Mem Cogn (2017) 45:480–492
DOI 10.3758/s13421-016-0666-z

Modulation of additive and interactive effects by trial history revisited

Michael E. J. Masson[1] · **Maximilian M. Rabe**[2] · **Reinhold Kliegl**[2]

Published online: 27 October 2016

Abstract Masson and Kliegl (*Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 898–914, 2013) reported evidence that the nature of the target stimulus on the previous trial of a lexical decision task modulates the effects of independent variables on the current trial, including additive versus interactive effects of word frequency and stimulus quality. In contrast, recent reanalyses of previously published data from experiments that, unlike the Masson and Kliegl experiments, did not include semantic priming as a factor, found no evidence for modulation of additive effects of frequency and stimulus quality by trial history (Balota, Aschenbrenner, & Yap, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1563–1571, 2013; O'Malley & Besner, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1400–1411, 2013). We report two experiments that included semantic priming as a factor and that attempted to replicate the modulatory effects found by Masson and Kliegl. In neither experiment was additivity of frequency and stimulus quality modulated by trial history, converging with the findings reported by Balota et al. and O'Malley and Besner. Other modulatory influences of trial history, however, were replicated in the new experiments and reflect potential trial-by-trial alterations in decision processes.

✉ Michael E. J. Masson
mmasson@uvic.ca

✉ Reinhold Kliegl
kliegl@uni-potsdam.de

[1] Department of Psychology, University of Victoria, Room A234, Cornett Building, P.O. Box 1700 STN CSC, Victoria, British Columbia V8W 2Y2, Canada

[2] Department of Psychology, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

Keywords Additive and interactive effects · Effects of trial history · Lexical decision · Data transformation · Linear mixed models

Various accounts of word reading include assumptions about the potential for dynamic adjustments to component processes. At a general level, it has been shown that correct responses leading up to an error are progressively faster, but a correct response following an error is slow (Allain, Burle, Hasbroucq, & Vidal, 2009; Rabbitt, 1966, 1989). A more specific adjustment to reading subprocesses was put forward by Besner, O'Malley, and Robidoux (2010), who proposed that when stimulus quality is low, the nonlexical route to pronunciation is less active, thereby allowing more efficient naming of exception words (items whose correct pronunciation conflicts with output from the nonlexical route; e.g., *pint*). Similarly, the application of a checking strategy in the lexical decision task has been shown to be differentially applied to low- and high-frequency word targets in a mixed list (Yap, Balota, Tse, & Besner, 2008). In addition, the level of difficulty in processing the target item on one trial can influence processing speed on the subsequent trial (Kinoshita, Mozer, & Forster, 2011).

Building on these ideas, we examined in an earlier article the possibility that characteristics of the target stimulus on the immediately preceding trial could affect the processing operations applied on the current lexical decision trial (Masson & Kliegl, 2013). In particular, we demonstrated five such influences when subjects were responding to word targets: (a) faster responses following a trial with a word target; (b) faster responses when the stimulus quality on the previous and current trials was the same, but only if the previous target was a word; (c) over- or underadditive interactions between word frequency and semantic priming depending on the nature of the previous target; (d) over- or underadditive interactions

between word frequency and stimulus quality depending on the nature of the previous target; and (e) speed-up across trials was found only when the previous target was a word.

The finding of nonadditivity between word frequency and stimulus quality that is modulated by trial history is of particular theoretical interest because of previous demonstrations of additivity between these factors (e.g., Becker & Killion, 1977; Yap & Balota, 2007) and the implications of this additivity for computational accounts of word reading (e.g., Besner, Wartak, & Robidoux, 2008; Plaut & Booth, 2000, 2006). Besner and colleagues (e.g., Besner et al., 2008; Borowsky & Besner, 2006) have argued that connectionist models of word reading, because of their inherently interactive modules, cannot account for additive effects using realistic parameter values. They instead argued that separate processing stages may be involved, at least in some word-reading contexts, and that additivity between factors such as stimulus quality and word frequency arises from serial processing across stages rather than interactions between them. Masson and Kliegl (2013) suggested the possibility that additivity between two factors, such as word frequency and stimulus quality, may be generated by two opposing patterns of interaction (one overadditive and one underadditive) that occur under different conditions of trial history—specifically, the characteristics of the target on the previous trial. This idea was encouraged by previous demonstrations of opposing patterns of interaction between word frequency and stimulus quality obtained in a lexical-decision task using pseudohomophones (e.g., *brane*) as nonwords, whereby overadditivity occurred on trials with short response times but underadditivity was observed on trials with long response times (Yap et al., 2008).

Indeed, in two experiments, Masson and Kliegl (2013) observed that the additive effects of frequency and stimulus quality seen in aggregate data emerged from opposite-going interactions between these factors that were associated with different attributes of trial history. In their first experiment, stimulus quality of the previous target was correlated with the interaction pattern. In their second experiment, stimulus quality was manipulated in separate blocks so it could not participate in trial-to-trial variations in processing. Nevertheless, in that experiment over- and underadditive effects of frequency and stimulus quality were correlated with the lexical status of the previous target.

Subsequently, Balota, Aschenbrenner, and Yap (2013) and O'Malley and Besner (2013) reexamined data from previously published word identification experiments to test for possible influences of trial history on the joint effects of word frequency and stimulus quality. Besner and O'Malley found no evidence for modulation of the additivity of these effects by trial history in a word-naming task. Balota et al. also found no modulation of additivity when they reanalyzed data from three lexical-decision experiments. Unlike the Masson and Kliegl (2013) experiments, however, none of the experiments reanalyzed by Balota et al. or by O'Malley and Besner included semantic priming as a factor, and only the Balota et al. experiments used the lexical decision task. Moreover, Scaltritti, Balota, and Peressotti (2013) obtained an overadditive interaction between frequency and stimulus quality that was restricted to trials following an unrelated semantic prime. This result was attributed to a retrospective checking process that was especially time-consuming for the most difficult condition of low-frequency words presented in degraded form. When only unrelated primes were included in an additional experiment, then Scaltritti et al. obtained an additive pattern. These findings indicate that the relationship between the effects of word frequency and stimulus quality depend on the nature of semantic primes presented in advance of target items. It is plausible, then, that the modulation of the frequency by stimulus-quality interaction by trial history reported by Masson and Kliegl is restricted to a particular set of conditions (e.g., when semantic priming is manipulated) and does not generalize. Even so, cross-trial dependencies of this kind would be important to understand because they potentially could impact other manipulations.

Alternatively, given that the modulation effects observed by Masson and Kliegl (2013) were small, it is possible that they likely require substantial power to detect. Whereas Masson and Kliegl used sample sizes of about 70 in each experiment, the experiments reanalyzed by Balota et al. (2013) either manipulated stimulus quality between subjects or used smaller sample sizes (28 for one experiment and 56 for the other). O'Malley and Besner (2013) aggregated data across multiple experiments for a total sample size of 96, but these subjects performed a naming task rather than lexical decision.

