Modulation of Additive and Interactive Effects in Lexical Decision by Trial History

Michael E. J. Masson
University of Victoria

Reinhold Kliegl
University of Potsdam

Additive and interactive effects of word frequency, stimulus quality, and semantic priming have been used to test theoretical claims about the cognitive architecture of word reading processes. Additive effects among these factors have been taken as evidence for discrete-stage models of word reading. We present evidence from linear mixed-model analyses applied to two lexical decision experiments indicating that apparent additive effects can be the product of aggregating over- and underadditive interaction effects that are modulated by recent trial history, particularly the lexical status and stimulus quality of the previous trial's target. Even a simple practice effect expressed as improved response speed across trials was powerfully modulated by the nature of the previous target item. These results suggest that additivity and interaction between factors may reflect trial to trial variation in stimulus representations and decision processes rather than fundamental differences in processing architecture.

Keywords: additive and interactive effects, effects of trial history, lexical decision, linear mixed models

Many formal models of word reading processes assume the existence of separate processing modules that are responsible for computing different types of information (e.g., deriving orthography from visual input; computing phonology from orthographic patterns; selecting semantic information that corresponds to a presented word). A fundamental issue in the design of such models is the manner in which information is shared between the constituent modules. It is commonly assumed that input between processing modules is cascaded, so that partial information moves between modules even before a module has completed its operations (e.g., McClelland, 1979), and most current models of visual word identification include feedback from higher (e.g., lexical) to lower (e.g., letter) levels of processing (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Perry, Ziegler, & Zorzi, 2007; Plaut & Booth, 2000).

These architectural features play an important role in allowing computational models to simulate some fundamentally important behavioral results involving

word identification tasks, and in particular, interactions between independent variables that influence the speed with which words are identified. In word identification tasks such as naming aloud or lexical decision, semantic priming of target words is enhanced if the targets are presented in a visually degraded form such as low contrast (e.g., Becker, 1979; Borowsky & Besner, 1991, 1993; Stanovich & West, 1983). The typical form of this interaction is *overadditive*, whereby the influence of semantic priming is greater when performance is slowed by a low-contrast stimulus. In general, an overadditive interaction is one in which the simultaneous influence of two independent variables is larger than what would be expected by a simple summation of their individual effects. For example, unrelated primes and low stimulus quality both slow responding relative to related primes and high stimulus quality, respectively. When combined, the two former conditions may yield an increase in response times that is larger than would be produced by adding their two separate effects.

An example of how an overadditive interaction may be generated is illustrated in Figure 1A, where the influences of stimulus quality and priming on the accumulation of evidence and setting of response thresholds are shown. This depiction is consistent with Morton's (1969) logogen model, in which it is assumed that evidence in support of a particular word accrues over time and when the amount of evidence surpasses a threshold, a response (e.g., naming the word or a positive lexical decision) can be made. In Figure 1A, evidence accumulates more rapidly for a clear stimulus than for a degraded one, so the evidence threshold needed for a response is reached sooner. In the logogen model, the effect of semantic priming is to lower the response threshold for a primed word and this idea is represented in Figure 1A as a lower amount of evidence
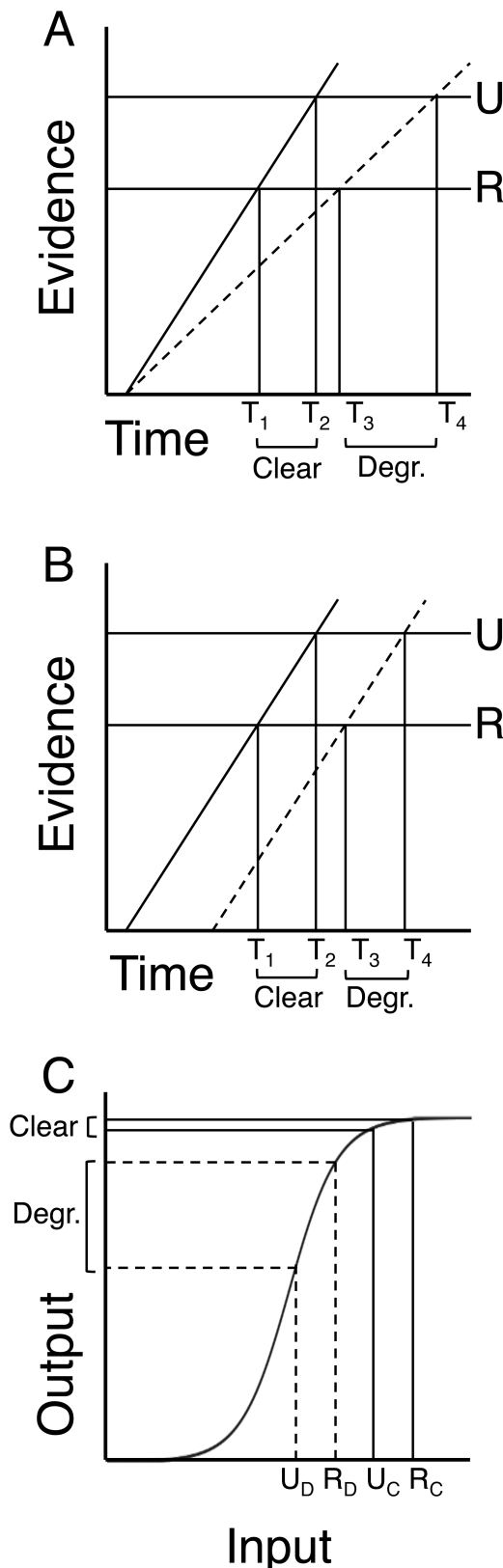
Figure 1. Schematic representation of accrual of evidence (A and B) or activation strength (C) under manipulations of stimulus quality (solid lines = clear, dashed lines = degraded) and semantic priming (R = related, U = unrelated). In A and B, the effect of stimulus quality is shown as a difference in slope or a difference in onset, respectively. Semantic priming affects the evidence threshold needed for a response. In C, the strength of input is affected by stimulus quality (weaker input for degraded stimuli, U_D and R_D, than for clear stimuli, U_C and R_C) and by semantic priming (weaker input for unrelated primes). The shape of the input-output function and the location of input activations along that function determine the combined effects of these two variables. See the text for an explanation of the effects produced by these models. (A and B are adapted from Borowsky & Besner, 1993; C is adapted from Plaut & Booth, 2000.)

required for a response when a related prime is presented. The faster rate of evidence accrual for clear targets leads to a smaller priming effect (T_2 - T_1) than for degraded targets (T_4 - T_3), as shown in the figure.

Another possibility, in keeping with the idea of additive factors (Roberts & Sternberg, 1993; Sternberg 1969), is that degrading the stimulus may delay the start of evidence accrual (perhaps to allow for stimulus "clean-up"), but once started, the rate is the same as for clear targets. As shown in Figure 1B, these assumptions lead to additive effects of stimulus quality and priming.

A particularly flexible approach to account for combined effects of independent variables is exemplified by the Plaut and Booth (2000) connectionist model of word processing. In this model, it is assumed that evidence from visual input is combined with contextual evidence such as semantic primes to generate input to a processing module that produces output leading to a response. The function relating input to output activation is sigmoidal, as shown in Figure 1C. Response time differences are determined by differences in the level of output activation. Depending on the degree of input activation, factors such as stimulus quality and semantic priming may yield interactive or additive effects. Figure 1C illustrates an overadditive interaction, in which priming is larger for degraded targets because input activation for those targets operates in a relatively steep section of the activation function. If both target types had input values in that steep, virtually linear region (imagine shifting the input values to the left in Figure 1C), an additive effect would be produced. Interestingly, if the input activation values were pushed to even lower levels so that the earliest part of the activation function were involved, an *underadditive* interaction would be generated, with larger priming effects found among clear targets.

The fact that computational models, such as the Plaut and Booth (2000) connectionist model, may produce different patterns of interaction or additivity

depending on the overall level of activation (Borowsky & Besner, 2006) raises an interesting possibility. Namely, an observed additive relationship in behavioral data may arise because of a combination of two different patterns of interaction that, when aggregated, yield no indication of an interaction. Some third factor, not considered in standard computational models, may be responsible for modulating the interaction, leading to a spurious additive pattern. In the experiments reported here, we demonstrate that features of the target on the previous trial in an experiment induce such pseudo-additive patterns. The resulting three-way interaction, including trial history as a factor, that ensues from a pair of two-way interactions of opposite pattern may serve as a challenge for future development of computational models of word recognition.

Models of Interactive and Additive Effects

The capability of simulating interactive effects with computational models that assume feedback between modules (e.g., Plaut & Booth, 2000), integration of information from independently operating modules (e.g., Massaro & Cohen, 1991), or combined effects of threshold setting and evidence accumulation rate (Morton, 1969), may become a liability when one considers factors that typically do not interact, such as word frequency and stimulus quality (e.g., Becker & Killion, 1977; Stanners, Jastrzembski, & Westbrook, 1975; Yap & Balota, 2007) or the existence of instances in which factors that usually interact, such as priming and stimulus quality, are additive (e.g., Brown & Besner, 2002; Ferguson, Robidoux, & Besner, 2009; Robidoux, Stolz, & Besner, 2010; Stolz & Neely, 1995). Even putting aside the question of the type of control structure needed to allow a computational model to shift between generating interactive and additive effects of factors on word identification, a critical issue is whether these models are able to produce additive effects at all.

