Journal of Experimental Psychology:
Learning, Memory, and Cognition
2009, Vol. 35, No. 1, 312–316

0278-7393/09/$12.00 DOI: 10.1037/a0014555

Additivity of Factor Effects in Reading Tasks Is Still a Challenge for Computational Models: Reply to Ziegler, Perry, and Zorzi (2009)

Derek Besner and Shannon O'Malley
University of Waterloo

J. C. Ziegler, C. Perry, and M. Zorzi (2009) have claimed that their connectionist dual process model (CDP+) can simulate the data reported by S. O'Malley and D. Besner. Most centrally, they have claimed that the model simulates additive effects of stimulus quality and word frequency on the time to read aloud when words and nonwords are randomly intermixed. This work represents an important attempt given that computational models of reading processes have to date largely ignored the issue of whether it is possible to simulate additive effects. Despite CDP+'s success at capturing many other phenomena, it is clear that CDP+ fails to capture the full pattern seen with skilled readers in these experiments.

Keywords: reading aloud, stimulus quality, word frequency, additive effects

Computational modeling of processes that subserve reading aloud in both skilled readers and those with developmental or acquired dyslexia has a relatively brief history in cognitive psychology. Nonetheless, there has been broad, deep, and rapid development, and these efforts have been highly successful on a number of fronts. For example, some of these models can simulate the well-documented main effects (in studies of skilled readers) of word frequency, spelling–sound regularity, lexical density, consistency, letter length, and pseudohomophony, among others. There are also successful simulations of prominent two-way interactions—such as Frequency × Regularity, Serial Position × Regularity, and Lexicality × Letter Length—and even three-way interactions between Orthography, Frequency, and Regularity, and between Repetition, Frequency, and Regularity (e.g., see Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Perry, Ziegler, & Zorzi, 2007).

To date, however, computational modelers of reading-related processes have largely ignored the issue of whether it is possible to simulate additive effects of factors on reaction time (RT; but see Besner, Wartak, & Robidoux, 2008; Plaut & Booth, 2000, 2006). It is unclear why there is so little interest in this issue, but one hypothesis is that—following the seminal work of McClelland and Rumelhart (1981; see also McClelland, 1987)—the dominant framework for language processing consists of *interactive activation.* The problem is that this framework does not lend itself to the standard interpretations of additive effects on RT: Namely, that (a) additive effects of two factors reflect two serially arranged and *discrete* processes, each affected by a different factor (e.g., Process A is affected by Factor *g* but not Factor *h*, and Process B is affected by Factor *h* but not Factor *g*; see Sternberg, 1969, 1998) or (b) *cascaded* processing in which Process A is affected by Factor *g* but not Factor *h*, and Process C is affected by Factor *h* but not Factor *g* (McClelland, 1979; see also Roberts & Sternberg, 1993; note also that there are some boundary conditions that need to be respected for a cascade account to produce additive effects on mean RT). Indeed, Besner (2006) has speculated that purely interactive activation (IA) models of reading processes may be unable to simulate systematically additive effects of two factors on RT. If this speculation is correct, then, to the extent that there are demonstrations of such additivity in skilled readers, a purely IA model would be ruled out. The computational issue has yet to be resolved (but see Besner, 2006, for a number of examples of additive effects in reading tasks, and Yap, Balota, Tse, & Besner, 2008, for new demonstrations of additive effects of stimulus quality and word frequency in lexical decision that extend through much of the RT distribution).

A related but distinct question is whether *any* current computational model of reading processes can simulate additive effects of two factors on RT. This is a distinct question because the most successful models do not consist of processes that are *all* engaged in IA. Rather, these models include some processes that are engaged in IA, others that are only cascaded, and still other processes that are discrete (thresholded). Both Coltheart et al.'s (2001) dual-route cascaded model (DRC) and Perry et al.'s (2007) connectionist dual process model (CDP+) are instances of such hybrid models.

Given this background, we now address the commentary provided by Ziegler, Perry, and Zorzi (2009; hereafter referred to as *ZPZ*) in response to O'Malley and Besner (2008). To recapitulate the point addressed by ZPZ, O'Malley and Besner reported a three-way interaction in the context of reading aloud in which the joint effects of word frequency and stimulus quality interact when only words appear in the list but are additive when nonwords are randomly intermixed with the words. O'Malley and Besner suggested that these results pose an important challenge for all current

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

This work was supported by Natural Sciences and Engineering Research Council of Canada Grant A0998 to Derek Besner. We thank J. Ziegler and D. Plaut for useful reviews.

