2015

Is reading-aloud performance in megastudies systematically influenced by the list context?

Jonathan Bruce Santo

The Quarterly Journal of Experimental Psychology

ISSN: 1747-0218 (Print) 1747-0226 (Online) Journal homepage: http://www.tandfonline.com/loi/pqje20

Is reading-aloud performance in megastudies systematically influenced by the list context?

Michael J. Cortese, Sarah Hacker, Jocelyn Schock & Jonathan B. Santo

To cite this article: Michael J. Cortese, Sarah Hacker, Jocelyn Schock & Jonathan B. Santo (2015) Is reading-aloud performance in megastudies systematically influenced by the list context?, The Quarterly Journal of Experimental Psychology, 68:8, 1711-1722, DOI: 10.1080/17470218.2014.974624

To link to this article: http://dx.doi.org/10.1080/17470218.2014.974624

Accepted author version posted online: 14 Oct 2014.
Published online: 18 Nov 2014.

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2015
Vol. 68, No. 8, 1711–1722, http://dx.doi.org/10.1080/17470218.2014.974624

Is reading-aloud performance in megastudies systematically influenced by the list context?

Michael J. Cortese, Sarah Hacker, Jocelyn Schock, and Jonathan B. Santo

Department of Psychology, University of Nebraska at Omaha, Omaha, NE, USA

(Received 22 January 2014; accepted 26 August 2014; first published online 18 November 2014)

To examine megastudy context effects, 585 critical words, each with a different orthographic rime, were placed at the beginning or end of a 2614-word megastudy of reading aloud. Sixty participants (30 participants in each condition) responded to these words. Specific predictors examined for change between beginning and end conditions were frequency, length, feedforward rime consistency, feedforward onset consistency, orthographic neighbourhood size, age of acquisition (AoA), and imageability. While it took longer to respond to items at the end of the experiment than items at the beginning of the experiment, there was very little change in the effects of the specific variables assessed. Thus, there is little evidence of list context effects influencing the estimates of the predictor variables in large-scale megastudies.

Keywords: Megastudy; Reading aloud.

Until fairly recently, computational models of word recognition (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Plaut, McClelland, Seidenberg, & Patterson, 1996) were deemed successful if they exhibited effects of frequency, consistency, orthographic neighbourhood size, length, and so on. Spieler and Balota (1997) revolutionized the field of visual word recognition when they computed mean reaction times (RTs) from 31 subjects for each of 2820 items and found that contemporary models (e.g., Plaut et al., 1996, Simulation 3) accounted for very little item-level variance (3%) in RT. Since then, accounting for item-level variance has become a key criterion for assessing word-processing models (see, e.g., Perry, Ziegler, & Zorzi, 2007). Furthermore, the recent proliferation of megastudies has been noteworthy, as has been the reanalysis of existing megastudy databases. Perhaps the most ambitious megastudy to date is the English Lexicon Project (hereafter, ELP; Balota et al., 2007), for which reading-aloud and lexical decision RTs and accuracy rates were collected for over 40,000 English words.
Assuming that these trends will continue to be popular, then one needs to determine the extent to which item means obtained from megatudies are systematically influenced by the list context (e.g., length of the study, characteristics of the list items). If the study context does systematically influence item responses, then the utility of the megastudy approach would come into question.

Given the implications, it is surprising that not many studies have addressed this issue. Notwithstanding, Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) extracted the monosyllabic words from the ELP that were also utilized by Balota et al. (2004) and found remarkable similarity in the patterns of predictors across studies for both reading-aloud and lexical decision measures. However, the similar contexts between

Correspondence should be addressed to Michael J. Cortese, Department of Psychology, University of Nebraska at Omaha, 6001 Dodge Street, Omaha, NE 68182, USA. E-mail: mcortese@unomaha.edu

studies could still be producing similar systematic patterns in RTs. For example, many orthographic sequences that are repeated within a study may be similar across studies and, thus, have similar effects. Furthermore, in both studies, most words were low frequency, and so the effect of the frequency composition of the stimuli on item responses may also be similar across studies. In a recent megastudy, Keuleers, Diependaele, and Brysbaert (2010) compared lexical decision performance between the first and last sessions and found little difference in the pattern of results.