We also considered the possibility that the modulation of the frequency by stimulus-quality interaction reported by Masson and Kliegl (2013), particularly given the small effect size of that interaction, might have been produced by a Type I error. Consequently, we conducted two further replications with sample sizes greater than 70 in each case (comparable to the sample sizes used by Masson & Kliegl). In our first replication experiment, we used the same materials as in Masson and Kliegl and in the second experiment we used a new set of nonwords carefully matched to the target words. Our primary interest was in assessing which of the trial-history effects reported in the Masson and Kliegl study could be consistently replicated. Although the modulation of the joint effects of word frequency and stimulus quality was of central importance, other influences of trial history, such as the modulation of the effect of stimulus quality and of speed-up across trials, were of interest as well.

In analyzing the data from these two new experiments, we also considered another issue raised by Balota et al. (2013) related to the application of transformations to response time data. Masson and Kliegl (2013) applied a reciprocal

transformation to response times so that the raw data would approximate a normal distribution, as required by the assumptions of the linear mixed-model analyses they applied. Balota et al. demonstrated that nonlinear transformations such as the reciprocal transformation may distort relationships between factors that are additive in the original response-time metric. Although Masson and Kliegl showed that the influence of trial history on the effects of their manipulated variables was virtually unchanged by the application of the reciprocal transformation, the fact remains that nonlinear transformations have the potential to distort the pattern of interaction between factors. Therefore, we included in our analyses an approach recently recommended by Lo and Andrews (2015) in which a generalized linear mixed-model analysis is applied, assuming a skewed (e.g., inverse Gaussian) rather than normal distribution of response times. In this analysis, a linear relationship is assumed to hold between the independent variables and response time, so there is no risk of effects being distorted through data transformation, but at the same time the requirement of a normal distribution of residuals can be maintained (see Lo & Andrews for details).

Experiment 1

Method

Subjects Seventy-three students at the University of Victoria participated in the experiment in return for extra credit in an undergraduate psychology course.

Materials The same word and nonword targets and word primes were used as in Experiment 1 of Masson and Kliegl (2013). The 240 word targets were classified as high frequency (M = 170,438) or low frequency (M = 16,594) based on the norms from the English Lexicon Project database (Balota et al., 2007). The words were four to seven letters in length. The related primes for high- and low-frequency target words had a similar average degree of forward association strength to their targets (.222 and .226, respectively; Nelson, McEvoy, & Schreiber, 2004). The backward associative strength for all but one related prime–target pair was zero, and the remaining item had a backward strength of .048. Unrelated primes were designated by reassigning primes to targets with the constraint that the new pairing appeared to be unrelated. This reassignment was done within sublists of 30 items. Assignment of these eight sublists (four each of high- and low-frequency targets) to the four experimental conditions (prime relatedness crossed with stimulus quality) was counterbalanced across subjects. Thus, among word targets presented to each subject, half were primed by a related word and half by an unrelated word.

The 240 nonword targets were pronounceable and were of similar length to the word targets. Their mean orthographic neighborhood size was 4.3 (range: 0–17). An English word was selected to serve as a prime for each of these items. An additional set of 32 prime–target pairs (half word targets and half nonword targets) was used for practice trials.

Procedure Subjects were tested individually using a Macintosh computer with items presented in black font on a white background. Subjects were instructed to classify uppercase letter strings as words or nonwords as quickly as possible while maintaining accuracy. On each trial, a fixation cross was presented for 250 ms, followed by a blank screen for 250 ms and a lowercase word prime for 200 ms. The target then appeared either in full contrast or in low contrast (20 % of maximum darkness) until a response was made. In case of an error, the message *ERROR* was presented for 1 s. The session began with 32 practice trials followed by a randomly ordered presentation of 480 critical trials.

Results and discussion

We present a standard analysis of variance (ANOVA) for the response-time and error data from trials with word targets, followed by linear mixed-model (LMM) analyses of response times. In the ANOVA, each subject's response time for a given condition was based on the mean response time across trials in that condition. In the LMM analysis, data from individual trials (subject–item combinations) were considered. The significance criterion was set at .05 for the ANOVA, and a Bayesian evaluation of effects was used for the LMM analyses. We excluded from the response-time analyses those trials on which response time fell outside the range of 300 to 3,000 ms (0.1 %). These boundaries were established so that no more that 0.5 % of the observations would be excluded (Ulrich & Miller, 1994). We also excluded trials on which a response error was made.

Analysis of variance Mean response time to word targets for each condition is shown in Fig. 1. The ANOVA indicated significant main effects of priming, $F(1, 72) = 20.01$, MSE = 1,495, word frequency, $F(1, 72) = 35.97$, MSE = 993, and stimulus quality, $F(1, 72) = 55.52$, MSE = 2,290, with shorter response times associated with related primes, high-frequency words, and clear stimulus quality. In addition, there was a significant interaction between priming and word frequency, indicating a larger priming effect among low-frequency relative to high-frequency words (21 ms vs. 7 ms), $F(1, 72) = 5.97$, MSE = 1,209. No other interactions were significant (Fs < 1).

The mean percentage error for word targets is shown for each condition in Table 1. An ANOVA applied to the error data indicated significant main effects (Fs > 10)

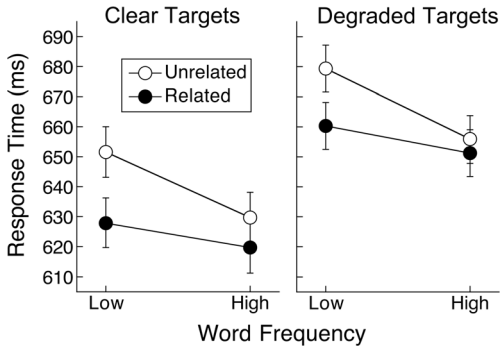

Fig. 1 Mean response time to word targets in Experiment 1 as a function of word frequency, prime, and stimulus quality. Error bars are 95 % within-subject confidence intervals appropriate for comparing condition means within a particular stimulus quality condition (Loftus & Masson, 1994; Masson & Loftus,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

corresponding to those found in the response-time analysis, and no significant interactions (Fs < 1.8); there was no indication of speed–accuracy trade-offs. For nonwords, the mean response times for clear and degraded conditions were 710 ms and 736 ms, respectively, and the overall error rate was 4.8 %.

These analyses provide evidence for the expected additive relationship between word frequency and stimulus quality. Furthermore, the additivity between priming and stimulus quality is consistent with the findings of Stolz and Neely (1995), who observed a similar result with weakly associated pairs (mean forward strength = .175) using the same stimulus onset asynchrony and relatedness proportion as we applied. The mean forward associative strength for our pairs was .224, which is much more similar to their weak pairs than to the strongly associated pairs they used (mean = .560). Unlike the results reported by Masson and Kliegl (2013), the interaction between priming and word frequency appeared in the response-time data and was not restricted to error rates.