Correspondence concerning this article should be addressed to Derek Besner, Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada. E-mail: dbesner@uwaterloo.ca

computational models of the processes involved in reading aloud. Their claim was that the observed pattern is most easily explained in terms of the idea that when nonwords are mixed with the words, the output of the letter level is thresholded so as to help prevent lexicalization errors (reading a nonword as a word), particularly so when the stimulus is low quality. However, when words are blocked, then processing across feature, letter, and input lexical levels is, minimally, cascaded (and may be engaged in IA).

ZPZ take up O'Malley and Besner's challenge in the context of their CDP+ model. ZPZ's central claim is that their CDP+ model does indeed simulate additive effects of these two factors in the context of reading aloud. We comment briefly on ZPZ's results, and we report some new simulations. Our conclusion is that the inferences that ZPZ draw from their results are not warranted. CDP+ does not, at present, capture additivity of factor effects from the experiments reported by O'Malley and Besner (2008), and it does not capture the relation between stimulus quality and lexicality seen in these experiments. We make 2 main points.

1. The Error Problem

When cognitive psychologists do speeded reading aloud experiments they typically emphasize accuracy, and the primary dependent measure is RT. However, they also report and consider the error data. Claims about additivity of factor effects in the RT data are, naturally, constrained by the pattern in the error data. ZPZ do not discuss the distribution of errors across the four conditions in their simulation of O'Malley and Besner's (2008) Experiment 3, and they do not report them in their figures. However, they do provide an item appendix that contains this information. The percentage of error for the four conditions can be seen in Table 1. This table shows that there is a large interaction in which CDP+ makes the most errors in the low-stimulus quality condition for low-frequency words (17.4%), whereas the other three conditions yield a negligible number of errors (less than 3.0%). This interaction is significant, $F(1, 134) = 8.6$, $MSE = 0.026$, $p < .01$. The additive effects seen in the cycles to criterion data are therefore qualified by the presence of this substantial interaction in the error data. Note that when the item set from Experiment 1 of O'Malley and Besner is run through CDP+ it also produces an interaction in the error data, $F(1, 132) = 7.0$, $MSE = 0.024$, $p < .01$. These data can also be seen in Table 1.

Table 1
Percentage of Errors Made by CDP+ to the Single-Syllable Words From O'Malley and Besner (2008) in Experiments 1, 2 and 3, in Which the Letter to Orthography Parameter Is Changed From .075 to .0598 and Low-Stimulus Quality Is Simulated by Reducing Feature to Letter Level Activation From .005 to .001

	Error %	
Item type	Clear	Degraded
Items from Experiments 1 and 2		
Low frequency	2.9	12.9
High frequency	4.6	4.6
Items from Experiment 3		
Low frequency	2.9	17.4
High frequency	1.5	1.5

Note. CDP+ = Perry, Ziegler, and Zorzi's (2007) connectionist dual process model.

It might, at first blush, be argued that the additivity between word frequency and stimulus quality in the cycles data and the interaction seen in the error data is not problematic for CDP+ because this is the pattern seen in 2 of the 3 experiments reported by O'Malley and Besner (2008). However, O'Malley and Besner discussed this issue at some length. They pointed out that when a median split of the data is performed on the basis of overall errors across all three of their experiments, the group that made more errors produced additive effects in RT but an interaction in the error data, whereas the group that made fewer errors had additive effects in both RT and errors. They therefore suggested that when there is additivity in RT but an interaction in the error data, subjects are trading speed for accuracy. The critical data to pay attention to are when the additivity in the RT data is not contradicted by an interaction in the error data. These data are not simulated by CDP+ given that the simulation produces a very large interaction in the error data, but subjects produce no such interaction.