Besner and colleagues (Besner & O'Malley, 2009; Besner, Wartak, & Robidoux, 2008; Borowsky & Besner, 2006) and Plaut and Booth (2006) have examined the ability of the Plaut and Booth (2000) connectionist model, which includes interactive feedback between modules, to simulate additive effects of word frequency and stimulus quality. A critical aspect of this model is that response latency is a linear function of the output of processing units in the model's semantic module, and that output activation is a sigmoid function of the strength of input into the units of that module. Plaut and Booth (2006) were able to simulate additive effects of word frequency and stimulus quality by restricting the level of input to the semantic module to be within a certain range (the linear region of a sigmoid-shaped activation function that translates input activation to output activation). Besner

et al. (2008) demonstrated, however, that in the Plaut and Booth model, as the strength of the simulated manipulation of stimulus quality increased, the pattern of effects changed from an *underadditive* interaction (stronger influence of stimulus quality on high-frequency words than on low-frequency words) to additivity and then to an overadditive interaction (stimulus quality had a larger impact on low-frequency words). Besner et al. noted that the underadditive interaction was a result that has not been seen with skilled readers and it was deemed problematic that the model generated such a result.

Ziegler, Perry, and Zorzi (2009) examined the relationship between stimulus quality and word frequency in the context of the Perry et al. (2007) connectionist dual-process model (CDP+) of word reading. In this dual-route model, words may be read either through a lexical route (learned orthographic patterns directly activate lexical entries) or a nonlexical route in which a phonological code is assembled from orthographic input. Processing in the nonlexical route is thresholded so that activation of units at the letter level must reach a minimum threshold before activation is passed on to the grapheme level. If processing of letter strings is dominated by the nonlexical route, additivity between stimulus quality and word frequency can arise because the effects of reduced stimulus quality can be resolved at the letter level before input is passed to later processing modules. This stage-wise processing configuration is critical in producing additive effects (Borowsky & Besner, 2006). Ziegler et al. argued that in the word naming task, if pure word lists are used, then both the lexical and the nonlexical routes would contribute to the computation of a target word's pronunciation. Because the lexical route is not thresholded, an overadditive interaction between word frequency and stimulus quality (low-frequency words show larger effects of stimulus quality) was produced by the model. A mixed list of words and nonwords, however, was argued to lead to a de-emphasis of the lexical route because it cannot be used to compute the pronunciation of nonwords. Under these conditions, the CDP+ model produced additivity. The shift from overadditive interaction to additivity, tied to the nature of the target items, simulated the response time results in a naming task reported by O'Malley and Besner (2008). We note, however, that Besner and O'Malley (2009) pointed out that although the CDP+ model simulated additive effects of stimulus quality and word frequency in response time, the model produced an interactive effect in error rates. We agree that this fact raises some doubts about this particular model's capabilities, but for present purposes our primary concern is the principle that a trade-off between a thresholded and a non-thresholded process may modulate additivity and interaction between factors.

A parametric examination of the behavior of the

CDP+ model by Ziegler et al. (2009) revealed an additional interesting outcome. If the influence of the lexical route was reduced sufficiently, the word frequency effect became quite small under conditions of low stimulus quality, yielding an underadditive interaction between word frequency and stimulus quality, as had been observed with the Plaut and Booth (2000) model (Besner et al., 2008). This aspect of the model's behavior led Ziegler et al. to propose that individual differences between subjects might be characterized by varying degrees of reliance on the lexical route when naming printed letter strings, and that these differences might be manifest in different patterns in the joint effects of word frequency and stimulus quality. Thus, some subjects might show additivity whereas others could show overadditive and perhaps even underadditive interactions. Aggregating across subjects would likely result in an additive pattern.

Dynamic Adjustment of Processing

Although Besner et al. (2008) pointed out that the underadditive interaction between word frequency and stimulus quality generated by the Plaut and Booth (2000) model had not been observed in human data, the Ziegler et al. (2009) analysis suggests that an interaction of that type may be hidden behind the overall mosaic seen when data are aggregated in the typical manner. Indeed, Besner, O'Malley, and Robidoux (2010) found an underadditive interaction in a naming task when stimulus quality was manipulated along with spelling-sound regularity: exception words (e.g., *pint*) were pronounced more slowly than regular words (e.g., *hint*), and the effect of stimulus quality was stronger for regular words. In dual-route models such as CDP+, exception words take longer to read because they invite two opposing pronunciations, the correct one from the lexical route and an erroneous one from the nonlexical route. Besner et al. (2010) proposed that under low stimulus quality, the nonlexical route is less active, so its contribution to the computation of a pronunciation response is reduced, paving the way for a more efficient, correct pronunciation of an exception word based on the lexical route. They suggested that the involvement of the nonlexical route might be adjusted dynamically as a low-contrast stimulus is encountered. This idea is quite plausible, given earlier demonstrations that response time is sensitive to recent trial history, such as slowing after commission of an error (e.g., Allain, Burle, Hasbroucq, & Vidal, 2009; Laming, 1979; Rabbitt, 1966, 1989).

Another example of dynamic adjustment of processing that can lead to an underadditive interaction was reported by Yap, Balota, Tse, and Besner (2008). They found the usual additive effect of stimulus quality and word frequency in the aggregate data from a lexical decision task. An analysis of response time distributions based on Vincentized data from experiments in which pseudohomophones (e.g., *brane*) were used as nonword targets, however, revealed a more complex pattern. For this particular case, they observed opposing interaction effects, with an overadditive interaction occurring when response times were relatively short, but for longer response times there was an underadditive interaction. Yap et al. suggested that underadditivity may have occurred because subjects sometimes engaged in a checking operation for low-frequency words, regardless of stimulus quality, whereas this operation would be applied to high-frequency words only when stimuli were low contrast. Checking would lead to generally longer response times, and its differential application in the case of high-frequency words would increase the influence of stimulus quality on those items.

Dynamic processing adjustments may also influence the joint effects of stimulus quality and semantic priming. As discussed above, these variables have been shown to yield either an overadditive interaction or additivity. One factor that determines the nature of these joint effects is the proportion of trials on which related primes are used (Stolz & Neely, 1995), with an additive pattern emerging when a low relatedness proportion is in effect. Although no computational model that implements dynamic changes in word processing mechanisms triggered by features of the current stimulus has yet emerged, there are recent formal models that incorporate sensitivity to events or processing fluency experienced on recent trials. For example, Kinoshita, Forster, and Mozer (2008) proposed a model (*Adaptation to the Statistics of the Environment*; ASE) to account for sensitivity to the proportion of repetition-prime trials in masked priming experiments. The model simulated that sensitivity by adapting its response-initiation processes to recent trial difficulty. In particular, evidence about when a response should be initiated by the model is derived from the current trial as well as from recent trials and the results are combined to determine when a response is produced. Given that repetition primes allow for faster responding, a high proportion of such trials makes it likely that the current trial will have been immediately preceded by one or more repetition-prime trials. If the current trial is also a repetition-prime trial, then the combined evidence would encourage a short latency for initiating a response.

The Kinoshita et al. (2008) model leads to the prediction that response time on trial N should be sensitive to the characteristics of trial N - 1. Kinoshita, Mozer, and Forster (2011) tested this prediction in a series of masked priming experiments in which repetition and unrelated primes were used. To provide a statistically powerful examination of their data at the level of individual trials, Kinoshita et al. (2011) used a

linear mixed-model analysis (LMM; e.g., Baayen, Davidson, & Bates, 2008; Kliegl, Masson, & Richter, 2010) and found that as response time on trial $N - 1$ increased, so did response time on the current trial.

The potential for trial to trial variation in processing difficulty to influence the current trial raises an important possibility with respect to patterns of interaction and additivity seen with word frequency, stimulus quality, and priming. Namely, an empirical demonstration of an additive relationship between two factors may conceal an underlying pattern of opposing interactions that, when aggregated, yield the appearance of additivity. A clear example of this possibility was reported by Yap et al. (2008). As discussed earlier, an aggregate additive relationship between word frequency and stimulus quality turned out to be composed of a combined overadditive and underadditive interaction when nonwords were pseudohomophones, with the type of interaction dependent on whether response times were generally short or long, respectively. But the Kinoshita et al. (2011) results suggest a more fine-grained possibility: the combined effects of two factors may vary depending on recent trial history (e.g., whether processing on the previous trial was difficult or easy).

From the perspective of the Plaut and Booth (2000) connectionist model, one might propose that the amount of output activation required for responding varies depending on what was experienced on the preceding trial. This variation would position the system at different points along the function that relates input activation to output activation in semantic processing units (Figure 1C). For example, encountering fluent processing on the previous trial might lead to a lower threshold for responding on the current trial, thereby requiring a lower level of input/output activation. The Plaut and Booth model could produce additivity or either type of interaction, depending on where the system is positioned on the relevant activation function when it reaches a sufficient level of activation for responding.

Moreover, consideration should be given to how subjects adapt to the task of identifying words, particularly when stimulus quality is varied. Under these conditions, rapidly discriminating words from nonwords in a lexical decision task, for example, may depend on stimulus learning that accrues over the course of a substantial number of trials. Turner, Van Zandt, and Brown (2011) have proposed a formal model of stimulus discrimination (signal vs. noise) in which subjects develop signal to noise likelihood ratios for various points along a stimulus strength axis. These ratios are modified by recent trial events in the following way. Presentation of a signal stimulus which has a particular strength value will increase the likelihood that any stimulus with a strength value similar to that of the presented stimulus is also in the

signal category. This system is essentially a Hebbian learning algorithm in which the strength of the connection between a value on the signal strength axis and a response category (e.g., signal) is increased when a signal is presented nearby. Thus, two closely spaced (similar) stimulus values are likely to come to have similar signal response strengths and the recent presentation of one will strengthen the other.

With these possibilities in mind, we used a lexical decision task to examine more closely the relationships between three factors that have been shown to influence word identification processes and that are known to have additive or interactive effects: word frequency, stimulus quality, and semantic priming. Like Kinoshita et al. (2011), we used LMMs to evaluate possible influences of recent trial history on the nature of these relationships. We anticipated that an additive relationship emerging from an analysis of aggregated data might mask a more complex relationship that varies as a function of the difficulty of the preceding trial. Uncovering layered effects of this nature would be an important step in understanding how additive effects arise and whether they should be taken as evidence of serial processing stages (e.g., Borowsky & Besner, 2006) or whether a more elaborate account is required.