Linear mixed-model analysis Following Masson and Kliegl (2013), for the LMM analysis of word-target response times we applied a reciprocal transformation to reaction time (–1/RT, where RT = response time in seconds) to meet the assumption of normally distributed residuals. The analysis was run using the *lmer* function in the *lme4* package in R (Bates, Mächler, Bolker, & Walker, 2015) and the *rePCA* function in the *RePsychLing* package (Bates, Kliegl, Vasishth, & Baayen, 2015). Because of concerns regarding how patterns of interactions between factors might be influenced by nonlinear data transformations (e.g., Balota et al., 2013), parallel analyses using raw response time as the dependent measure were also performed, one using LMM and another using generalized linear mixed models (GLMM) in which we assumed an inverse Gaussian distribution of the data (cf. Lo & Andrews, 2015). We report the results of those two additional analyses only where they differ substantially from the analysis of the transformed data. Degrees of freedom are generally not precisely known for t ratios in LMMs, making significance testing problematic. Moreover, given the recent discussions of shortcomings of null-hypothesis significance testing in general (e.g., Kruschke, 2013; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016), we report the 95 % highest posterior density interval (computed using the *profile* function in the *lme4* package) for each fixed effect. We were restricted to using t tests to evaluate the results of the GLMM, however, because the R function available for that analysis (*glmer*) does not yet allow for application of Bayesian tests when an inverse Gaussian distribution is assumed. A t ratio with absolute value greater than 2 was deemed to be significant (e.g., Kliegl, Masson, & Richter, 2010).

Fitting the maximal LMM for this experiment requires the estimation of 697 parameters (32 fixed effects, including the intercept; 32 variance components and 496 correlation parameters for the subjects random factor; 16 variance components and 120 correlation parameters for the items random factor; 1 residual variance). To break this number down: The subjects factor yields 32 (i.e., 2^5) variance components, consisting of the mean reciprocal RT (intercept) plus one for each of the main effects and interactions of the five within-subject factorial design. The items factor yields 16 variance components (i.e., 2^4) for the four-factor within-item factorial design (word frequency is a between-word factor). In addition, the maximal model includes 496 correlation parameters for the subject factor (i.e., (32)(31)/2) and 120 correlation parameters for the item factor (i.e., (16)(15)/2).

Fitting this maximal LMM with 697 parameters took 103 hours and 50 minutes on a computer cluster (3.5 GHz) at the University of Potsdam, running R 3.3.0 on Scientific Linux 6.5. We also had to choose a simpler optimization solution to obtain the result (i.e., we did not compute the gradient and Hessian of nonlinear optimization solution; rather, we specified the argument "control = lmerControl(calc.derivs = F)" in the *lmer* function call). The fit yielded a convergence warning, but estimates looked reasonable. Nevertheless, such a complex model is highly likely to be overidentified (degenerate); that is, parameters are not supported by the data (Bates, Kliegl, et al., 2015). One way to check for model overparameterization is to determine the dimensionality of the variance–covariance matrix of random effects, specifically to determine the number of principle components (PCs)

Table 1 Mean percentage error for word targets in Experiment 1 as a function of word frequency, stimulus quality, and prime relatedness

Word freq.	Clear		Degraded	
	Related	Unrelated	Related	Unrelated
High	2.3	3.0	3.1	4.3
Low	3.8	5.0	4.8	6.9

accounting for some nonzero amount of variance. We used the rePCA() function of the *RePsychLing* package (Bates, Kliegl, et al., 2015) to this end. For the subjects random factor, the first 16 PCs accounted for 99 % of the variance, but, somewhat surprisingly, all of the 32 PCs were different from zero (even if only slightly so). For the items random factor, the first eight PCs accounted for 99.5 % of the variance and 15 of the 16 PCs were different from zero. Thus, the overparameterization was nominally much smaller than expected.

Even if the maximal LMM is (barely) identified, a large number of model parameters may contribute negligibly to the goodness of fit of the model. In the long run, removing redundant parameters (i.e., fixing them at zero) yields better statistical power for fixed effects, even if the true value of these parameters is different from zero (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2015; for a different perspective on this issue, i.e., a preference to keep it maximal rather than parsimonious, see Barr, Levy, Scheepers, & Tily, 2013). A reasonable first step is to check whether the correlation parameters contribute to the goodness of fit. The estimation of the zero-correlation-parameter (zcp) LMM took 139 hours and 43 minutes when computing the gradient and Hessian of nonlinear optimization solution. For the subjects random factor, the first 11 PCs accounted for 99 % of the variance and the last 15 PCs were very close zero (i.e., < 0.02 %). For the items random factor, the first six PCs accounted for 99.7 % of the variance and the last seven of the 16 PCs were zero. Thus, the overparameterization was actually much more pronounced in this less complex zcp LMM than the maximal LMM. This could be due to the less precise optimization routine in the latter. Nevertheless, in a likelihood ratio test, there was no evidence for a loss of goodness of fit of the zcp LMM relative to the maximal LMM due to dropping 616 correlation parameters, $\Delta\chi^2(616) = 281$.

In the next step, we compared the zcp LMM with the LMM reported in Masson and Kliegl (2013). Similar to recommendations in Bates, Kliegl, et al. (2015), Masson and Kliegl had used an iterative procedure to arrive at a parsimonious LMM. This model included three variance components and one correlation parameter for the subjects random factor and two variance components for the items random factor. Obviously, this LMM was not overparameterized. A likelihood ratio test for the difference between the zcp LMM and the parsimonious LMM yielded a $\Delta\chi^2(42) = 9.2$ – indicating no loss of goodness of fit. In a separate test, we ascertained that the correlation parameter (which is not estimated in the zcp LMM) was significant, $\Delta\chi^2(1) = 9.06$, $p < .01$. Thus, the 39-parameter parsimonious LMM accounts for the data as well as the 697-parameter maximal LMM.

The estimates of variance components for the random effects of subjects and items generated by the parsimonious LMM are listed in the upper section of Fig. 2. These values are quite similar in magnitude to those reported by Masson and Kliegl (2013, Exp. 1). In addition, the parameter for the correlation between the intercept (mean) and the effect of stimulus quality (-.42) was very similar to the value found by Masson and Kliegl.

The results for fixed effects are shown in the lower part of Fig. 2. The intercept for fixed effects was –1.64, and is not depicted in Fig. 2 because it is beyond the limits of the scale used to show the other fixed effects. The estimates did not depend on the random-effect structure (i.e., the same pattern of significance was obtained for maximal, zero-correlation, and parsimonious LMM specifications). As in the ANOVA, all three primary main effects (frequency, priming, and stimulus quality) were reliable using the criterion of the 95 % highest probability density interval (HPDI) not including zero, and there was an interaction between frequency and priming. This interaction is shown in Fig. 3, where it can be seen that it takes the same form as in the raw score means (see Fig. 1). Two other two-way interactions were reliable: stimulus quality interacted with both trial-history factors (stimulus quality and lexical status of the previous target). The three-way interaction between stimulus quality and the two trial-history factors was also reliable, and this interaction is plotted in Fig. 4. The pattern of this interaction was the same as in Experiment 1 of Masson and Kliegl (2013), whereby subjects responded faster if the stimulus quality on the current trial matched that of the previous trial, but only if the previous target was a word. Exactly the same pattern of fixed effects was found when raw response time rather than the transformed measure was analyzed and when GLMM was applied with an inverse Gaussian distribution of scores assumed.