On the Nature of the Errors Produced by CDP+

It is also instructive to consider the nature of the items that produce errors in the low-frequency, low-stimulus quality condition in the simulation of Experiment 3 by CDP+. Examination of these items (which can be seen in the appendix of the ZPZ's article) reveals that *10 of 11* of them are *exception* words according to CDP+ (though note that 5 of these items are not exception words according to DRC). That is, they violate the typical spelling–sound correspondence rules (e.g., *PINT* would be pronounced as in *MINT, DINT, LINT*; *HAVE* would be pronounced as in *CAVE, RAVE, SAVE, GAVE*). In contrast, in Experiment 3 of O'Malley and Besner (2008), none of the subjects produced an error to 8 of 11 of these items. Furthermore, errors by subjects were spread across a wide range of the remaining 58 items selected by ZPZ. In short, the distribution of the errors in this condition by skilled readers looks nothing like the errors produced by CDP+.

A more important point concerns *why* CDP+ produces errors to exception words in this condition. The answer is straightforward. A central and well-known issue in dual-route models with lexical and nonlexical routines (as in CDP+ and DRC) concerns the balance between these routes (i.e., their relative strength). If the nonlexical route is too strong relative to the lexical route, then the model will make errors to many low-frequency exception words, whereas if the lexical route is too strong relative to the nonlexical route, then the model will not produce what is seen with skilled readers: a regularity effect (slower responses to lower frequency exception words than to matched regular words). What ZPZ did to try and simulate additivity between word frequency and stimulus quality when words and nonwords were mixed together was to reduce the strength of the lexical route. In short, they ignored the balance problem identified above with the consequence that the model makes a number of errors to the low-frequency exception words in the low-stimulus quality condition. In other words, the particular parameter setting that ZPZ chose is contraindicated for the result that they wish to produce when the item set contains low-frequency exception words.