There are additional advantages to using LMM instead of the usual analysis of variance (ANOVA) to analyze data from word recognition experiments. First, the inclusion of item factors, such as word frequency, carries with it the question of whether observed effects may be generalized across a broad population of items. This question usually is addressed within ANOVA by computing two sets of F ratios, one using subjects as the random factor and the other using items as the random factor (see Raaijmakers, Schrijnemakers, & Gremmen, 1999, and Raaijmakers, 2003, for a critique of this approach). A more efficient and justifiable method, however, is to model potential random effects of both subjects and items simultaneously using LMM (e.g., Baayen et al., 2008).

Second, the LMM approach allows us to test for the possibility that there are stable individual differences between subjects with respect to semantic priming effects. Stolz, Besner, and Carr (2005) evaluated the reliability of semantic priming effects using split-half and test-retest measures to determine whether individual subjects consistently generated large or small priming effects. Their analyses indicated that semantic priming effects had low reliability, particularly when priming was likely to have only an automatic influence on performance (e.g., because a short SOA was used). Stolz et al. concluded that automatic semantic priming yields little in the way of systematic individual differences. These measures of reliability, however, depend on computing correlations between difference scores, which can themselves have

rather low reliability. In our study, by directly estimating the variance for between-subject semantic priming effects as an LMM parameter, we correct for the unreliability of the subject-based difference scores. Moreover, if we can establish that there are reliable differences between subjects in semantic-priming effects, we may also observe significant correlations of these effects with differences between subjects in mean RT or other experimental effects. These correlations are also estimated as LMM parameters (Kliegl et al., 2010; Kliegl, Wei, Dambacher, Yan, & Zhou, 2011).

Finally, the LMM approach allows a seamless integration of a continuous covariate such as trial number and its interaction with experimental factors of semantic priming, word frequency, and stimulus quality. Such interactions will inform us about how subjects' sensitivity to events or processing fluency changes across the experiment.

Experiment 1

To enhance the possibility of finding additive effects of factors in aggregate data, our manipulation of semantic priming in a lexical decision task consisted of comparing low-associate primes to unrelated primes. In addition, the prime-target SOA was restricted to 200 ms. Under these conditions, priming and stimulus quality have been shown to have additive effects (Stolz & Neely, 1995). We anticipated that word frequency and stimulus quality would produce additive effects when aggregate data are considered, as has commonly been found (e.g., O'Malley et al., 2007; Yap et al., 2008). The potential for semantic priming and word frequency to produce an overadditive interaction was considered to be high given previous results (e.g., Becker, 1979; Borowsky & Besner, 1993), although Yap, Tse, and Balota (2009) reported finding additive effects for these factors among subjects with relatively high vocabulary knowledge.

In addition to examining effects in aggregated data, we used LMMs to determine whether trial history influences the pattern of additivity versus interaction that holds between these three independent variables. Difficulty of responding on a given trial may be influenced by any of these factors, with relatively high difficulty arising from low stimulus contrast, low frequency, or an unrelated prime. In testing the possibility that the joint effects of these factors on the current trial are modulated by difficulty of the previous trial, we anticipated that a more complex account of additive effects in particular might be required.

Method

Subjects. Seventy-two University of Victoria students volunteered to participate in the experiment to earn extra credit in an undergraduate psychology course.

Materials. A list of 240 target words, each item comprised of 4 to 7 letters, along with a related prime word for each target was constructed by supplementing the list of pairs provided by Tse and Neely (2007). Half of the word targets were of relatively high frequency ($M = 170,438$) and the other half were of low frequency ($M = 16,594$), according to frequency norms generated by the English Lexicon Project data base (Balota et al., 2007). These frequencies are based on a corpus of approximately 131 million word tokens (Lund & Burgess, 1996). The mean forward associative strength for related prime-target pairs was .226 for high-frequency targets and .225 for low-frequency targets (Nelson, McEvoy, & Schreiber, 2004). This degree of associative strength is similar to the level of associative strength for items considered to be low strength by Stolz and Neely (1995) who found additive effects of priming and stimulus quality with low strength pairs. A list of 240 pronounceable nonwords was constructed and served as nonword targets. They were one or two syllables and were of similar letter length to the word targets. None were pseudohomophones, but they followed English orthography. An English word was selected to serve as a prime for each of these items. These prime words were similar to those used as primes for the word targets. An additional 32 prime-target pairs (half word targets and half nonword targets) were used as practice items.

The list of high- and low-frequency target items was broken into four sublists of 30 targets of each frequency category. Within each sublist, a second pairing of primes and targets was created to produce unrelated prime-target pairs by reassigning primes to alternative targets within a sublist. Assignment of these sublists to the four experimental conditions created by factorially varying prime relatedness (related vs. unrelated) and stimulus quality (clear vs. degraded) was counterbalanced across subjects so that each target appeared equally often with its related and with its unrelated prime in each of the four conditions. Nonwords were broken into two sublists and assignment of these lists to clear and degraded conditions was counterbalanced across subjects.

Procedure. Subjects were tested individually in a quiet room, seated at computer monitor that was controlled by a Macintosh computer. They were instructed that their task was to classify a series of letter strings as words or nonwords. They were informed that on each trial a target letter string would appear in uppercase letters, but it would be preceded by a briefly presented word in lowercase letters to which no response was to be made. Subjects lightly held the forefinger of each hand on a response button mounted on a box that was connected to the computer. A response with the right hand indicated that the target was a word and a left-hand response was used for

nonwords.

Each trial began with a fixation cross presented for 250 ms, followed by a blank screen for 250 ms, then the prime word in lowercase letters for 200 ms. The target string was then presented in uppercase and remained in view until a response was made. The next trial followed immediately after a correct response. If an incorrect response was made, the message "ERROR" was presented on the screen for 1 s. The fixation cross and primes were presented in black font against a white background, as were targets that were assigned to the clear stimulus-quality condition. For targets in the degraded condition, the black level of the letters was reduced to 20% of the maximum darkness possible, yielding a light gray, low-contrast image.

Subjects were first presented with 32 practice trials with equal numbers of word and nonword targets and within each item class, equal numbers of items were tested in each possible condition. The practice trials were followed by a randomly ordered sequence of 480 critical trials consisting of 240 word and 240 nonword targets. Subjects were provided evenly spaced breaks throughout the course of testing.

Results and Discussion

We begin by presenting a standard analysis of response times and error rates based on ANOVA, followed by LMM analyses specifically designed to test the possibility that trial history modulates the joint effects of word frequency, stimulus quality, and priming.

Analysis of variance. Response times for word targets less than 200 were classified as spoils and not included in the analysis. Response times longer than 1,700 ms were excluded as outliers. This cutoff was chosen so that no more than 0.5% of correct responses would be excluded (Ulrich & Miller, 1994). Mean correct response time for word targets in each of the eight conditions was computed for each subject and the means obtained by averaging across subjects are shown in Figure 2. An ANOVA indicated that all three main effects were significant $Fs(1,71) > 43$, $ps < .001$, with shorter response times in the expected cases: clear stimulus quality (64-ms effect), related prime (16-ms effect), and high-frequency target (23-ms effect). None of the interactions were significant, $Fs < 1$.

The overall error rate was 3.9%. An analysis of error rates indicated significant main effects corresponding to the main effects found in the response time data. In addition, however, there was a significant interaction between frequency and priming that did not appear in the response time data, $F(1, 71) = 7.59$, $p < .01$. The pattern of means indicated a larger benefit of related primes for low-frequency targets (2.4% effect) than for high-frequency targets (0.7% effect). Mean response times for nonwords was 693 ms in the clear condition and 747 ms in the degraded condition, and the
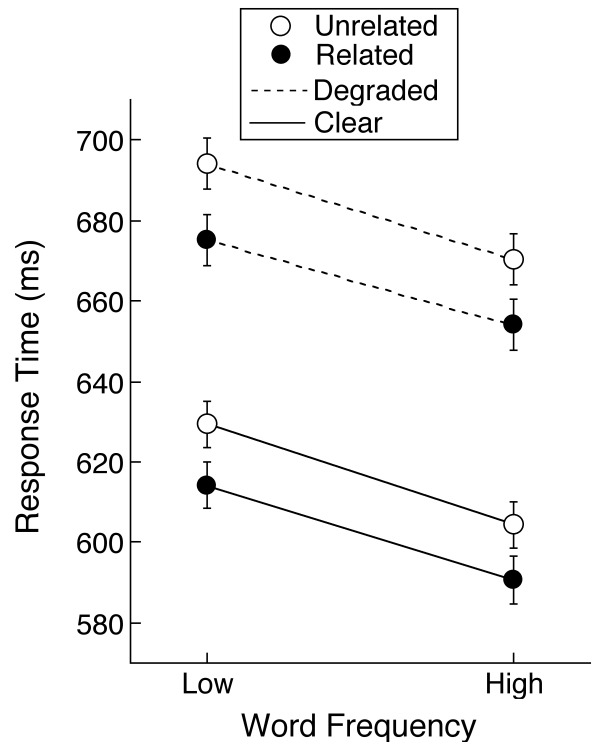

Figure 2. Mean response time in Experiment 1 as a function of word frequency, prime, and stimulus quality. Error bars are 95% within-subject confidence intervals appropriate for comparing condition means within a particular stimulus quality condition (Loftus & Masson, 1994; Masson & Loftus, 2003).

mean error rate was 5.0%.

The additive joint effects of stimulus quality with frequency and with priming were anticipated, given prior results, and particularly previous studies that have used low-strength associates as primes (Stolz & Neely, 1995). Although the overadditive interaction of priming and frequency did not appear in response latencies, it was present in the error data. This outcome could reflect a speed-accuracy trade-off that prevented the interaction from emerging in the response time data.

Linear mixed-model analysis. The critical question we address with the next analysis is whether the additive effects involving stimulus quality that are apparent in the aggregate data may be masking interactive effects that are modulated by recent trial history. In particular, we examined the influence of two characteristics of the immediately preceding trial on the current trial: whether the previous trial required a word or nonword response and the stimulus quality of the target on the previous trial.