Although Balota et al. (2004) and Keuleers et al. (2010) found megastudy RT data to be reliable, Sibley, Kello, and Seidenberg (2009) questioned its reliability. Specifically, they compared item RTs from studies that investigated spelling-to-sound consistency and frequency to those obtained from the ELP and three other megastudy datasets. Sibley et al. (2009) concluded that the megastudy data were unreliable because consistency effects across four different megastudy datasets varied. However, Sibley et al. computed item means from raw RTs when *z*-score RTs are more appropriate because different participants responded to different sets of stimuli (for discussion, see Balota, Yap, Hutchison, & Cortese, 2012; Faust, Balota, Spieler, & Ferraro, 1999). Utilizing ELP *z*-score-transformed RTs, Balota et al. (2012) found that all of the original patterns reported in the literature were replicated, and the statistical analyses were mostly replicated. Finally, Adelman, Marquis, Sabatos-DeVito, and Estes (2013) did not find practice effects when subjects read aloud the same 2820 words 50 times each.

We also note that Courrieu and colleagues (e.g., Courrieu, Brand-D'Abrescia, Peereman, Spieler, & Rey, 2011) have examined the reliability of megastudy data via the intraclass correlation coefficient. This technique involves repeatedly selecting data from two sets of *N* randomly selected subjects and computing the average correlation between the two sets. Using this technique, Courrieu and colleagues (2011) estimated that the reproducible variance of the data analysed by Yap and Balota (2009) from the ELP (where 25 subjects contributed to the mean of each item) was .769. This outcome indicates that the ELP data are reliable and highly reproducible. Note that the studies conducted by Courrieu and colleagues do not address whether or not the list context systematically influences the overall pattern of data obtained. It could be that systematic list context effects are part of what is reliable and reproducible.

Despite these results, there are reasons to think that megastudies of reading aloud may be systematically influenced by the list context. For example, in a smaller scale reading-aloud study, Seidenberg, Waters, Barnes, and Tanenhaus (1984) found increased RTs for regular inconsistent words (e.g., *mint*) when their irregular/inconsistent neighbours (e.g., *pint*) preceded them in the experiment. In contrast, irregular words, which had longer RTs than regular inconsistent words, were unaffected by prior presentation of their regular inconsistent neighbours. Many subsequent studies have controlled for this effect by not repeating orthographic rimes in an experiment (e.g., Jared, McRae, & Seidenberg, 1990). However, in megastudies, orthographic rimes are repeated multiple times (e.g., *all* appears in both *ball* and *shall*).

In addition, there are numerous examples of context effects in reading-aloud studies (e.g., Baluch & Besner, 1991; Lupker, Brown, & Colombo, 1997; Zevin & Balota, 2000). While many of these studies have been done in non-English languages, or nonwords were used, the composition of words alone can also produce context effects. For example, Lupker et al. (1997, Experiment 3) found a much larger frequency effect when low- and high-frequency words were presented in separate blocks than when they were mixed within one block.

There are two hypotheses used to explain list context effects in reading aloud. The route selection hypothesis (e.g., Zevin & Balota, 2000) posits that attention can either emphasize or de-emphasize lexical/semantic or sublexical/phonological information. When the context includes mostly irregular/inconsistent words, one may emphasize lexical/semantic information, and when the context includes mostly nonwords, one may emphasize sublexical/phonological information. According to the deadline hypothesis (Lupker et al., 1997), readers may adjust a deadline to respond based on the average difficulty of the list items. When there are mostly difficult items, the deadline will be relatively long, and the reader will respond more slowly to the less difficult items than when there are mostly easy items. Conversely, if the list contains mostly easy-to-process items, the deadline is shortened, and responses to difficult items will be faster than when the list contains mostly difficult items.

Megastudies usually contain a relatively high proportion of lower frequency words. Thus, lexical/semantic access may be typically slow. And since most of the words will be spelling–sound consistent, readers may rely on sublexical/orthographic-to-phonological processing. According to the deadline hypothesis, the average item may be relatively difficult. Thus, the reader will slow down for easier words relative to if easier words were more typical. This characteristic would diminish the effect of any factor that covaries with item difficulty.