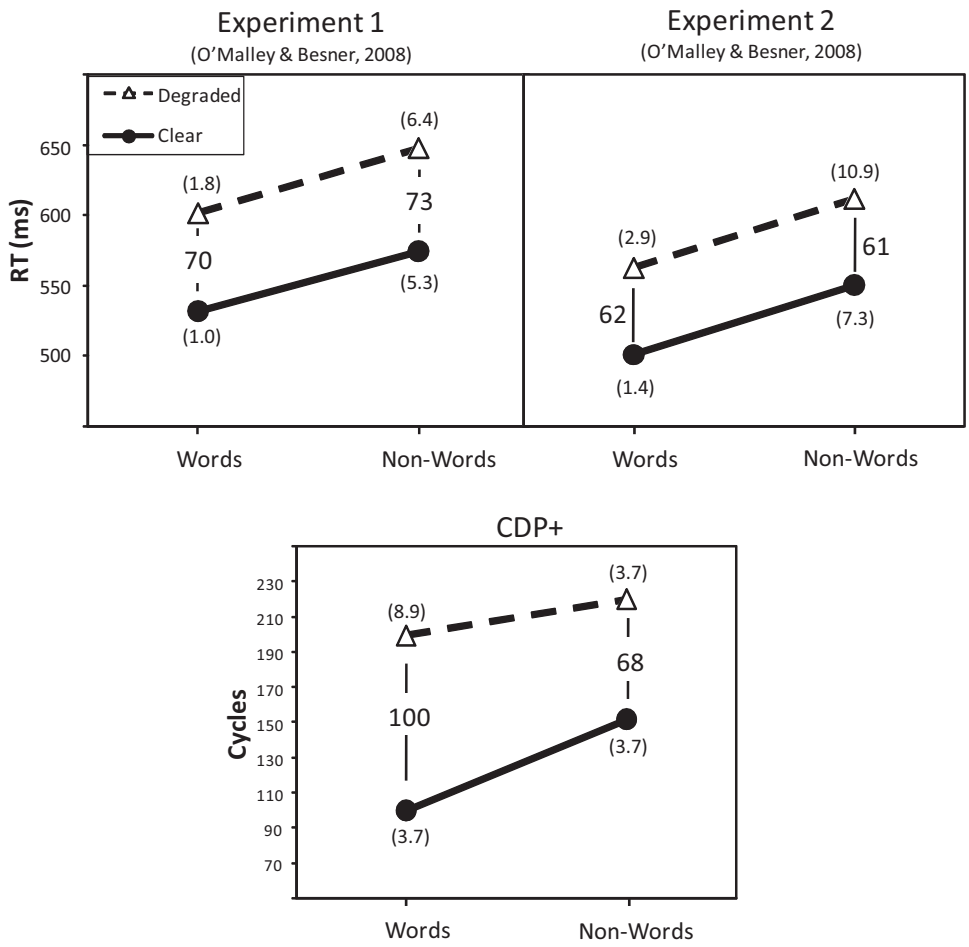

Figure 1. Mean reaction times (RTs; in milliseconds) and percentage errors (in parentheses) from O'Malley and Besner's (2008) Experiments 1 and 2 as a function of stimulus quality and lexicality along with mean cycles to criterion and percentage error in CDP+ for the single-syllable items from these experiments. CDP+ = Perry, Ziegler, and Zorzi's (2007) connectionist dual process model.

Some readers might assume that the outcome that ZPZ wish to produce could be obtained by a smaller reduction in the strength of the lexical pathway than used at present. We do not see how this could work given that using CDP+'s default parameter set produces an interaction between stimulus quality and *regularity* in which exception words are slowed more by low-stimulus quality and produce more errors than regular words. Weakening the connections between letter and orthographic input lexicon simply makes that interaction larger (Besner, O'Malley, Robidoux, & Fox, 2008). We therefore conclude that an approach that only reduces the strength of the connections between the letter level and the orthographic lexicon so as to produce additive effects of stimulus quality and word frequency when reading aloud is not workable. Whether this approach would work when combined with some other parameter adjustment(s) is currently unknown.

2. The Joint Effects of Stimulus Quality and Lexicality

ZPZ emphasized that the CDP+ model's nonlexical route is *thresholded*, whereas the lexical route is cascaded and engaged in feedback between several levels. They also emphasized that this thresholding is central to any attempt to simulate additivity of stimulus quality and word frequency. We would like to emphasize that whatever approach is adopted in the future will also need to address other results that are seen within the experiments reported by O'Malley and Besner (2008)—in particular, whether such a model simulates the relation between the effect of stimulus quality and lexicality (words vs. nonwords). Below, we report several analyses of variance (ANOVAs) for both the subject and item analysis for all three experiments reported by O'Malley and Besner along with simulation data from CDP+.

In Experiment 1, there was a main effect of stimulus quality, $F_1(1, 31) = 59.1$, $MSE = 2{,}805$, $p < .001$, $F_2(1, 398) = 1193.6$, $MSE = 885$, $p < .001$; a main effect of lexical status, $F_1(1, 31) = 47.9$, $MSE = 1{,}347$, $p < .001$, $F_2(1, 398) = 158$, $MSE = 2{,}668.5$, $p < .001$; but no interaction, $F_1(1, 31) = 0.6$, $MSE = 134.6$, $p > .40$, $F_2(1, 398) = 0.35$, $MSE = 885$, $p > .50$. In the error analysis, there was a main effect of stimulus quality, $F_1(1, 31) = 5.8$, $MSE = 5.5$, $p < .05$, $F_2(1, 398) = 8.4$, $MSE = 23.5$, $p < .01$; a

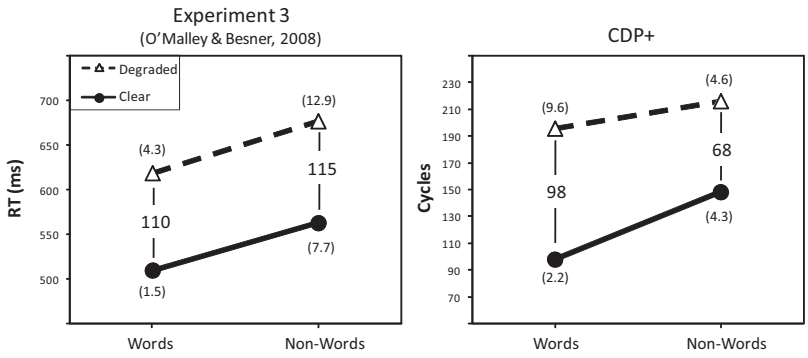

Figure 2. Mean reaction times (RTs; in milliseconds) and percentage errors (in parentheses) from O'Malley and Besner's (2008) Experiment 3 as a function of stimulus quality and lexicality along with mean cycles to criterion and percentage error in CDP+ for the single-syllable items from this experiment. CDP+ = Perry, Ziegler, and Zorzi's (2007) connectionist dual process model.

main effect of lexicality, $F_1(1, 31) = 103.1$, $MSE = 6.1$, $p < .001$, $F_2(1, 398) = 78.3$, $MSE = 23.5$, $p < .001$; and no interaction, $F_1(1, 31) = 0.10$, $MSE = 4.4$, $p > .50$, $F_2(1, 398) = 0.13$, $MSE = 23.5$, $p > .70$.