(a) Dependent variable. Estimates of LMM variance and covariance parameters critically depend on residuals being normally distributed. Therefore, in a reanalysis of masked priming data using LMMs, Kliegl et al. (2010) considered a number of methods of

transforming raw response time data to produce a measure for which model residuals fit a normal distribution. They found that for lexical decision tasks, a reciprocal transformation applied to raw response time data yielded residuals that much more closely approximated a normal distribution than untransformed or log-transformed response times. Kinoshita et al. (2011) also found the reciprocal transformation to be the most appropriate one for their lexical decision data. Consequently, we examined the reciprocal transformation for the data reported here and found again that it produced residuals with a good approximation to a normal distribution. The analyses we report, therefore, are based on a reciprocal transformation of the raw data, specifically -1/RT (where RT refers to response time measured in seconds) so that higher scores will continue to correspond to longer response times. Further, we note that over the last few years we have analyzed data from well over 20 experiments using the lexical decision task. Without exception, the reciprocal transformation yielded residuals that were in much better agreement with normal distribution assumptions than were raw or log-transformed response times.[1] Note that without the multiplication by (-1), reciprocal RT corresponds to a perfectly valid metric in physics, namely, speed. Perhaps lexical decision effects will prove to be easier to interpret in general when conceptualized as differences in response speed rather than as differences in response time.

To examine possible interaction and additive effects between priming, word frequency, and stimulus quality, the LMM analyses included these three manipulated variables as fixed effects, along with the main effects of stimulus quality on the previous trial and lexical status of the target on the previous trial. In addition to these five main effects, all of the interactions (from two-way interactions up to the five-way) were included.

(b) Significance criterion. We took as the criterion for significance of fixed effects a *t* ratio of 2.0, following Kliegl et al. (2010). In LMM, the degrees of freedom for *t* ratios are not known exactly, although with the very large number of observations in data sets such as the ones we report here, the *t* distribution converges to the normal distribution. Therefore, using two standard errors as a significance criterion corresponds closely to the .05 level of significance. Before we turn to details about fixed-effect estimates, we describe how we determined statistically reliable

variance components and correlation parameters associated with them.

(c) Variance components and correlation parameters (random effects). Along with fixed effects, our LMM allows the estimation of two sets of variance components and correlation parameters for the random factors of subject and word. In principle and ignoring interactions between fixed effects, the design afforded the following parameters for the two random factors. For the subject factor, there were 6 variance components (mean RT [intercept] plus 5 within-subject effects) and 15 correlation parameters for the possible correlations between each pair of these 6 components. For the random factor word, there were 5 variance components (mean RT [intercept] plus 4 within-word effects; frequency is a between-word effect) and 10 correlation parameters for correlations between these components. Counting also the residual variance, the model parameters added up to a maximum of 37 variance components and correlation parameters. These are simply too many parameters given our data, so we proceeded in two steps to determine significant parameters. In the first step, we determined the significant variance components by deleting in turn each variance component from a model containing all variance components and checking for a significant decrement in goodness of fit using likelihood ratio tests, the Akaike Information Criterion, and the Bayesian Information Criterion. Models with successively fewer variance components were tested until it was determined that dropping a particular variance component produced a significantly worse fit to the data (see Baayen et al., 2008, and Quené & van den Bergh, 2008, for detailed examples of how this procedure is applied to LMMs).

(d) LMM results for variance components and correlation parameters. The final model included variance components for the mean speeds for item and subjects (i.e., the intercepts), a variance component for the priming effect for items (indicating that priming was differentially successful across items), and variance components for the effects of stimulus quality on the current trial and lexical status of the target on the previous trial for subjects. In addition, the final model included a correlation parameter for the random effects of stimulus quality and of intercept for subjects. This correlation was -.41, indicating that subjects who responded faster generally showed larger effects of stimulus quality. This somewhat counterintuitive outcome probably reflects greater reliability in the measurement of stimulus quality effects for subjects who respond faster (and probably with less variability). These random effects are summarized in the upper section of Table 1.

Note that the set of random components retained in the best fitting LMM does not include a component for variation between subjects with respect to the effect of

---

[1]We repeated the LMM analyses reported below using response time instead of the reciprocal transformation and found essentially the same results. This was true for both Experiment 1 and Experiment 2. A few interactions that were significant with the reciprocal measure were not significant in the response time analysis, probably because of lower statistical power due to heterogeneity of residuals.

Table 1

Linear Mixed Model Estimates of Coefficients, Standard Errors, and t Ratios for Fixed Effects, and Variances, Standard Deviations, and Correlation for Random Effects in Experiment 1

Random effects	Variance	St. dev.	r
Items			
Intercept	0.0044	0.0664	
Prime	0.0020	0.0451	
Subjects			
Intercept	0.0261	0.1617	
Stimulus quality	0.0042	0.0650	-0.409*
Last-Trial Target	0.0019	0.0435	

Fixed effects	Coefficient	St. error	t
Intercept	-1.6376	0.0197	-83.29
Frequency (F)	0.0546	0.0097	5.64
Prime (P)	0.0384	0.0054	7.17
Stimulus quality (Q)	0.1647	0.0089	18.54
Last-trial stimulus quality (LQ)	0.0023	0.0045	0.50
Last-trial target (LT)	0.0250	0.0069	3.64
F x P	-0.0033	0.0107	-0.31
F x Q	-0.0131	0.0090	-1.46
P x Q	0.0050	0.0090	0.55
F x P x Q	-0.0053	0.0180	-0.29
F x LQ	0.0012	0.0091	0.13
F x LT	0.0118	0.0091	1.30
P x LQ	-0.0146	0.0091	-1.60
P x LT	-0.0037	0.0091	-0.40
Q x LQ	-0.0344	0.0091	-3.78
Q x LT	0.0085	0.0091	0.94
LT x LQ	0.0141	0.0091	1.55
F x P x LQ	0.0446	0.0182	2.45
F x P x LT	-0.0032	0.0182	-0.18
F x Q x LQ	0.0433	0.0182	2.38
F x Q x LT	0.0010	0.0182	0.06
P x Q x LQ	0.0195	0.0182	1.07
P x Q x LT	-0.0246	0.0182	-1.35
F x LQ x LT	-0.0137	0.0182	-0.75
P x LQ x LT	-0.0113	0.0182	-0.62
Q x LQ x LT	0.0924	0.0182	5.08
F x P x Q x LQ	-0.0076	0.0364	-0.21
F x P x Q x LT	0.0094	0.0364	0.26
F x P x LQ x LT	0.1089	0.0364	2.99
F x Q x LQ x LT	0.0059	0.0365	0.16
P x Q x LQ x LT	0.0405	0.0364	1.11
F x P x Q x LQ x LT	0.0598	0.0729	0.82

*Correlation between intercept and stimulus quality.

semantic priming, even though there were components for variation between subjects in the effects of stimulus quality on the current trial and lexical status of the previous trial's target. The lack of evidence for a random effect of priming among subjects, suggesting low reliability in our measurement of priming variability between subjects, is consistent with the finding of Stolz et al. (2005) that semantic priming effects have very low reliability. If there were stable, meaningful variationbetween subjects with respect to the magnitude of semantic priming, then the fit of the LMM should have been improved by including a random subjects effect for priming.

(e) LMM results for fixed-effects. The estimated coefficients for the fixed effects of the final model are shown in the lower part of Table 1. The fixed effects of the three factors, priming, frequency, and stimulus quality, were significant and we observed the same pattern of additivity among those factors as was seen in the ANOVA. In addition, however, the lexical status of the target on the previous trial influenced response time as a main effect: there was slower responding if the previous target was a nonword. Moreover, both lexical status and stimulus quality of the previous trial's target entered into interactions with the three primary factors.[2]

For example, the effect of current stimulus quality was modulated by both the prior target's lexical status and its stimulus quality, as indicated by the interaction between prior and current stimulus quality and the three-way interaction between these two factors and lexical status of the prior target. The three-way interaction is plotted in Figure 3. The plot shows that responding was faster if the stimulus quality on the previous and current trials was the same rather than different, but this effect held only when the previous target was a word. The difficult processing encountered when the previous target was degraded did not generally slow response time on the current trial, as might be expected from the Kinoshita et al. (2008) ASE model, but instead its influence depended on the quality of the current trial's target. This result is consistent with the implications of the Turner et al. (2011) proposal regarding how stimulus representations are learned and are influenced by recent trial history. Two degraded word targets can be considered relatively close on the stimulus strength axis (as compared, for example, to a degraded word target and a degraded nonword target), and a signal (word) response to one of these two word targets may be expected on the Turner et al. account to elevate the signal response strength of other nearby targets, leading to a faster response.

[2]The occurrence of a response error on a trial immediately preceding one of the critical trials might introduce some extraneous influence on our measure of the effects of trial history. To test this possibility, we repeated the LMM analyses for both Experiment 1 and Experiment 2, omitting critical trials that were preceded by an error. The results in both cases were consistent with those we report for data that were not filtered for an error on trial N - 1.
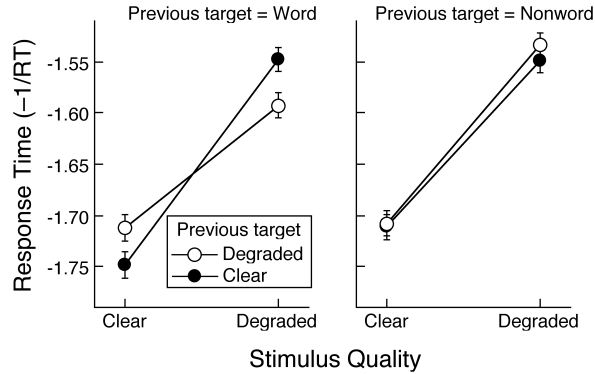
Figure 3. Mean transformed response time in Experiment 1 as a function of stimulus quality of the target, and stimulus quality and lexical status of the target on the previous trial. Error bars are estimated 95% within-subject confidence intervals appropriate for comparing condition means.