To help readers interpret the three-way interaction between stimulus quality and the two trial history factors, we present in Table 2 the mean response time in ms for each of the relevant conditions. It can be seen that when the previous trial's target was a word, there is an average benefit of about 13 ms in responding on the current trial if stimulus quality is the same on the previous and current trials. In addition, we provide the percentage error for each of those conditions to verify that the interaction is not the product of a speed–accuracy trade-off.[1] Differences in response time that favor repetition of stimulus quality across trials are accompanied by small differences in percentage error that also favor such repetition.

Two fixed-effect interactions reported by Masson and Kliegl (2013) failed to materialize in this experiment. One was a four-way interaction between frequency, prime, and the trial-history factors. In this experiment, we instead obtained the standard interaction between frequency and prime, with no evidence that that interaction was modulated by trial

[1] We thank Derek Besner for suggesting that we present this information and check for a possible speed–accuracy trade-off, here and in Experiment 2.

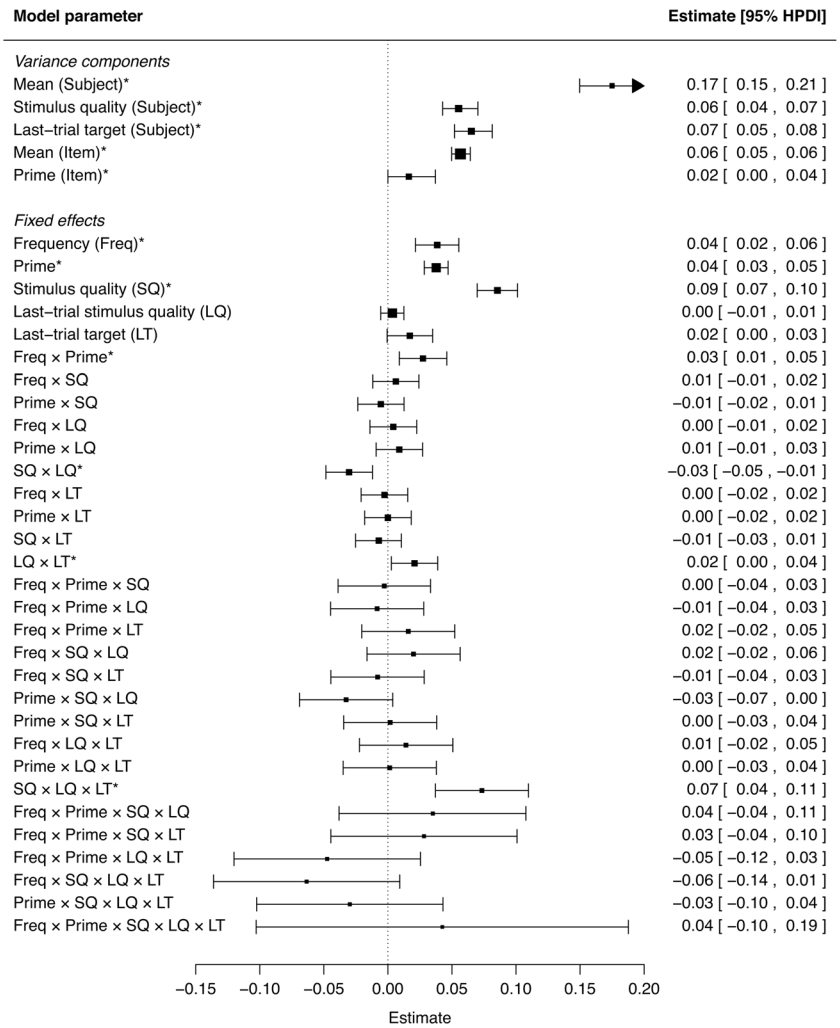

Fig. 2 Parameter estimates and 95 % highest posterior density intervals (HPDIs) for square root of variance components (standard deviations) and fixed effects produced by the parsimonious LMM for Experiment 1. Not shown are the estimates of the mean (intercept): -1.64 [-1.68, -1.60], the residual: 0.294 [0.291, 0.297], and the correlation parameter for the subject-related mean and the effect of stimulus quality: -0.42 [-0.64, -0.16]. Symbol sizes are in proportion to the precision of the estimates. The plot is based on the *forest* function of the *metafor* package in R (Viechtbauer, 2010)

history. Similarly, the three-way interaction between frequency, stimulus quality, and previous trial stimulus quality that Masson and Kliegl obtained was not replicated here. Instead, we observed additivity between frequency and stimulus quality, unmodulated by trial history. This outcome is consistent with the reanalyses of data reported by Balota et al. (2013) and by O'Malley and Besner (2013). We were mindful of the Scaltritti et al. (2013) finding that frequency and stimulus quality interacted when unrelated semantic primes were used. We tested for this possibility in our data by analyzing separately related and unrelated prime trials, but still no frequency by stimulus-quality interaction emerged in either case. There are a number possible reasons that we did not replicate the Scaltritti et al. finding. First, we used a lexical-decision task, whereas their task was speeded pronunciation. It is known that pure word lists used in the pronunciation task can produce an interaction between frequency and stimulus quality (O'Malley & Besner, 2008). Second, Scaltritti et al. attributed their result to retrospective checking that was especially time-consuming for low-frequency words presented in degraded form. Such a process would be less likely in our experiment because we used a much shorter stimulus onset asynchrony than Scaltritti et al. (200 ms vs. 800 ms), and our manipulation of degradation was apparently much weaker than theirs; response time to degraded targets in their experiment was about 100 ms longer than in our study, despite the fact that pronunciation response times are usually noticeably shorter than lexical-decision times.

An additional LMM analysis included trial as a covariate to investigate the change in response time over the course of the testing session. Masson and Kliegl (2013) reported that response time to word targets decreased across trials, but only in cases where the preceding trial's target was a word. For this analysis, we included centered trial number as a factor along

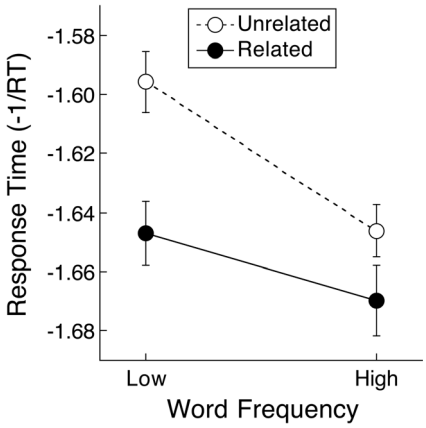

Fig. 3 Mean transformed response time to word targets in Experiment 1 as a function of word frequency and prime. Error bars are 95 % within-subject confidence intervals appropriate for comparing all condition means (Morey, 2008)

Table 2 Mean response time (ms) and percentage error for word targets in Experiment 1 as a function of stimulus quality and trial history

Stimulus	Trial history			
	Word		Nonword	
quality	Clear	Degraded	Clear	Degraded
Resp. time				
Clear	627	634	635	639
Degraded	670	651	661	667
% error				
Clear	3.6	3.7	3.6	3.1
Degraded	4.5	4.3	4.3	6.0

Note. Trial history refers to the lexical status and stimulus quality of the target on the previous trial

with its interactions with all other fixed effects. Overall, subjects' response times decreased over the course of the experiment (coefficient = –2.02, 95 % HPDI: [–2.60, –1.44]). The modulation of this speed-up effect by the previous target's lexical status was reliable for transformed response time (coefficient = 0.00008, 95 % HPDI: [0.00001, 0.00014]), but not for raw response time. The pattern of the modulation was similar to that found by Masson and Kliegl, with greater improvement across trials when the previous target was a word rather than a nonword (see Fig. 5).