The present study examined megastudy context effects via reading-aloud RT and accuracy. The entire stimulus set consisted of 2614 monosyllabic words. A critical subset of 585 words (each with a different orthographic rime) appeared at the beginning of the experiment for one group of 30 subjects or the end of the experiment (i.e., the end condition) for another group of subjects ($n = 30$). Within this subset, critical unique (CU) words ($n = 197$) did not share an orthographic rime with any other word in the entire stimulus set, and critical nonunique (CN, $n = 388$) words had at least one rime neighbour in the stimulus set. Many of the CU words are “strange” words that do not have any orthographic neighbours in English (e.g., *yacht*). For multiple word neighbourhoods, one neighbour was randomly selected to be the critical word.

Based on the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2011), we expected fatigue and general slowing to occur as the experiment progressed. Thus, we predicted slower RTs in the end condition than in the beginning condition. For this prediction, RTs for the CU items were compared to rule out orthographic rime priming. Once this prediction was confirmed, many subsequent analyses were performed on *z*-score RTs. To examine context effects, frequency, length, feedforward rime consistency, and orthographic neighbourhood size effects were compared between the beginning and end conditions for the CN items. These particular variables were selected, in part, because they are powerful predictors of megastudy performance (see, e.g., Cortese & Khanna, 2007), but also because including more variables could diffuse the effects of critical set location across the set of variables, limiting the power to observe an effect for any one variable. A second set of analyses also assessed age of acquisition (AoA), imageability, and onset consistency in addition to the variables assessed in the initial round of analyses. For all the variables considered, a change in their effects would be indicative of a shift toward either lexical/semantic or sublexical processes. Specifically, more emphasis on lexical/semantic information would produce larger frequency, AoA, and imageability effects, whereas a shift to sublexical processing would produce larger feedforward consistency, orthographic neighbourhood size, and length effects. Finding a change in one or more of these predictors of reading aloud would indicate that the megastudy context produces a systematic influence on RTs. Finding no change in these predictors would suggest that context effects in megastudies may be negligible.

In a mega recognition memory study that employed most of these words, Cortese, Khanna, and Hacker (2010) contended that subjects emphasized sublexical processing because many of the words were lower frequency and spelling-to-sound consistent. Given that the critical stimuli contain many strange words, it is unlikely that beginning condition readers would emphasize sublexical/orthographic–phonological processing. However, the end condition readers will have read many low-frequency consistent words first, and they might emphasize sublexical processes for the critical set at the end. If readers shift toward sublexical/orthographic–phonological processing, one would expect smaller frequency, AoA, and imageability effects and larger length, feedforward consistency, and orthographic neighbourhood size effects in the end condition than in the beginning condition. In contrast, a shift toward lexical/semantic processing would produce the opposite pattern. In addition, employing a large set of words and examining these factors expressed as continuous variables provide a powerful test of context effects. No significant change in these variables as a function of condition would indicate that reading-aloud megastudies are reliable and valid.

EXPERIMENTAL STUDY

Method

Participants

Sixty undergraduates (34 females) from the University of Nebraska at Omaha participated for course credit. Their mean age was 22.3 years ($SD = 4.26$), and the mean number of years of education was 13.7 ($SD = 1.28$).

Stimuli

The stimuli consisted of 2614 monosyllabic and monomorphemic words. See Table 1 for item characteristics. Within this corpus, one word belonging to each orthographic rime set was randomly selected to be among the critical set of 585 words. CU words ($n = 197$) had no orthographic rime neighbours in the corpus, and CN words ($n = 388$) each had at least one neighbour. Seventeen homographic words (e.g., *bass*) were in the critical set, but were not the focus of the study and were not part of any analysis reported below.

Measures

Measures were as follows:

Word length (i.e., number of letters).

Subtitle frequency (i.e., log of the Brysbaert & New, 2009, word frequency per million words estimate).

Feedforward consistency (i.e., spelling-to-sound rime and onset consistency values from Kessler, Treiman, & Mullennix, 2008). We calculated consistency values separately for words in our corpus that were not listed in Kessler et al. (2008), using the Zeno, Ivens, Millard, and Duvvuri (1995) norms.