More importantly, apparent additive effects involving frequency and the other two primary factors dissolved into interactive relationships when data were conditionalized on features of the previous trial's target. Additivity between frequency and priming turned out to be composed of two, opposite going interactions that depended on the quality of the previous target. This dependency was revealed by an interaction between these two factors and quality of the previous target, and by a four-way interaction that included lexical status of the previous target. The pattern of this four-way interaction is shown in Figure 4. Although priming and frequency were additive when the previous target was a word, interactions emerged when it was a nonword. Specifically, a typical overadditive interaction between priming and frequency occurred when the previous nonword target was degraded (bottom-right panel), but an underadditive interaction was seen when the previous nonword target was clear (top-right panel). A re-parameterization of the LMM with frequency and priming nested within the four cells defined by the stimulus quality and lexical status of the target on the previous trial showed that both of these two-way interactions were significant (*t*s > 2.3). There was no significant interaction when the previous target was a word (*t*s < 1). Combining the data across all trials, ignoring trial history, led to a pseudo-additive pattern in the aggregate. The overadditive component of this interaction (seen only when the previous target was a degraded nonword) is consistent with past literature (e.g., Becker, 1979; Borowsky & Besner, 1993), but the underadditive outcome is quite surprising.

Although we anticipated the possibility that opposite going interactions might underlie an additive effect, there is little or no theoretical guidance for predicting what form these interactions might take. A speculative possibility is that the process in the Plaut and Booth (2000) model that relates input and output activation (see Figure 1C) may be sensitive to recent trial events. For example, a particularly difficult stimulus, such as a degraded nonword, may lead to a increase in the amount of evidence required to produce a response on the next trial. This is one of the principles incorporated into the ASE model of dynamic changes in response times (Kinoshita et al., 2011). Requiring more evidence (i.e., a higher level of output activation for a response) would move the criterion for responding further up the sigmoid function in Figure 1C, into a region that would produce an overadditive interaction, as depicted in the figure. A less demanding experience on trial *N* - 1 might allow the output activation threshold to be lowered, moving the criterion back down the sigmoid function. Shifting to a sufficiently low level would create the potential for an underadditive interaction, which was observed when the target on the preceding trial was a clear nonword. Although this scenario fits with the case where the previous target was a nonword, no such dynamic process was evidence when the target on trial *N* - 1 was a word. We consider below the question of why lexical status of the previous target might modulate these dynamics.

The additive relationship between frequency and stimulus quality, a result that is central to the debate about the ability of computational models to capture additive effects (e.g., Borowsky & Besner, 2006; Plaut
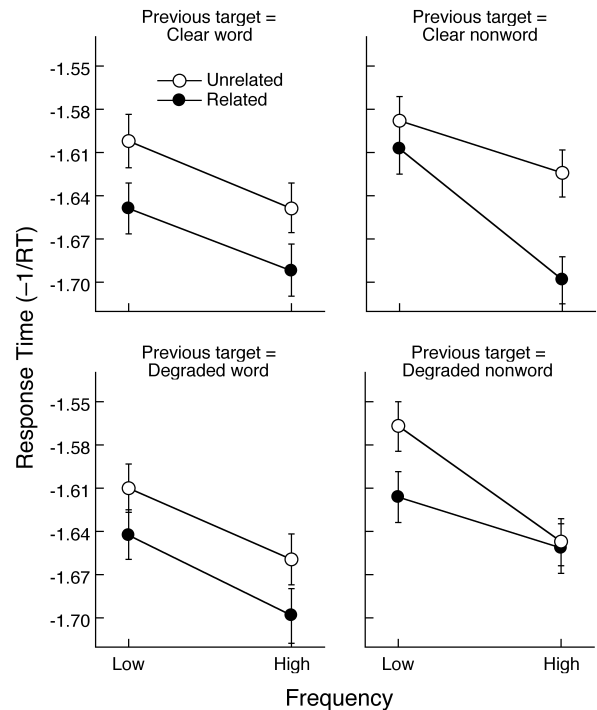

Figure 4. Mean transformed response time in Experiment 1 as a function of frequency and prime condition for the current target, and stimulus quality and lexical status of the target on the previous trial. Error bars are estimated 95% within-subject confidence intervals appropriate for comparing condition means.
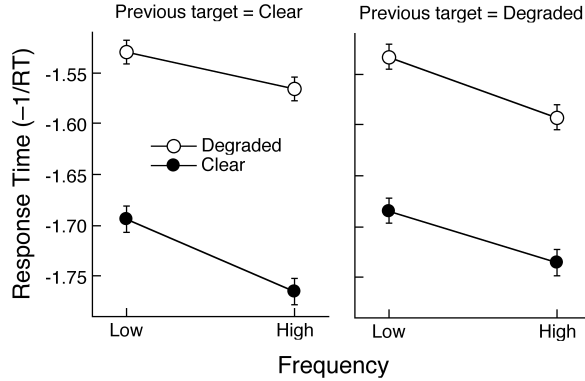
Figure 5. Mean transformed response time in Experiment 1 as a function of frequency and stimulus quality for the current target, and stimulus quality of the target on the previous trial. Error bars are estimated 95% within-subject confidence intervals appropriate for comparing condition means.

& Booth, 2006), turned out to be the product of aggregating two, opposite-going interaction effects, as indicated by the three-way interaction between frequency, stimulus quality, and stimulus quality of the previous target. This interaction is shown in Figure 5, and comprises a significant underadditive interaction when the previous target was clear (left panel; $t = 2.71$) and a non-significant, but numerically overadditive interaction between frequency and stimulus quality when the target on the previous trial was degraded (right panel; $t = 0.67$), as indicated by a re-parameterization of the LMM with frequency and stimulus quality nested within levels of last-trial stimulus quality.

The underadditive interaction between frequency and stimulus quality obtained when the previous trial's target was clear (left panel of Figure 5), combined with the underadditive interaction between frequency and stimulus quality reported by Yap et al. (2008), raises the distinct possibility that the underadditivity produced in some simulations with the Plaut and Booth (2000) model (e.g., Besner et al., 2008) might be a valid reflection of behavior. As with the combination of over- and underadditivity seen with frequency and priming, the modulation of the frequency by stimulus quality interaction could, in principle, be captured in a model like the Plaut and Booth connectionist model. With a difficult (degraded) target on the previous trial, the response threshold might be elevated, requiring stronger output for a response. This setting would move the system to a higher point on the sigmoid function shown in Figure 1C, near a region in which overadditivity is seen. When the prior target was less demanding (clear), the threshold might be reduced, bringing the system into the region of the activation function that yields underadditivity.

(f) Including Trial as a covariate in the LMM. The results described in the previous section clearly

document that characteristics of the previous trial (i.e., the lexical status and the stimulus quality of the last target) exert strong effects on response speed in the current trial. According to Turner et al. (2011), subjects may develop stimulus representations over time and this learning increases the ability to discriminate between stimulus classes–words and nonwords. We investigated the possibility that these effects would change across trials in the experiment. In a second LMM, we added trial number (centered, so that the middle trial was coded as zero) and its interactions as a sixth factor to the primary model described above to assess condition-specific change in response time over the course of the entire experiment. This analysis indicated that trial number interacted with lexical status of the target on the previous trial. Response time to word targets did decrease over trials, but only if the previous trial's target was a word (see Figure 6). The t ratio for the effect of trial number when the previous trial had a word target was substantial, $t = -6.26$, whereas the effect was clearly not significant when the immediately preceding target was a nonword, $t < 1$. None of the other interactions with trial were significant. These data indicate that the tendency to respond more efficiently to word targets as subjects gained more experience with the task and with the nature of the stimuli was strongly modulated by recent trial history.

The sensitivity of improvement across trials to the lexical status of the previous trial's target offers possible insight into the four-way interaction involving frequency and priming shown in Figure 4. That interaction revealed that the relationship between frequency and priming was modulated by the stimulus quality of the previous target, but only if that target was
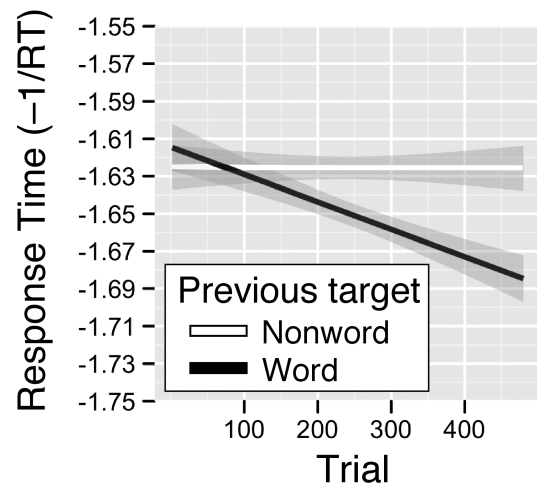

Figure 6. Mean transformed response time in Experiment 1 as a function of trial number and lexical status of the target on the previous trial. Continuous error bars (shown as gray bands around trend lines) are 95% within-subject confidence intervals appropriate for comparing condition means.

a nonword. This finding, combined with the trial effects shown in Figure 6, suggests that the application of dynamic adjustments to the word recognition system, such as altering response thresholds, may be modulated by ongoing events. In particular, with consecutive presentation of word targets, the system appears to become increasingly efficient. In contrast, when a nonword target appears, the efficiency gains evident following a word target appear to be lost, a phenomenon that we suggest is comparable to the increased caution that contributes to post-error slowing (e.g., Allain et al. 2009; Dutilh et al., 2012; Rabbitt, 1966, 1989). We suggest that the presentation of a nonword target may interrupt the relatively smooth processing that develops when consecutive word targets are experienced, and that adjustments to decision processes are made to reduce the possibility of an error. A similar principle is used to guide response initiation processes in the ASE model (Kinoshita et al., 2008, 2011).