Experiment 2

Given the failure to replicate the modulation of the relationship between word frequency and stimulus quality in Experiment 1, we attempted an additional replication in Experiment 2. For this experiment, we modified the nonword items to make them more similar to the target words with respect to letter length, length of subsyllabic segments, and bigram transition frequencies. This change was made to further examine the influence of trial history on the reduction of response time to words across trials seen in Experiment 1 and in Masson and Kliegl (2013). Less improvement was seen if the previous trial's target was a nonword (see Fig. 5). Yap, Sibley, Balota, Ratcliff, and Rueckl (2015) have shown that lexical-decision responses to nonword targets are systematically affected by orthographic characteristics and base-word frequency in the case of nonwords derived from a specific base word. Our interest in Experiment 2 was to determine whether nonwords that were more word-like would modulate the improvement in responses to word targets in the same way that the less well controlled nonwords did in Experiment 1.

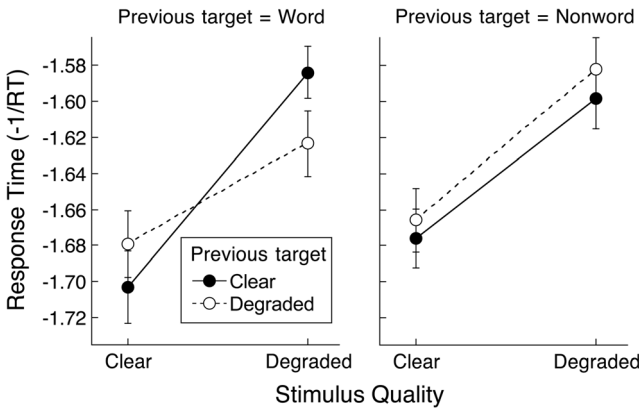

Fig. 4 Mean transformed response time to word targets in Experiment 1 as a function of stimulus quality and lexical status and stimulus quality of the target on the previous trial. Error bars are 95 % within-subject confidence intervals appropriate for comparing all condition means (Morey, 2008)

Method

Subjects A new sample of 72 students from the same source as in Experiment 1 participated in the experiment.

Materials and procedure The same materials were used as in Experiment 1 except that the nonword targets were replaced by items matched to the word targets using the *Wuggy* application (Keuleers & Brysbaert, 2010). Items were matched on letter length, length of subsyllabic segments, and bigram transition frequencies, making this set of nonwords more word-like than those used in Experiment 1, although their mean orthographic neighborhood size, 5.4 (range: 0–20), was similar to that of the nonwords used in Experiment 1. Subjects were tested using the same procedure as in Experiment 1.

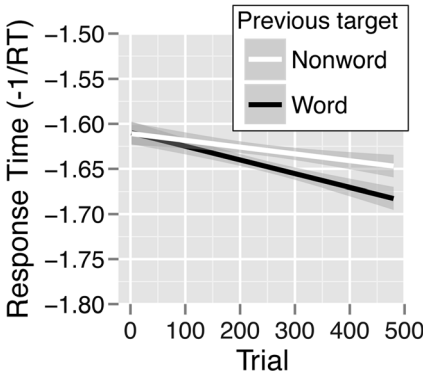

Fig. 5 Mean transformed response time to word targets in Experiment 1 as a function of trial number and lexical status of the target on the previous trial. Continuous error bars (shown as gray bands) are 95 % confidence intervals for the regression lines

Results and discussion

The same upper and lower bounds for response times were set as in Experiment 1 (300 ms and 3,000 ms), which led to the exclusion of less than 0.1 % of trials with correct responses.

Analysis of variance Mean response time for word targets in each condition is shown in Fig. 6. In general, the pattern of means was quite similar to that found in Experiment 1. An ANOVA indicated that all three main effects were significant, priming, $F(1, 71) = 41.87$, $MSE = 1{,}319$, word frequency, $F(1, 71) = 43.32$, $MSE = 1{,}026$, and stimulus quality, $F(1, 71) = 58.70$, $MSE = 4{,}235$. Unlike Experiment 1, the interaction between priming and word frequency was not significant, $F(1, 71) = 2.02$, $MSE = 790$, although there was a significant interaction between priming and stimulus quality, $F(1, 71) = 5.28$, $MSE = 1{,}014$, with larger priming for degraded targets (25 ms vs. 14 ms). This interaction was not expected, based on previous studies using the short prime duration and weak associative strength between primes and targets employed here (Masson & Kliegl, 2013; Stolz & Neely, 1995). Its appearance in this experiment, but not in previous work, might be due to

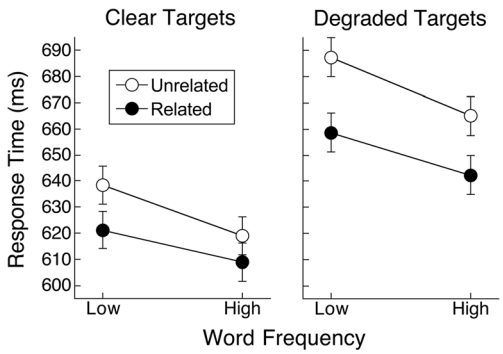

Fig. 6 Mean response time to word targets in Experiment 2 as a function of word frequency, prime, and stimulus quality. Error bars are 95 % within-subject confidence intervals appropriate for comparing condition means within a particular stimulus quality condition

lower power to detect a small effect in earlier experiments. It is also possible that differences between samples of subjects with respect to the experienced associative strength between primes and targets could be responsible for the variation in results. Stolz and Neely obtained an interaction between priming and stimulus quality with strongly, but not weakly, associated pairs. Our items were on average somewhat higher in strength than were those of Stolz and Neely, although not greatly so (.224 vs. .175). Still, it is possible that subjects in Experiment 2 experienced the pairs as more strongly related than implied by the normative data. Thomas, Neely, and O'Connor (2012) have shown that interactions between priming and stimulus quality arise from backward associations between primes and targets. We can be sure, however, that this factor cannot account for the interaction we obtained because for our materials, the strength of backward associations was virtually zero.

The mean percentage error for word targets is shown in Table 3 for each condition. An ANOVA applied to the error data indicated significant main effects (Fs > 20) corresponding to those found in the response-time analysis. There was weak evidence for a priming by frequency interaction, $F(1, 71) = 3.62$, $MSE = 10.95$, $p < .07$, suggesting that in Experiment 2 that interaction was partly manifest in response time and partly in error rates rather than concentrated fully in response times, as in Experiment 1. All other interactions were nonsignificant (Fs < 2.6).

