too.[8]

Our simulations are successful in that they provide an existence proof that, when configured this way, both models produce the same pattern of RTs and regularization data as produced by skilled readers in Experiments 2 and 3. Of course, this assumption (that nonlexical processing is more affected by low stimulus quality than is the lexical route when only words are being read) is entirely post hoc.[9] Moreover, although the data can be simulated by the existing models discussed here by simply changing a single parameter, to do so they require a module that neither model possesses at present. That is, our hypothesis is that a module exists that monitors rate of evidence accumulation at feature and/or letter level; when it is too slow (as when stimulus quality is low) this either directly serves to change the single parameter in the nonlexical route or signals some local "control" module, which in turn does this. This hypothesis strikes us as simple, plausible on its face, and sufficient in that it allows the models to simulate the highly unusual data pattern seen in Experiments 2 and 3.

Conclusions

The results of the three experiments reported here constitute a set of novel observations that any viable theory of reading aloud needs to be able to explain. Neither of the computational models discussed here are, in their current form, able to simulate all the results from the present experiments (nor related ones in the literature). However, we have advanced the idea that when words are read alone (i.e., with no nonwords), low stimulus quality has a specific influence on the nonlexical route over and above its effect on feature and letter level processing (which is common to both routes). When this idea is tested with a simulation, it produces both the RT and regularization patterns that we observed with human subjects in Experiments 2 and 3. We have also appealed to a thresholding account (offered previously in related contexts) to accommodate additive effects of stimulus quality and regularity and stimulus quality and lexicality like those observed in Experiment 1 when words and nonwords are randomly intermixed.

It remains to be seen whether any of the proposals offered here will be implemented in these or any other computational models or whether other ways of understanding and simulating these effects will emerge. Most generally, the present results can be taken as a

[7] These results were obtained despite the fact that errors and error types were not our main dependent variable of interest. Indeed, instructions to the participant are typically designed to avoid having complicated data patterns in the errors to facilitate interpretation of the RT data.

[8] There is one fact that does not fit this empirical generalization: Repetition interacts with stimulus quality for words (but not nonwords) in the presence of nonwords (Blais & Besner, 2007). It remains to be seen whether these findings hold up when word frequency is also manipulated within the same experiment and whether word frequency will add or interact with stimulus quality in that context.

[9] One might be tempted to suppose, therefore, that skilled readers should yield larger effects of stimulus quality when reading only nonwords compared with only exception words, on the grounds that nonwords must be read via the nonlexical route and only the lexical route can correctly read aloud exception words. However, this prediction is predicated on the assumption that the nonlexical route is more affected by low stimulus quality than the lexical route in all cases; our theory argues that this is true when processing along the nonlexical route is not thresholded. When nonwords (blocked or mixed) are read, we argue that the nonlexical route is thresholded and make no claims about the effects of stimulus quality along that route. In short, the nonword cell cannot be tested, because when nonwords are read aloud the assumption is that both the lexical and nonlexical routes are thresholded to explain the data from Experiment 1 (as well as the data from Besner & Roberts, 2003; Reynolds & Besner, 2004).

set of phenomena that elude computational models in their currently implemented forms. These data also reinforce the conclusion that what seem like small changes in context can have profound effects on how some of the underlying processes unfold over time. Put differently, the processes underlying visual word recognition and reading aloud appear considerably more dynamic than generally envisioned to date.

References

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., Piepenbrock, R., & van Run, H. (1993). The CELEX lexical database [Database]. Philadelphia, PA: University of Pennsylvania, Linguistic Data Consortium.

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

Besner, D., & O'Malley, S. (2009). Additivity of factor effects in reading tasks is still a challenge for computational models: Reply to Ziegler, Perry, and Zorzi (2009). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 312–316.

Besner, D., Reynolds, M. G., & O'Malley, S. (2009). When underadditivity of factor effects in the psychological refractory period paradigm implies a bottleneck: Evidence from psycholinguistics. *Quarterly Journal of Experimental Psychology, 62,* 2222–2234.

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., & Smith, M. C. (1992). Models of visual word recognition: When obscuring the stimulus yields a clearer view. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 468–482.

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.

Bub, D. N., Masson, M. E. J., & Lalonde, C. E. (2006). Cognitive control in children: Stroop interference and suppression of word reading. *Psychological Science, 17,* 351–357.

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. (2009). Reading aloud: Evidence for contextual control over lexical activation. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 499–507.

Fiset, D., Arguin, M., Bub, D., Humphreys, G. W., & Riddoch, J. M. (2005). How to make the word length effect disappear in letter-by-letter dyslexia: Implications for an account for the disorder. *Psychological Science, 16,* 535–541.

Fiset, S., Arguin, M., & Fiset, D. (2006). An attempt to simulate letter-by-letter dyslexia in normal readers. *Brain and Language, 98,* 251–263.

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 program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers, 35,* 116–124.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106,* 491–528.

Herdman, C. M., Chernecki, D., & Norris, D. (1999). Naming case alternated words. *Memory & Cognition, 27,* 254–266.

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

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

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

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

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

O'Malley, S., & Besner, D. (2008). Reading aloud: Qualitative differences in the relation between stimulus quality and word frequency as a function of context. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1400–1411.

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., 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.

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., & Coltheart, M. (1998). Whammies and double whammies: The effect of length on nonword reading. *Psychonomic Bulletin, 5,* 277–282.

R Development Core Team. (2004). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

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). Contextual control over lexical and sublexical routines when reading English aloud. *Psychonomic Bulletin & Review, 12,* 113–118.

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.

Ridderinkhof, K. R. (2002). Activation and suppression in conflict tasks: Empirical clarification through distributional analyses. In W. Prinz & B. Hommel (Eds.), *Attention and Performance XIX: Common mechanisms in perception and action* (pp. 494–519). Oxford, England: Oxford University Press.

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.

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

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.

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

Vincent, S. B. (1912). The function of vibrissae in the behavior of the white rat. *Behavioural Monographs, 1*(5), 1–84.

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.

Ziegler, J., Perry, C., & Zorzi, M. (2009). Additive and interactive effects of stimulus quality: No challenge for CDP+. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 306–311.

Received October 7, 2009
Revision received January 11, 2010
Accepted January 18, 2010 ■