Experiment 2

In Experiment 1, we found that recent trial history, embodied by the characteristics of the target on the immediately preceding trial, altered something as fundamental as the improvement in performance across trials. Moreover, the stimulus quality of the previous trial's target modulated the relationship between frequency and the other two primary factors, priming and stimulus quality. Such a complex constellation of results requires a follow-up to assess its reliability and generalizability. Therefore, in Experiment 2, we investigated the question of whether this modulation was the result of an ineluctable characteristic of the word processing architecture, or instead was the product of dynamic adaptation to varying trial difficulty, in line with models such as the ASE (Kinoshita et al., 2008, 2011) or the Turner et al. (2011) account of learning stimulus representations. Rather than randomly mixing clear and degraded targets, subjects in Experiment 2 experienced these two types of display in separate blocks of trials. With this arrangement, subjects would not be exposed to trial by trial variation in processing difficulty arising from stimulus degradation, although there would still be potential variation due to lexical status of the target on the previous trial. Finally, we also aimed to demonstrate reliable individual differences in semantic priming effects, given a task context without random trial by trial fluctuation in stimulus quality.

Method

Subjects. Seventy-two new subjects drawn from the same source as in Experiment 1 participated in the experiment.

Materials and procedure. The same materials and procedure were used as in Experiment 1, except that clear and degraded targets were presented in two separate blocks, each consisting of 16 practice trials and 240 critical trials. The order of presentation of the two blocks was counterbalanced across subjects.

Results and Discussion

Analysis of variance. Response times were filtered as in Experiment 1, with the upper bound set at 1,600 ms so as to exclude no more than 0.5% of the correct responses as outliers. Mean correct response times are shown in Figure 7. An ANOVA with frequency, prime condition, and stimulus quality as repeated measures factors indicated that all three main effects were significant, $Fs(1, 71) > 35$, $ps < .001$. In addition, there was an overadditive interaction between frequency and prime, indicating larger priming for low-frequency targets, $F(1, 71) = 7.10$, $p < .01$. None of the remaining interactions were significant. Although the pattern of means in Figure 7 suggests that the frequency x prime interaction was restricted to degraded targets, implying a three-way interaction, that interaction was not significant, $F(1, 71) = 2.08$, $p > .15$.

The overall error rate was 3.1%. An ANOVA applied to these data found significant main effects of frequency and prime, $Fs(1, 71) > 13$, $ps < .01$, but the
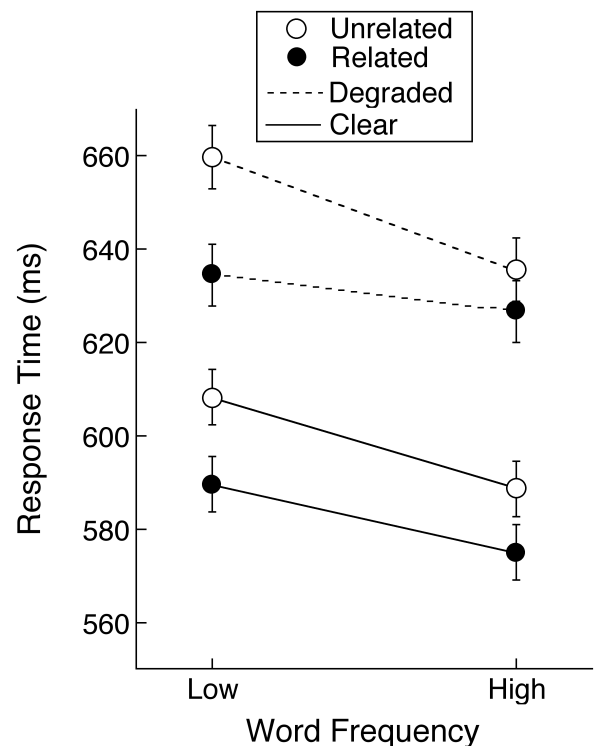

Figure 7. Mean response time in Experiment 2 as a function of word frequency, prime, and stimulus quality. Error bars are 95% within-subject confidence intervals appropriate for comparing condition means within a particular stimulus quality condition.

stimulus quality effect was not significant, *F*(1, 71) =3.02, *p* = .09. There was one significant interaction which was an underadditive pattern involving prime and stimulus quality, whereby the priming effect was larger for clear targets (1.5% vs. 0.5%), *F*(1, 71) = 5.72, *p* < .05. Mean response time for nonword targets was 672 ms for the clear condition and 718 ms for the degraded condition and the overall error rate was 4.3%.

In summary, the response time data replicated the typical interaction between frequency and prime. Even though there was an additive relationship between prime and stimulus quality in response times, the error data present a complication in the form of an underadditive interaction between these two factors. The interaction could reflect a ceiling that subjects placed on their willingness to make response errors, so that the error rate in the most susceptible condition (unrelated prime/degraded target) did not reach what otherwise might have been its full extent.   This suggestion is similar to one of the accounts of the underadditive interaction between frequency and stimulus quality that Yap et al. (2008) offered for their finding with response time data. But if this were so, then it implies that going beyond the imposed error rate ceiling would be associated with faster responding in the unrelated/degraded condition. That change would, in turn, potentially generate an underadditive interaction in the response time measure. The interpretation of the prime by stimulus quality interaction in error rates, then, is unclear. Given that the error rates were very low overall, with many cells (43%) having zero errors, the interaction may be a type I error.

Linear mixed-model analysis.   The LMM analysis was applied to response data after they were converted using the reciprocal transformation (-1/RT), as in Experiment 1. Stimulus quality of the previous trial's target was not included as a factor in this analysis because that factor was a blocked manipulation and stimulus quality did not vary from trial to trial within a block. All of the other fixed effects and interactions included in the LMM analysis reported in Experiment 1 were included in the present analysis. Using the same method as in Experiment 1, we determined which variance components warranted inclusion in the final model, based on adequacy of the fit to the data.

(a) LMM results for variance components and correlation parameters.   In the final model, the retained variance components were the item and subject intercepts, the effect of prime for items, and the effects of stimulus quality and lexical status of the target on the previous trial for subjects. All of these random effects were also included in the optimal model for Experiment 1. In addition, the model for Experiment 2 included random effects of frequency and prime for subjects, as well as a parameter for the correlation between prime and intercept for subjects. Details of the random effects

Table 2
Linear Mixed Model Estimates of Coefficients, Standard Errors, and t Ratios for Fixed Effects, and Variances, Standard Deviations, and Correlation for Random Effects in Experiment 2

Random effects	Variance	St. dev.	*r*
Items			
Intercept	0.0042	0.0651	
Prime	0.0014	0.0371	
Subjects			
Intercept	0.0262	0.1617	
Frequency	0.0011	0.0328	
Prime	0.0008	0.0280	-0.507*
Stimulus quality	0.0073	0.0853	
Last-Trial Target	0.0038	0.0619	

Fixed effect	Coefficient	St. error	*t*
Intercept	-1.7011	0.0194	-86.60
Frequency (F)	0.0422	0.0103	4.12
Prime (P)	0.0445	0.0060	7.41
Stimulus quality (Q)	0.1375	0.0110	12.54
Last-Trial Target (LT)	0.0531	0.0085	6.22
F x P	0.0204	0.0100	2.04
F x Q	-0.0122	0.0088	-1.39
P x Q	0.0019	0.0088	0.21
F x P x Q	0.0166	0.0176	0.95
F x LT	-0.0008	0.0089	-0.09
P x LT	-0.0014	0.0089	-0.15
Q x LT	0.0013	0.0089	0.15
F x P x LT	-0.0148	0.0178	-0.83
F x Q x LT	-0.0183	0.0178	-1.03
P x Q x LT	-0.0230	0.0178	-1.30
F x P x Q x LT	-0.0770	0.0355	-2.17

*Correlation between intercept and prime.

are shown in the upper section of Table 2. The estimate of the correlation parameter (-.51) indicates that the effect of priming was larger for subjects who responded faster. As with the correlation between stimulus quality and response speed in Experiment 1, we suggest that the reliability of the data is greater for faster subjects.

Although the variance component for the random effect of priming for subjects was quite small (indeed it was the smallest random effect obtained in either experiment), it significantly improved the fit of the LMM. These results indicate that with a sufficiently sensitive measure of individual differences, evidence can be obtained for stable sources of variation in the magnitude of semantic priming. We note that with a prime-target SOA of 200 ms and a prime relatedness proportion of .5 (as was used here), Stolz et al. (2005) found a low, but significant split-half reliability estimate for semantic priming. In addition, the signifi-
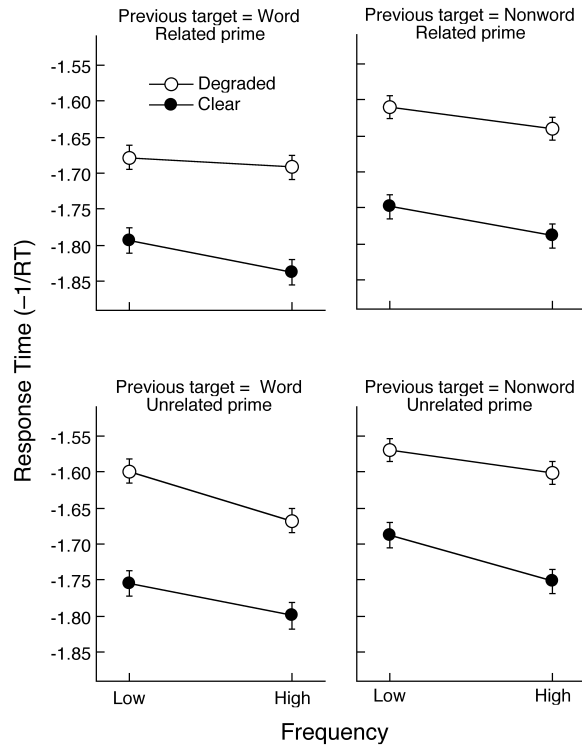
Figure 8. Mean transformed response time in Experiment 2 as a function of frequency, prime condition, and stimulus quality for the current target, and lexical status of the target on the previous trial. Error bars are estimated 95% within-subject confidence intervals appropriate for comparing condition means.

cant correlation we obtained between priming effect size and overall speed of responding indicates that there is sufficient reliability in measures of semantic priming to establish relationships with other variables. This finding, in turn, encourages further examination of possible mechanisms underlying individual differences in semantic priming.