Orthographic neighbourhood size (i.e., Coltheart's *N*; Coltheart, Davelaar, Jonasson, & Besner, 1977). Orthographic *N* refers to the number of words that can be derived from a target word by changing one letter while preserving the identity and position of the other letters in the word.

Equipment

Stimuli were presented on a 19-inch computer monitor that was controlled via a microcomputer running the E Prime software (Schneider, Eschman, & Zuccolotto, 2002).

Procedure

Subjects participated in a reading-aloud task. Each participant received all 2614 items in two 2-hour sessions occurring on separate days within seven days' time, resulting in the presentation of 1307 items in each session. Participants were allowed breaks after each block of (approximately) 150 trials. Half ($n = 30$) of the participants received the critical items at the beginning of their first session, followed by all other items. The other half ($n = 30$) of the participants received the critical items at the end of their second session preceded by all the other items. Therefore, those in the beginning condition received critical and noncritical items in the first session and only noncritical items in the second session. Conversely, those in the end condition received only noncritical items in the first session and more noncritical items followed by critical items in the second session. Within the critical and noncritical sets, the presentation order was random.

On each trial, a fixation mark (+) appeared at the centre of the screen for 1000 ms, followed by a word, which remained there until the voice key registered an acoustic signal. Participants were instructed to name each word as quickly and accurately as possible. Accuracy was coded online by a researcher as correct, incorrect, or noise.

Table 1. *Characteristics of items by category*

				Critical items	
Characteristic	*All items*	*Noncritical items*	*Critical items*	*Nonunique*	*Unique*
n	2614	2029	585	388	197
Length					
M	4.37	4.32	4.52	4.43	4.68
SD	0.90	0.88	0.95	0.92	0.99
Range	2–8	2–8	2–8	2–7	2–8
Mdn	4	4	4	4	5
Rime consistency					
M	0.90	0.89	0.93	0.90	0.97
SD	0.21	0.22	0.17	0.20	0.11
Range	0.04–1.00	0.04–1.00	0.08–1.00	0.08–1.00	0.03–1.0
Mdn	1	1	1	1	1
Frequency					
M	233.49	219.50	281.57	213.86	411.93
SD	1447.54	1202.99	2083.55	1254.32	3110.95
Range	0.02–41,857.12	0.04–29,449.18	0.02–41,857.12	0.04–18,896.31	0.02–41,857.12
Mdn	8.71	9.51	6.90	6.90	6.35
Log frequency					
M	1.18	1.20	1.10	1.09	1.11
SD	0.84	0.84	0.84	0.82	0.87
Range	0.01–4.62	0.02–4.47	0.01–4.62	0.02–4.28	0.01–4.62
Mdn	0.99	1.02	0.89	0.89	0.87

Note: Length is in letters. Rime consistency refers to the feedforward consistency measures of Kessler et al. (2008), based on the Zeno et al. (1995) frequency norms. Frequency refers to the frequency of occurrence per million words based on the subtitle WF (word frequency) norms (Brysbaert & New, 2009). Log frequency is the log of frequency per million words estimate.

Results and discussion

Before the major analyses, the RT data were screened for outliers and voice key errors. First, for each subject, RTs less than 300 ms and greater than 1750 ms were removed (0.15% of the data). Then, RTs greater than and less than 2.5 standard deviations from the subject's mean were eliminated (1.9% of the data). Also, voice key errors were removed (3.9% of the data). In all, this screening process eliminated 6.0% of the RTs. The elimination of data did not differ significantly, for any of the criteria, between beginning and end conditions (all ps $> .18$). RT analyses were conducted on correct responses only.

Analytic strategy

For each measure assessed, one set of analyses was conducted using a two-level multilevel modelling framework in hierarchical linear modelling (HLM) (Raudenbush, Bryk, Cheong, & Congdon, 2000) with words (Level 1) nested within each participant (Level 2). When only the CN set was analysed, the first level was within subjects and consisted of 23,280 (the 388 critical nonunique items × 60 subjects) words. When only the CU set was analysed, the first level was within subjects and consisted of 11,820 (the 197 critical unique items × 60 subjects) words. In both cases, the second level consisted of the between-subjects analyses of the 60 participants. Model building began with an unconditional model without any predictors to demonstrate the proportion of variance within the individual (Level 1) and between individuals (Level 2). An advantage of multilevel modelling is that lower levels (significant or not) effects can vary significantly at higher levels (as measured using a chi-squared test) and that variability can be accounted for. The dependent variables were RT, accuracy, and standardized RT. At Level 1 (differences between types of words), all of the predictor variables were added as a block. At Level 2 (between-group differences), the location in the experiment of the critical set (beginning vs. end) was included in the model. All variables were grand centred and entered as random variables (meaning that we assumed that there would be between-subject variability in the Level 1 predictors). The results are presented in Tables 2 and 3.