Linear mixed-model analysis The LMM analysis on reciprocal-transformed response times was carried out on word targets, as described for Experiment 1. Again, the maximal LMM (time for estimation: 103 hours, 23 minutes) and zcp LMM (time for estimation: 135 hours, 11 minutes) were overparameterized, as determined with the *rePCA* function. For the parsimonious LMM, we started as in Experiment 1, but the model fit improved after dropping the parameter for the correlation of intercept and stimulus quality for subjects, and this was the only difference in the final model for Experiment 2 relative to the model used for Experiment 1. Once more, there was neither a difference in goodness of fit between maximal LMM and zcp LMM, $\Delta\chi^2(616) = 291$, nor a difference in goodness of fit between zcp LMM and parsimonious LMM, $\Delta\chi^2(42) = 26.4$. Thus, the 38-parameter

Table 3 Mean percentage error for word targets in Experiment 2 as a function of word frequency, stimulus quality, and prime relatedness

Word freq.	Clear		Degraded	
	Related	Unrelated	Related	Unrelated
High	1.3	2.5	2.8	3.1
Low	3.2	4.5	4.3	6.5

parsimonious LMM accounted for the data as well as the 697-parameter maximal LMM.

The variance components for the parsimonious LMM appear in the upper part of Fig. 7. Item-related variance components were very similar to what was reported by Masson and Kliegl (2013, Exp. 1) and to those in the present Experiment 1. The fixed effects are shown in the lower part of Fig. 7. The estimates did not depend on the random-effect structure (i.e., the same pattern of significance was obtained for maximal, zero-correlation, and parsimonious LMM specifications). All three main effects were reliable, as in Experiment 1. In addition, responses were faster if the previous trial's target was a word, and there was a priming by stimulus-quality interaction, which is shown in Fig. 8. Priming was somewhat larger for degraded targets, as discussed in the ANOVA results. As in Experiment 1, stimulus quality interacted with both trial history factors, and in addition, this effect was modulated by the priming factor, as shown in Fig. 9. This four-way interaction indicates that maintaining the same level of stimulus quality from the previous trial reduced response time, but only if the previous target was a word and the current prime was related to its target. Relative to Experiment 1, then, prime relatedness was added as a new constraint on the benefit of repeating stimulus quality across trials. In the two analyses of raw response time (LMM and GLMM assuming an inverse Gaussian distribution), this four-way interaction was not significant, but otherwise the pattern of significant effects was the same across all three analyses. As in Experiment 1, to assist with interpretation of the clearly reliable three-way interaction between stimulus quality and the two trial history factors (which appeared in all three analyses), we present the corresponding mean response times in ms in Table 4, along with the percentage error for each condition. Once again, the response-time benefit (9 ms) of repeating stimulus quality from one trial to the next when word targets appear on both trials was not a consequence of a speed–accuracy trade-off; the benefit is also apparent in small differences in percentage error that favor repetition of stimulus quality.

We conducted separate analyses for trials with related versus unrelated primes, as in Experiment 1, to check for a possible replication of the frequency by stimulus-quality interaction reported by Scaltritti et al. (2013). Once again, neither analysis obtained an interaction between these factors.

When trial was included as a covariate, we found that response times decreased across trials (coefficient = –2.83, 95 % HPDI: [–3.40, –2.25]). Unlike Experiment 1, there was no reliable evidence that this effect was modulated by the lexical status of the previous trial's target (see Fig. 10), despite the fact that responses were generally faster following a trial with a word target, as described above. The same pattern of results was found when raw response time was used as the dependent variable. The use of nonwords that more closely conform to orthographic rules may have prevented the influence of the previous trial's target on the general speed up in responding over the course of the experiment. Alternatively, given the relatively small size of the influence of the previous target on speed-up over trials in Experiment 1, the lack of an effect here may simply have been due to low power. In support of this latter possibility, we note that this interaction was significant in the two experiments in Masson and Kliegl (2013). Moreover, for nonword targets, Experiments 1 and 2 showed very comparable effects of trial speed up and modulation of that enhancement by the previous target. Figure 11 shows the nature of this modulation for each experiment, whereby a significantly greater speed-up over the course of the experiment was observed when the previous target was a nonword (Exp. 1: coefficient = .00015; Exp. 2: coefficient = .00011), although responding was slower overall in that case.

General discussion

The two replication experiments reported here found no evidence for modulation of the joint effects of word frequency and stimulus quality due to trial history. This outcome is not consistent with the original modulation effect reported by Masson and Kliegl (2013), but converges with the reanalyses of previously published data reported by O'Malley and Besner (2013) and by Balota et al. (2013). Taken together, these results, coupled with the relatively small size of the modulation of additivity found by Masson and Kliegl, suggest that that effect should be attributed to a Type I error. Another feature of the modulation of additivity that Masson and Kliegl obtained that supports the validity of this conclusion is the inconsistency across their two experiments with respect to the factors that produced the modulation. In their Experiment 1, the modulating factor was the previous trial's stimulus quality, whereas in their second experiment both prime relatedness and the lexical status of the previous target modulated the frequency by stimulus-quality interaction. In retrospect, this inconsistency in the modulating factors across experiments might be taken as grounds for doubting that the modulation is genuine. Thus, the additivity between word frequency and stimulus quality appears to continue to stand as a benchmark result to be accounted for by models of word reading.

Masson and Kliegl (2013) also reported an influence of the lexical status and stimulus quality of the previous trial's target on the effect of stimulus quality on the current trial, whereby maintaining the level of stimulus quality across trials when both targets were words led to faster responding. This effect was replicated in both experiments here, and a similar result has recently been reported by Balota, Aschenbrenner, and Yap (in press). Masson and Kliegl proposed an explanation for this effect based on an account of stimulus learning and classification developed by Turner, Van Zandt, and Brown (2011).