(b) LMM results for fixed-effects. The estimated coefficients for the fixed effects of the optimal model in Experiment 2 are shown in lower section of Table 2. The *t* ratios shown in the table indicate that along with the significant main effects of all three primary factors and the effect of the previous target's lexical status, there was an interaction between frequency and prime, matching the overadditive interaction seen in the ANOVA and in Figure 7 reported earlier. In addition, however, the four-way interaction was significant. The pattern of means in this interaction is shown in Figure 8 where the means are plotted to highlight the potential interaction patterns between frequency and stimulus quality. The existence of a four-way interaction implies that these two factors are not consistently additive, contrary to what is suggested by the analysis of aggregated data reported above.

Indeed, it can be seen in Figure 8 that there are two

contexts (word target on previous trial, related prime on current trial; nonword target on previous trial, unrelated prime on current trial) where there appears to be an underadditive interaction between word frequency and stimulus quality. Combining these two contexts to examine that interaction yielded a significant effect, $t = -2.51$. For the other two contexts, no significant interaction between word frequency and stimulus quality was obtained, neither when the contexts were considered separately nor when combined. The weak tendency toward an overadditive interaction when the prior target was a word and the current trial's prime was unrelated ($t = 1.38$) appears to have helped to prevent an overall frequency by stimulus quality interaction from emerging.

The underadditive interactions that appear in two of the panels of Figure 8 occurred under circumstances that represent particularly fast and particularly slow responding. This pattern is difficult to explain, even with a flexible framework such as the Plaut and Booth (2000) model. For the case in which the current target has a related prime and the previous target was a word, which is where the fastest responses were made, one could imagine that the required level of output activation for a response might be relatively low, which means the system is operating in a region of the input-output activation function that is amenable to an underadditive interaction. But the other circumstance in which underadditivity was found, prior nonword target and unrelated prime, is associated with slow responses. That feature suggests that relatively more evidence is required for responding, which would presumably move the system further along the input-output activation function, out of the region where an underadditive interaction would be expected. We have no compelling explanation for this somewhat anomalous outcome, but one possibility is that it is the product of a ceiling effect on response time in the slowest condition (low frequency, degraded target). As can be seen in Figure 8, response times were longer in this condition than in any other.

(c) Including Trial as a covariate in the LMM. As was done in Experiment 1, we examined the change in response time across trials by including trial number as a covariate along with its interactions with the other independent variables. Trial number ran from the first trial of the first block (clear targets for half of the subjects, degraded targets for the other half) to the last trial of the second block. As in Experiment 1, the lexical status of the target on the previous trial strongly modulated the improvement in response time across trials. This result is illustrated in Figure 9. Note that although response time decreased over trials when the previous target was a word, $t = -2.07$, the opposite trend held when the previous target was a nonword, $t = 2.07$. This analysis was repeated with stimulus quality of the current target as a factor, and trials running only from
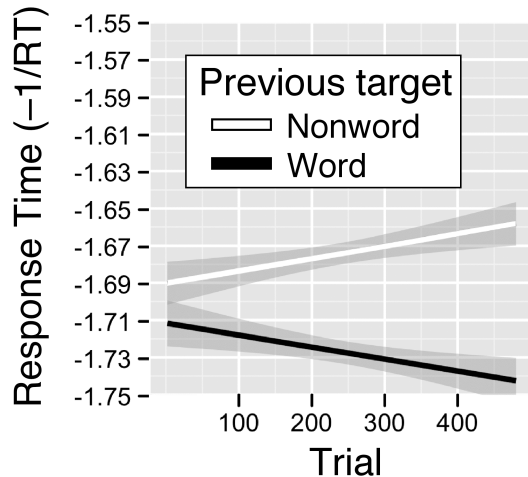
Figure 9. Mean transformed response time in Experiment 2 as a function of trial number and lexical status of the target on the previous trial. Continuous error bars (shown as gray bands around trend lines) are 95% within-subject confidence intervals appropriate for comparing condition means.

the first to the last trial within a block (i.e., 1 to 240), given that stimulus quality was a blocked factor in this experiment. The same pattern of change across trials was seen for both levels of stimulus quality, with no interaction between stimulus quality, trial, and lexical status of the previous target. The improvement seen when the previous target was a word is consistent with the results of Experiment 1, but the slowing over trials when the earlier target was a nonword is novel and surprising. It suggests that even though an error is not committed, responding positively to a word target may be done with elevated caution because of the heightened sensitivity to the possibility of an error occasioned by handling a nonword target. Given that this result did not appear in Experiment 1, however, it would be prudent to wait for replications of this result before drawing strong conclusions as to its meaning.

General Discussion

The primary result obtained in the two experiments reported here is the contrast between apparently additive effects obtained in aggregated data and the underlying interactions (both over- and underadditive) between these same effects when the data were examined with respect to recent trial history. In both Experiments 1 and 2, aggregate data indicated that the effects of word frequency and stimulus quality were additive, in keeping with previous findings (e.g., Becker & Killion, 1977; Stanners, Jastrzembski, & Westbrook, 1975; Yap & Balota, 2007). But the LMM analyses revealed that in both cases, these additive patterns were generated by underlying interactive effects that were modulated by recent trial history.

Variations in the pattern of interaction between word frequency and both semantic priming and stimulus quality were demonstrated.

We have suggested that these modulations potentially can be captured in a model such as the Plaut and Booth (2000) connectionist model of word recognition by assuming that experience on a particular trial can influence the evidence threshold required for responding on the next trial, similar to what is assumed in the ASE model of dynamic adjustment of response initiation (Kinoshita et al., 2008, 2011). Changes in threshold correspond to changes in degree of output activation needed for responding. As shown in Figure 1C, the sigmoid function relating input to output activation can generate additive, overadditive, or underadditive relationships between factors, depending on the region of the activation function that maps onto the current level of required evidence (output activation). According to the Besner et al. (2008) simulation results using the Plaut and Booth (2000) model, an underadditive interaction between frequency and stimulus quality was obtained when a relatively weak manipulation of quality was applied, but additivity and then an overadditive interaction were obtained in that simulation as the manipulation was amplified. In our experiments, however, the strength of the stimulus quality manipulation was constant and therefore cannot be mapped onto the conditions under which we obtained additivity and interactions. It was not the nature of the manipulation of stimulus quality that modulated the relationships between factors that we observed, but instead recent trial history. One of the challenges, then, in adapting a model such as that of Plaut and Booth (2000) to accommodate results like those reported here is to develop a principled account of dynamic responses to recent trial events and of how the application of cognitive control over these responses is regulated (e.g., Botvinick, Braver, Barch, Carter, & Cohen, 2001).

Another challenge is to determine how a computational model might accommodate the degree of processing flexibility implied by the dynamic changes in relationships between variables seen in our experiments. For example, the four-way interaction depicted in Figure 4 shows all three possible relationships between word frequency and semantic priming (additivity, overadditivity, and underadditivity). Although the sigmoid function shown in Figure 1C can generate all three patterns, there is a problem. Namely, the two types of trial history that produced the two interaction effects in Figure 4 both involve relatively long response times. The conditions yielding additivity have shorter response times. On a simple interpretation of the input-output activation function in Figure 1C, the average response time in the conditions that produce additivity should lie between the response times for the over- and underadditive
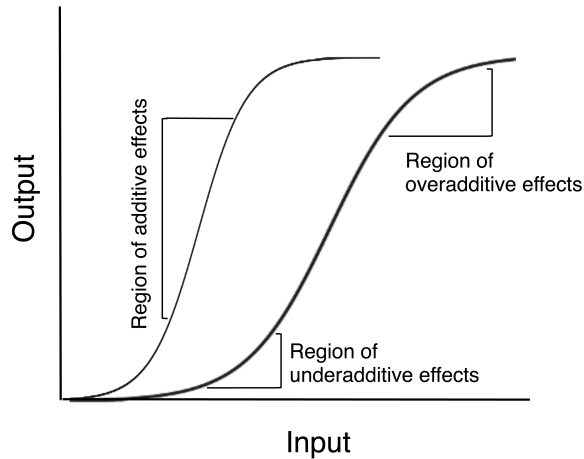
Figure 10. Schematic representation of an adjustable input-output activation function. The variant on the left is invoked when the previous trial's target was a word and the version on the right is invoked when the previous target was a nonword. In the latter case, more input activation (and time) is required to achieve a particular amount of output activation relative to the version of the function on the left. Additive effects (e.g., of frequency and priming) can be generated by the left function under response times that are shorter (an input values that are lower) than those for the right function when it generates under- and over-additive effects.

conditions. Clearly, this is not what we found. Another component would have to be added to the model to accommodate the set of results seen in Figure 4. As one example, we suggest that the input-ouput activation function may itself be dynamic, such that the slope of the function might become more shallow in cases where caution is warranted or processing is more challenging. This notion is depicted in Figure 10, which embodies the assumption of a slower-rising activation function when the previous trial presented a nonword target. On those trials, more input strength is required to produce a response (presumably increasing overall response time). As shown in the figure, this modification creates a system in which what is experienced on the previous trial can lead to different interaction patterns.

Trial-to-Trial Influences on Practice Effects

A particularly striking feature of our results is the demonstration that something as fundamental as improvement with practice across trials is strongly modulated by differences between successive trials. In both experiments, we found that improvements in speeded lexical decisions to word targets were seen only when the previous trial's target was also a word. When the previous target was a nonword, either no improvement (Exp. 1) or even an elevation in response time (Exp. 2) was seen as the experiment progressed. These findings raise interesting questions about the mechanisms underlying skill learning and models of those processes. Improvement with practice generally

follows a power law, with the logarithm of response time decreasing as a linear function of the logarithm of practice trial number (A. Newell & Rosenbloom, 1981).