In addition, we report separate subjects and items analyses that were conducted on CU word RTs and accuracy when assessing practice and fatigue effects. In both cases, the results of the items analyses (but not the subjects analyses) were different from the HLM multilevel modelling results. Also, regression analyses (see Lorch & Myers, 1990) were performed on each subject's responses individually. A series of *t* tests were performed on the standardized betas obtained for each predictor to compare performance across beginning and end conditions. However, these results never contradicted the HLM multilevel modelling approach, and so they are not reported.

Finally, stepwise multiple regression analyses were conducted on standardized item mean RTs and accuracy levels collapsed across subjects. These analyses considered most of the predictors assessed by Cortese and Khanna (2007). With these items, subjective frequency (Balota, Pilotti, & Cortese, 2001) was highly correlated with log frequency ($r = .969$) and was not included. We also did not assess feedback consistency as it was not considered in any of the initial analyses that examined systematic influences of the megastudy context. We did not include feedback consistency in the initial analyses because a change in its strength and/or direction would not clearly be indicative of more or less emphasis on lexical/semantic or sublexical processes, but, rather, would be reflective of more or less feedback from phonology to orthography. In these analyses, the item means produced by the beginning and end conditions were analysed both separately and combined.

Fatigue and/or practice effects

Practice and fatigue effects were assessed by comparing mean RT and mean accuracy between beginning and end conditions for the CU items. Figures 1a and 1b present the mean RTs and accuracy rates as a function of item type (CN, CU) and critical set location (beginning, end). RTs were

Table 2. *Length, frequency, rime consistency, onset consistency, neighbourhood size, AoA, and imageability on standardized reaction times*

Factor	b	SE	t	*PRPE*	$\Delta\chi^2(1)$
Intercept	0.0020	0.03			
Beg/end	−0.2008	0.06			
Length	0.0676	0.01	5.79***		
Beg/end	−0.0272	0.02	−1.17	1.81	2.15
Frequency	−0.0716	0.01	−6.05***		
Beg/end	−0.0237	0.02	−1.00	6.87	1.03
Rime consistency	−0.2526	0.03	−6.91***		
Beg/end	0.0874	0.07	1.20	16.18	1.49
Onset consistency	0.0297	0.08	0.37		
Beg/end	0.0801	0.16	0.49	6.42	0.24
Neighbourhood size	−0.0057	0.00	−3.61**		
Beg/end	0.0037	0.00	1.19	0.00	0.89
AoA	0.0848	0.01	8.23***		
Beg/end	−0.0096	0.02	−0.46	12.59	0.12
Imageability	−0.0107	0.00	−2.06*		
Beg/end	−0.0000	0.01	−0.00	11.76	−0.01
Level 1 model				1.95	2668.15***

Note: AoA = age of acquisition; beg = beginning. PRPE = proportional reduction in prediction error (in percentages).
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 3. *Length, frequency, rime consistency, onset consistency, neighbourhood size, AoA, and imageability on accuracy*

Factor	b	SE	t	*PRPE*	$\Delta\chi^2(1)$
Intercept	0.9701	.00			
Beg/end	0.0042	.00			
Length	−0.0052	.01	−3.80**		
Beg/end	0.0045	.00	1.69	50	2.04
Frequency	0.0009	.00	0.56		
Beg/end	−0.0045	.00	−1.42	0.00	1.39
Rime consistency	0.0617	.00	8.66***		
Beg/end	−0.0105	.01	−0.74	2.70	0.83
Onset consistency	−0.0276	.00	−2.87**		
Beg/end	−0.0022	.01	−0.12	0.00	0.01
Neighbourhood size	0.0000	.00	0.02		
Beg/end	0.0004	.00	0.79	0.00	0.32
AoA	−0.0171	.00	−7.52***		
Beg/end	−0.0058	.00	−1.29	0.00	2.41
Imageability	0.0005	.00	0.74		
Beg/end	−0.0023	.00	−1.49	0.00	1.35
Level 1 model				1.99	−369.10***

Note: AoA = age of acquisition; beg = beginning. PRPE = proportional reduction in prediction error (in percentages).
*$p < .05$. **$p < .01$. ***$p < .001$.