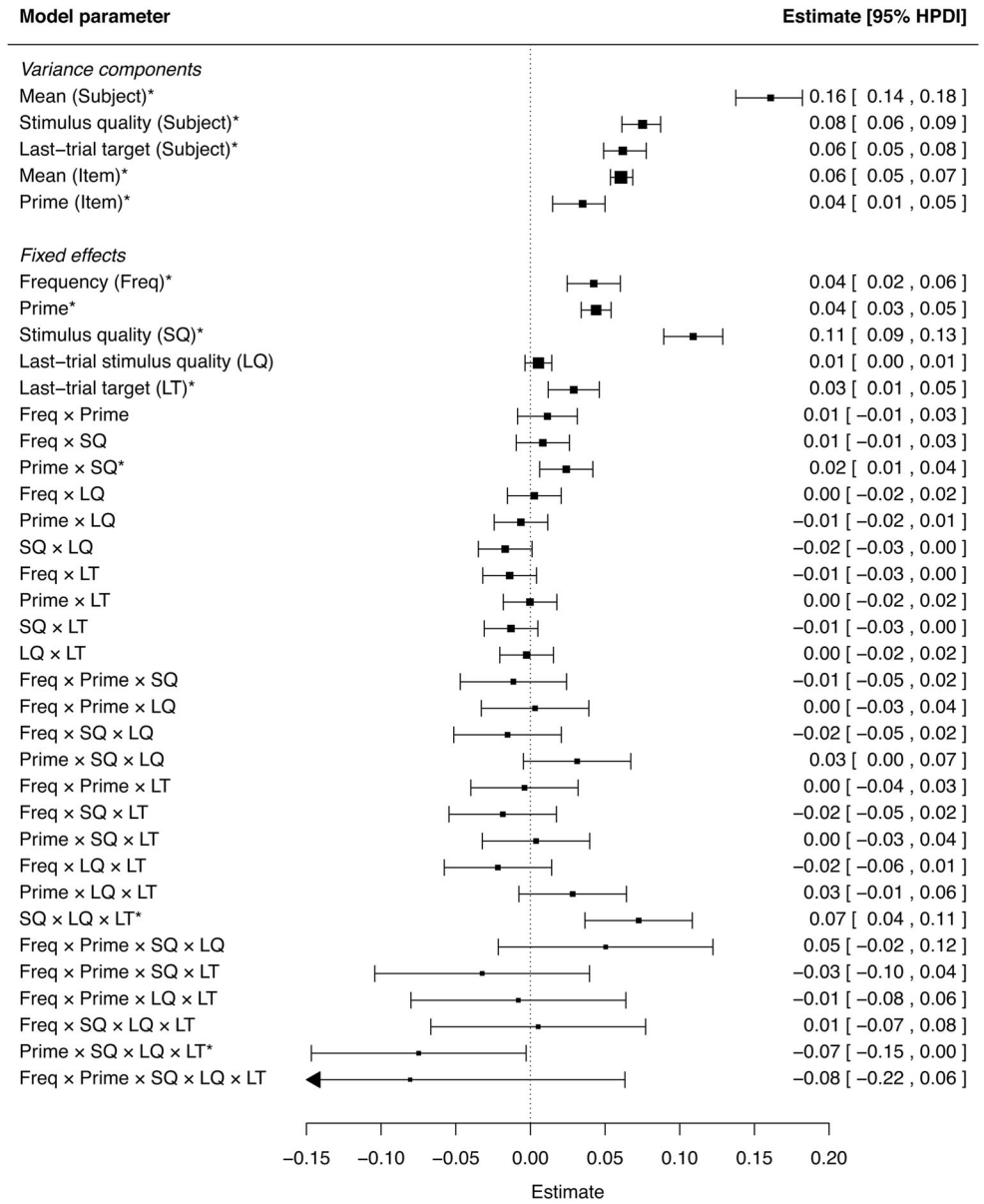

Fig. 7 Parameter estimates and 95 % highest posterior density intervals (HPDIs) for square root of variance components (standard deviations) and fixed effects produced by the parsimonious LMM for Experiment 2. Not shown are estimates of the mean (intercept): -1.64 [-1.68, -1.61] and the residual: 0.289 [0.286, 0.292]. Symbol sizes are in proportion to the precision of the estimates. The plot is based on *forest* function of *metafor* package in R (Viechtbauer, 2010)

On that account, subjects generate signal-to-noise likelihood ratios for different points on a stimulus strength axis. These ratios are modified by experience, such that presentation of a stimulus occupying a particular point on the strength axis and belonging to the signal category will increase the likelihood ratio for any stimulus with a similar strength value. Therefore, when two signal stimuli have similar strength values (e.g., two degraded word targets), the presentation of one will strengthen the signal-to-noise likelihood ratio of the other.

Another possible interpretation of the modulation of stimulus-quality effects by trial history draws upon the diffusion model of lexical decision (Ratcliff, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). In this model, evidence accumulation follows a random walk toward one of two possible response boundaries (word and nonword in the case of a lexical-decision task). When a boundary is reached, the corresponding response is made. The distance between response boundaries can vary across trials, and this provides a mechanism for explaining speed–accuracy trade-offs. Bringing boundaries closer together means that less evidence is required to make a response, but possibly at the expense of a higher risk of error.

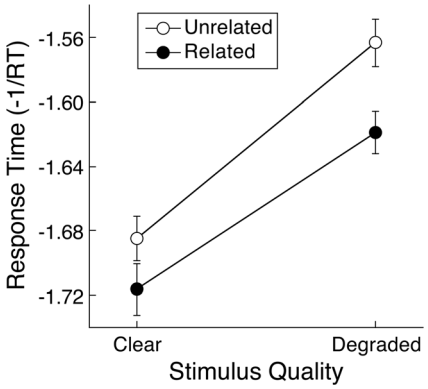

Fig. 8 Mean transformed response time to word targets in Experiment 2 as a function of stimulus quality and prime. Error bars are 95 % within-subject confidence intervals appropriate for comparing all condition means (Morey, 2008)

In one application of this idea, Dufau, Grainger, and Zielger (2012) proposed a leaky competing accumulator model of lexical decision in which trial-by-trial adjustments to response boundaries in a random walk process can be made. In this model, each correct response leads to a small reduction in the response thresholds for word and nonword responses. We suggest that there may be separate response thresholds for clear and degraded stimuli, and that an adjustment process

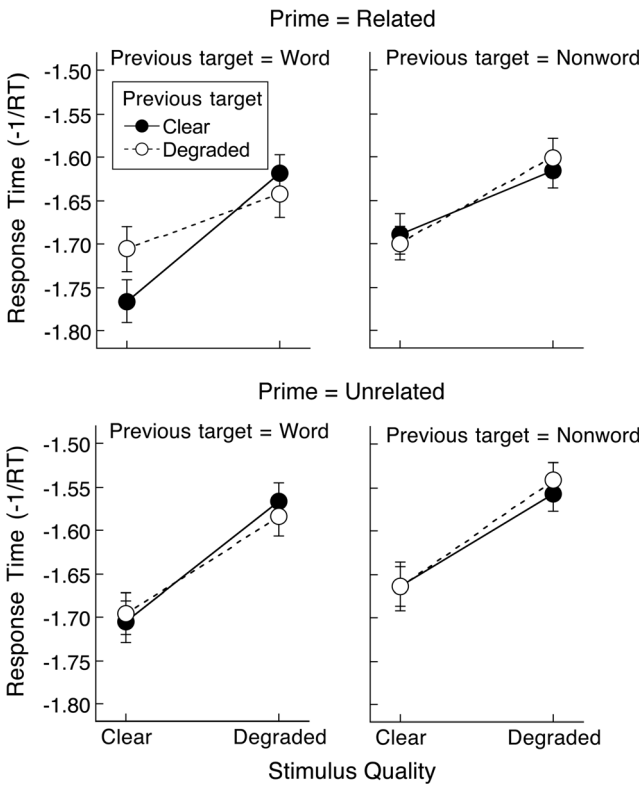

Fig. 9 Mean transformed response time to word targets in Experiment 2 as a function of prime, stimulus quality, and lexical status and stimulus quality of the previous target. Error bars are 95 % within-subject confidence intervals appropriate for comparing all condition means (Morey, 2008)

Table 4 Mean response time (ms) and percentage error for word targets in Experiment 2 as a function of stimulus quality and trial history

Stimulus	Trial history			
	Word		Nonword	
quality	Clear	Degraded	Clear	Degraded
Resp. time				
Clear	611	626	632	626
Degraded	663	660	664	669
% error				
Clear	2.8	3.3	2.8	2.6
Degraded	5.1	3.9	3.4	4.3

Note. Trial history refers to the lexical status and stimulus quality of the target on the previous trial

of sort described by Dufau et al. might be specific to the stimulus quality of the target. Thus, when a correctly classified word target is followed by another target of the same stimulus quality, response boundaries are moved closer together, enabling a faster word response. It is noteworthy that improved response speed found on trials where stimulus quality was repeated did not incur a cost with respect to accuracy (see Tables 2 and 4). This finding implies that subjects were initially operating with a relatively conservative set of thresholds that they could afford modestly to reduce without an elevated risk of error.