Our results indicate that this lawful relationship may not hold and might even be reversed, depending on what occurred on the previous trial. It would be of substantial interest to determine which components of skill acquisition underlie such contingencies. Moreover, Logan (1988) developed a detailed mathematical model of instance-specific learning that he applied to the lexical decision task. The model provided an account of the power function speed-up when responding to repeatedly presented items by assuming that subjects relied on memory for prior instances involving those items. Both word and nonword items showed these systematic effects, but our findings raise the possibility that item-specific learning functions may be modulated by the nature of the previous trial's target. If so, then models of memory-based skill acquisition (e.g., Logan, 1988; Rickard, 2004) would have to be revised.

Implications for Processing Architecture

At present, we lack explanatory frameworks powerful enough to generate a priori expectations about how trial history should influence interactive effects of manipulations such as frequency, stimulus quality, and priming. The current approach to explaining patterns of additivity and interaction has been dominated by attempts to choose between different architectural assumptions about processing modules. For example, in the Plaut and Booth (2000, 2006) connectionist model, changes in interaction patterns can be attributed to variation in activation levels along the sigmoid function relating input to output of processing units. In the additive factors approach championed by Besner and colleagues (e.g., Besner et al., 2008, 2010; Yap et al., 2008), flexibility in the application of a thresholding or normalization mechanism when processing degraded input accounts for observed relationships between factors that may or may not be additive.

It is unclear, however, whether dynamic adjustments to processing architecture will be the most fruitful approach to explaining the modulation of additivity and interaction between factors. Indeed, serious questions have been raised about whether we can empirically distinguish between predictions of discrete stage models and continuous processing models (McClelland, 1979; Stafford & Gurney, 2011; Thomas, 2006). An alternative direction endorsed by the present results is grounded in an examination of the influence of trial history on processing dynamics. Models like the ASE (Kinohsita et al., 2008, 2011), that provide coherent accounts of how recent trial events can influence decisional processes, may turn out to be especially useful. In approaches such as these, the critical issue is not dynamic changes in architectural

operation, but instead modulation of response initiation processes. In the ASE, recent trial history provides an estimate of how rapidly information about a stimulus can be assumed to accumulate and therefore about approximately when it would be safe to make a speeded response without too much risk of an error. In this model, there are conditions under which easy items are more sensitive to trial history than are hard items (the opposite relationship is not predicted by the model). The underadditive interaction between frequency and stimulus quality in Experiment 1, resting as it does on modulation of high-frequency word targets in particular, may reflect this differential sensitivity.

A related approach to accounting for patterns of additivity and interaction emphasizes the nature of stimulus representations that enter into decision processes. As we have discussed, Turner et al. (2011) have proposed that changes in task performance across trials, including the influence of recent trial history, are determined in part by the development of stimulus representations. We suggest that differential learning about stimuli, potentially as a function of any of the three major factors considered here, can contribute to the nature of the relationship observed between those factors without appeal to any modulation of processing architecture. Although the Turner et al. model has been developed to account for performance accuracy, not response time, they point out that other models exist that account for how response latency is affected by stimulus representations that change over time (e.g., Lee & Dry, 2006; B. R. Newell & Lee, 2011).

A specific example of how finding additivity or interaction in response time data may depend on the way in which stimuli are coded was demonstrated in modeling work reported by Stafford and Gurney (2011). They simulated the influence of color saturation on the congruency effect in single-stage and discrete-stage variants of a model of the standard color-word Stroop effect (Cohen, Dunbar, & McClelland, 1990). Their results showed that a single-stage model, with no separate, thresholded processing stage to delay input to a decision stage for the sake of cleaning up perceptual input, produced additivity between color saturation and congruency, as seen in behavioral data (Stafford, Ingram, & Gurney, 2011). The additive effect was achieved by binding the color and word inputs submitted to the model. That is, the intensity of the word representation provided to the model was tied to the intensity (saturation) of the color representation. Even the discrete-stage variant of the Cohen et al. model was able to produce additivity between color saturation and congruency only when this binding of color and word representations was implemented. Given these simulation results, Stafford and Gurney suggested that "no simple inference from factor additivity, or its absence, to underlying architecture is possible" (p. 7).

Coupled with these simulation results, our finding that aspects of recent trial history can be a crucial determinant of whether one observes interactive or additive effects between factors indicates that a new way of interpreting the implications of such effects is needed. Theoretical developments of this nature will depend on building a clearer understanding of how trial history influences current trial performance. To this end, the next step should be to simplify the task environment by manipulating fewer independent variables than we have examined here. Indeed, preliminary assessment of trial history effects in the word naming task with no semantic primes found only weak evidence for the influence of trial history on interactions between frequency or spelling-sound regularity and stimulus quality (D. Besner and S. O'Malley, personal communication, May 14-15, 2012). Additional experiments with substantial power and designs involving just a pair of critical factors should provide helpful constraints on the nature of trial history effects.

In general, we suggest that additivity and interaction between factors may have stronger implications for the nature of stimulus coding or decision processes than for the architectural components of processing operations. Moreover, aggregating data when examining the relationships between factors may mask theoretically important undercurrents pertaining to trial history and stimulus representations and consequently may produce additive effects that turn out to be artifactual. We believe that further efforts to develop computational models of how trial-to-trial variations in stimulus characteristics affect response time and accuracy will be very fruitful.

References

Allain, S., Burle, B., Hasbroucq, T., & Vidal, F. (2009). Sequential adjustments before and after partial errors. *Psychonomic Bulletin & Review, 16*, 356-362.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Becker, C. A. (1979). Semantic context and word-frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 252-259.

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

Besner, D., & 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., O'Malley, S., & Robidoux, S. (2010). On the joint effects of stimulus quality, regularity, and lexicality when reading aloud: New challenges. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 750-764.

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.

Borowsky, R., & Besner, D. (1991). Visual word recognition across orthographies: On the interaction between context and degradation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 272-276.

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

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

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*, 624-652.

Brown, M., & Besner, D. (2002). Semantic priming: On the role of awareness in visual word recognition in the absence of an expectancy. *Consciousness and Cognition, 11*, 402-422.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97*, 332-361.

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.

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E.-J. (2012). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics, 74*, 454-465.

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.

Kinoshita, S., Forster, K. I., & Mozer, M. C. (2008). Unconscious cognition isn't that smart: Modulation of masked repetition priming effect in the word naming task. *Cognition, 107*, 623-649.

Kinoshita, S., Mozer, M. C., & Forster, K. I. (2011). Dynamic adaptation to history of trial difficulty explains the effect of congruency proportion on masked priming. *Journal of Experimental Psychology: General, 140*, 622-636.

Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition, 18*, 655-681.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*:238. doi:

10.3389/fpsyg.2010.00238

Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica, 43*, 381-412.

Lee, M. D., & Dry, M. J. (2006). Decision making and confidence given uncertain advice. *Cognitive Science, 30*, 1081-1095.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. Psychonomic Bulletin & Review, 1, 476-490.

Logan, G. D. (1988). Toward and instance theory of automatization. *Psychological Review, 95*, 492-527.

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

Massaro, D. W., & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology, 23*, 558-614.

Masson, M. E. J., & 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: A framework for analyzing processes in cascade. *Psychological Review, 86*, 287-330.

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

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida word association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*, 402-407.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Newell, B. R., & Lee, M. D. (2011). The right tool for the job? Comparing an evidence accumulation and a naïve strategy selection model of decision making. *Journal of Behavioral Decision Making, 24*, 456-481.

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.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*, 413-425.

Raaijmakers, J. G. W. (2003). A further look at the "language-as-a-fixed-effect fallacy". *Canadian Journal of Experimental Psychology, 57*, 141-151.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "The language-as-a-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*, 416-426.

Rabbitt, P. M. A. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology, 71*, 264-272.

Rabbitt, P. M. A. (1989). Sequential reactions. In D. H. Holding (Ed.), *Human skills* (2nd ed., pp. 147-170). Oxford: John Wiley & Sons.

Rickard, T. D. (2004). Strategy execution in cognitive skill learning: An item-level test of candidate models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 65-82.

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.

Robidoux, S., Stolz, J., & Besner, D. (2010). Visual word recognition: Evidence for global and local control over semantic feedback. *Journal of Experimental Psychology: Human Perception and Performance, 36*, 689-703.

Stafford, T., & Gurney, K. N. (2011). Additive factors do not imply discrete processing stages: A worked example using models of the Stroop task. *Frontiers in Psychology, 2*:287. doi: 10.3389/fpsyg.2011.00287

Stafford, T., Ingram, L., & Gurney, K. N. (2011). Piéron's law holds during Stroop conflict: Insights into the architecture of decision making. *Cognitive Science, 35*, 1553-1566.

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

Stanovich, K. W., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General, 112*, 1-36.

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

Stolz, J. A., Besner, D., & Carr T. H. (2005). Implications of measures of reliability for theories of priming: Activity in semantic memory is inherently noisy and uncoordinated. *Visual Cognition, 12,* 284-336.

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

Thomas, R. D. (2006). Processing time predictions of current models of perception in the classic additive factors paradigm. *Journal of Mathematical Psychology, 50*, 441-455.

Tse, C.-S., & Neely, J. H. (2007). Semantic priming from letter-searched primes occurs for low- but not high-frequency targets: Automatic semantic access may not be a myth. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 1143-1161.

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review, 118*, 583-613.

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General, 123*, 34-80.

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.-S., & 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.

Yap, M. J., Tse, C.-S., & Balota, D. A. (2009). Individual differences in the joint effects of semantic priming and word frequency revealed by RT distributional analyses: The role of lexical integrity. *Journal of Memory and Language, 61*, 303-325.

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