In Experiment 2, there was a main effect of stimulus quality, $F_1(1, 31) = 164.9$, $MSE = 730$, $p < .001$, $F_2(1, 398) = 548$, $MSE = 1{,}318$, $p < .001$; a main effect of lexical status, $F_1(1, 31) = 74$, $MSE = 1{,}059$, $p < .001$, $F_2(1, 398) = 141$, $MSE = 3{,}772$, $p < .001$; but no interaction, $F_1(1, 31) = 0.007$, $MSE = 135$, $p > .90$, $F_2(1, 398) = 0.004$, $MSE = 1{,}317$, $p > .90$. In the errors analysis, there was a main effect of stimulus quality, $F_1(1, 31) = 22.7$, $MSE = 28.4$, $p < .001$, $F_2(1, 398) = 37.4$, $MSE = 35.7$, $p < .001$; and a main effect of lexicality, $F_1(1, 31) = 55$, $MSE = 28.4$, $p < .001$, $F_2(1, 398) = 129.6$, $MSE = 76$, $p < .001$. There was also an interaction in which there was a larger effect of stimulus quality for nonwords, $F_1(1, 31) = 3.8$, $MSE = 9.1$, $p = .06$, $F_2(1, 398) = 6.5$, $MSE = 35.7$, $p < .05$.

Finally, in O'Malley and Besner's (2008) Experiment 3 (using only the data from the condition in which words and nonwords are intermixed), there is a main effect of stimulus quality, $F_1(1, 31) = 298$, $MSE = 1{,}364$, $p < .001$, $F_2(1, 398) = 2{,}501$, $MSE = 999$, $p < .001$; a main effect of lexicality, $F_1(1, 31) = 144$, $MSE = 682$, $p < .001$, $F_2(1, 398) = 193.6$, $MSE = 3{,}060$, $p < .001$; and no interaction, $F_1(1, 31) = 1.9$, $MSE = 157.4$, $p > .15$, $F_2(1, 398) = 1.3$, $MSE = 999$, $p > .20$. In the error analysis, there was a main effect of stimulus quality, $F_1(1, 31) = 36.5$, $MSE = 13.60$, $p < .001$, $F_2(1, 398) = 79.9$, $MSE = 38.7$, $p < .001$; and a main effect of lexicality, $F_1(1, 31) = 44.3$, $MSE = 38.3$, $p < .001$, $F_2(1, 398) = 135$, $MSE = 78.7$ $p < .001$. Again, there was an interaction in which there was a larger effect of stimulus quality for nonwords, $F_1(1, 31) = 6.1$, $MSE = 7.8$, $p < .05$, $F_2(1, 398) = 7.6$, $MSE = 38.7$, $p < .01$.

The RT data are clear. All three experiments yielded additive effects of stimulus quality and lexicality. The error data undermine the interpretation the claim that the RT data are genuinely additive. The interaction in the error data for Experiments 2 and 3 is consistent with the suggestion that subjects sometimes prematurely generate a pronunciation, particularly so when stimulus quality is low, and when they do so they generate an error. As we noted earlier, O'Malley and Besner (2008) discussed this possibility in detail when considering the error data in response to words by doing a median split based on total number of errors. Subjects who made more errors yielded additivity between stimulus quality and word frequency in RT but an interaction in the error data. Subjects who made fewer errors yielded additive effects in both RT and errors. We have not had a chance to explore whether this conclusion also applies to the nonwords data noted here. Further, there may be other reasons why the nonwords are more impaired by low stimulus quality than the words (e.g., the letters may be more confusable with each other than they are for the words; see Fiset, Arguin, Bub, Humphreys, & Riddoch, 2005). For present purposes, however, the important point is that nonwords are not *less* impaired by stimulus quality than are the words.