These experiments also showed that modulation of improved response times across trials by the lexical status of the previous target may depend on the type of nonword used. In Experiment 1, we used the same nonwords as Masson and Kliegl (2013) and replicated the general effect of trial history on response speed-up across trials. These nonwords conformed to English rules of pronounceability, but unlike the nonwords used in Experiment 2, they were not quantitatively matched to the word set with respect to orthographic

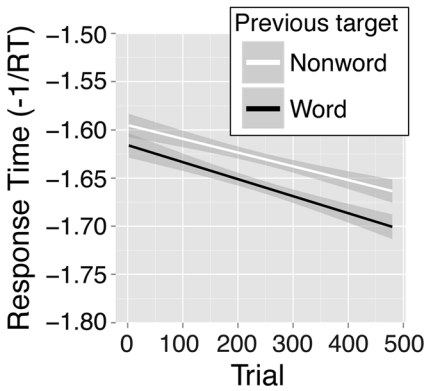

Fig. 10 Mean transformed response time to word targets in Experiment 2 as a function of trial number and lexical status of the target on the previous trial. Continuous error bars (shown as gray bands) are 95 % confidence intervals for the regression lines

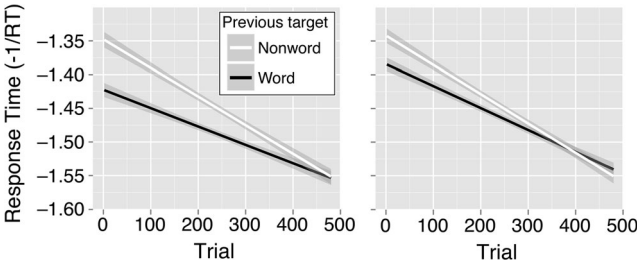

Fig. 11 Mean transformed response time to nonword targets in Experiments 1 (left) and 2 (right) as a function of trial number and lexical status of the target on the previous trial. Continuous error bars (shown as gray bands) are 95 % confidence intervals for the regression lines

patterns such as bigram transition frequencies. Masson and Kliegl suggested that the influence of the previous target's lexical status on increased speed across trials was a result of fluctuating response thresholds, with caution being applied after a nonword target. This type of adjustment might be captured in the Ratcliff et al. (2004) diffusion model. Wagenmakers et al. (2008) showed that manipulations that affect response criteria, such as speed–accuracy instructions or the proportion of word versus nonword items in the stimulus set, are captured by adjustments of the decision boundaries in the model. We suggest that the speed-up in responding across trials can be viewed as improved stimulus discrimination with continuing practice at the lexical decision task. The modulation of this speed-up by the lexical status of the previous target (see Fig. 5) can be attributed to changes in response boundaries. Specifically, for word targets in Experiment 1, as trials progressed, the decision boundary for "word" responses was moved further away from the starting point (requiring more evidence for a response) when the previous target was a nonword, or closer to the starting point when the previous target was a word. Experiment 2, however, did not replicate this type of boundary shift for word targets. For nonword targets, a boundary adjustment due to the nature of the previous target appears to have been in effect for early trials, but it dissipated as trials progressed. This adjustment could involve moving the boundary for "nonword" responses further away from the starting point after experiencing a nonword target and/or closer to the starting point after a word target.

Finally, we note that the results of our linear mixed-model analyses were quite consistent across the two variants of the dependent measure that we examined, reciprocal and raw response time. As Balota et al. (2013) pointed out, nonlinear transformations such as the reciprocal transformation have the potential to modify the pattern of additivity and interaction between independent variables. The present results, coupled with the consistency of effects across reciprocal and raw response times reported by Masson and Kliegl (2013), suggest that the use of a nonlinear transformation was not responsible for the effects they reported. The Masson and Kliegl finding of modulation of the joint effects of word frequency and stimulus quality by trial history, however, was clearly not replicated with either dependent measure in our experiments. As a result, we conclude that the additive effects of these two factors in the lexical-decision task appears to be a robust phenomenon.

Author Note This research was supported by a discovery grant to Michael E. J. Masson from the Natural Sciences and Engineering Research Council of Canada and a grant to Reinhold Kliegl from the Deutsche Forschungsgemeinschaft. Data and R scripts for all analyses are available upon request and at the Potsdam Mind Research Repository (http://read.psych.uni-potsdam.de/PMR2/). We thank Marnie Jedynak for assistance with data collection. We are also grateful to Derek Besner and Sachiko Kinoshita for very helpful comments on an earlier version of this article.

References

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

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: The influence of trial history and data transformation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1563–1571.

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (in press). Dynamic adjustment of lexical processing in the lexical decision task: Cross-trial sequence effects. *Quarterly Journal of Experimental Psychology.*

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Trieman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445–459.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models.* Retrieved from [stat.ME]

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed models using *lme4. Journal of Statistical Software, 67,* 1–48.

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., & 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. (2006). Parallel distributed processing and lexical-semantic effects in visual word recognition: Are a few stages necessary? *Psychological Review, 113,* 181–193.

Dufau, S., Grainger, J., & Zielger, J. C. (2012). How to say "no" to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1117–1128.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42,* 627–633.

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.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142,* 573–603.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyze reaction time data. *Frontiers in Psychology, 6*(1171). doi:10.3389/fpsyg.2015.01171

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

Masson, M. E. J., & Kliegl, R. (2013). Modulation of additive and interactive effects in lexical decision by trial history. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 898–914.

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

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2015). *Balancing Type I error and power in linear mixed models.* Retrieved from arXiv:1511.01864 [stat.ME]

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4,* 61–64.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23,* 103–123.

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.

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

O'Malley, S., & Besner, D. (2013). Reading aloud: Does previous trial history modulate the joint effects of stimulus quality and word frequency? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1321–1325.

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.

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, UK: Wiley.

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

Scaltritti, M., Balota, D. A., & Peressotti, F. (2013). Exploring the additive effects of stimulus quality and word frequency: The influence of local and list-wide prime relatedness. *Quarterly Journal of Experimental Psychology, 66,* 91–107.

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, M. A., Neely, J. H., & O'Connor, P. (2012). When word identification gets tough, retrospective semantic processing comes to the rescue. *Journal of Memory and Language, 66*, 623–643.

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.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the *metafor* package. *Journal of Statistical Software, 36,* 1–48.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58,* 140–150.

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., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 597–613.