Does CDP+ simulate the relation between lexicality and stimulus quality observed for the skilled readers? In a word: no. We conducted two simulations with CDP+ using the single-syllable items from O'Malley and Besner's (2008) experiments. Following ZPZ, the lexical route was de-emphasized (letter to orthographic input lexicon activation was reduced from .075 to .0598), and a reduction in stimulus quality was simulated by reducing feature to letter level activation from .005 to .001. The first simulation used the single-syllable words and nonwords from Experiments 1 and 2 of O'Malley and Besner (2008; all items means from the experiments and simulation means are available upon request). A 2 (Stimulus Quality) × 2 (Lexical Status) ANOVA yielded a main effect of stimulus quality, $F(1, 304) = 20{,}229$, $MSE = 50.9$, $p < .001$, and a main effect of lexicality, $F(1, 304) = 161$, $MSE = 1{,}184$, $p < 001$. Critically, there is an interaction between stimulus quality and lexicality in which the effect of low stimulus quality is *smaller* for nonwords than for words, $F(1, 304) = 719$, $MSE = 50.9$, $p < .001$. Obviously, there is nothing in the error data that undermines interpretation of the RT data (see Figure 1).

The second simulation used the single-syllable items from Experiment 3. A 2 (Stimulus Quality) × 2 (Lexical Status) ANOVA yielded a main effect of stimulus quality, $F(1, 297) = 18{,}388$, $MSE = 54.2$, $p < .001$, and a main effect of lexicality, $F(1, 297) = 210$, $MSE = 883$, $p < .001$. Critically, there is an interaction between stimulus quality and lexicality in which the effect of low stimulus quality is again *smaller* for nonwords than for words, $F(1, 297) = 598$, $MSE = 54.2$, $p < .001$. Again, there is nothing in the

error data that undermines interpretation of the RT data. The data from both the skilled readers and the simulations can be seen in Figures 1 and 2.

This discrepancy between what skilled readers produce and what CDP+ produces demonstrates that the model does not capture the pattern produced by skilled readers when reading mixed lists of words and nonwords in either the cycles or error measures. Why is this? One possibility (again) is that the parameters chosen by ZPZ are inappropriate. Perhaps there is another parameter set that would produce the right outcome. That said, we note that the default parameter set for CDP+ produces the same pattern as here. Our hypothesis is that the problem is more fundamental, having to do with the fact that the lexical route is cascaded and engaged in feedback between various levels, whereas the nonlexical route is thresholded. To put it another way, we are inclined to the hypothesis that the letter level is thresholded for both lexical and nonlexical routes.

Summary and Conclusions

By way of summary, CDP+ can simulate additive effects of stimulus quality and word frequency on the cycles to criterion measure. However, it is also clear that the error data from ZPZ's simulations undermine the claim that the model produces genuine additivity. CDP+ also fails to capture the pattern of joint effects of stimulus quality and lexicality from the same experiments in both cycles to criterion and error measures.

In conclusion, although the CDP+ model is very successful at simulating a large number of phenomena (arguably, it is currently among the best on the table) and ZPZ's work represents an important attempt at simulating the relation between factor effects observed in speeded reading aloud, additive effects still pose a fundamental challenge to computational endeavors.

References

Besner, D. (2006). Visual language processing and additive effects of multiple factors on timed performance: A challenge for the interactive activation framework? *PsyCrit: Critical Commentary and Review of Papers in Psychology.* Retrieved from http://psycrit.com/

Besner, D., O'Malley, S., Robidoux, S., & Fox, R. (2008, June). *On the joint effects of spelling-sound regularity and stimulus quality when reading aloud: New challenges for computational models.* Paper presented to the Canadian Society for Brain, Behaviour and Cognitive Science, University of Western Ontario, Canada.

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.

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.

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.

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.

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.

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.

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.

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.

Yap, M. J., Balota, D. A., Tse, C., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency in 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. C., Perry, C., & Zorzi, M. (2009). Additive and interactive effects of stimulus degradation: No challenge for CDP+: Comment on O'Malley and Besner (2008). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 306–316.

Received July 24, 2008
Revision received October 30, 2008
Accepted October 30, 2008 ■