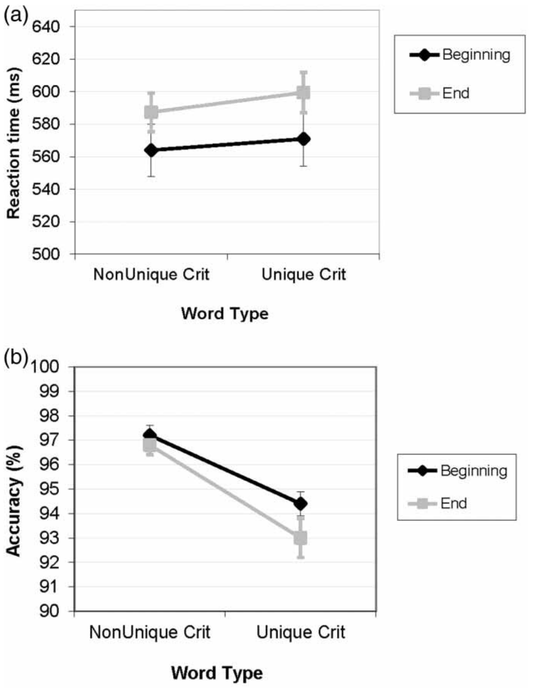

Figure 1. *(a) Mean item reaction time and (b) accuracy measures for nonunique critical (nonunique crit) and unique critical (unique crit) words as a function of critical set location (beginning, end).*

slower for the CU items than for the other item types. In addition, when the CU item RTs were analysed separately, the intraclass correlation for the unconditional model indicated that 1.83% of the variability was at Level 2, and that this variability between subjects was nevertheless statistically significant, $\chi^2(59) = 265.83$, $p < .05$. A second model included critical set location as a predictor variable. The effect of critical set location was not significant ($b = -26.7325$, $SE = 20.69$), $t = -1.292, p > .05$. Including critical set location provided a significantly better prediction of accuracy than using the intercept alone, $\chi^2(14) = 7.27$, $p < .05$, reducing prediction error by 1.74%. And so, the HLM multilevel modelling analyses did not find strong independent evidence for fatigue or practice. However, when analysed separately by subjects and items, RTs for the CU items did not differ significantly as a function of critical set location in the subject analysis, $t(58) = 1.37$, $p = .176$, but did differ significantly in the items analysis, $t(196) = 14.88$, $p < .001$. Because there is evidence for general fatigue at the item level, it would be appropriate to transform RTs into z-score RTs for subsequent items analyses.

Figure 1b shows that accuracy rates were remarkably similar for each item type as a function of beginning/end condition. In the HLM multilevel analysis conducted on accuracy rates for the critical unique item RTs, the intraclass correlation for the unconditional model indicated that 43.86% of the variability was at Level 2, and that this variability between subjects was statistically significant, $\chi^2(59) = 7761.12$, $p < .05$. A second model included critical set location as a predictor variable. The effect of critical set location was not significant ($b = 0.0123$, $SE = .00$), $t = 1.309$, $p > .05$. Including critical set location provided a significantly better prediction of accuracy than using the intercept alone, $\chi^2(1) = 11.38$, $p < .05$, reducing prediction error by 1.99%.

When analysed separately by subjects and items, the subjects analysis did not produce a main effect of critical set location, $t(58) = 1.47$, $p = .147$, whereas the items analysis did produce a main effect of critical set location, $t(196) = 2.96$, $p = .003$, such that CU words in the beginning condition (94.3%) were read slightly more accurately than those in the end condition (93.1%), Note that because the critical items had shorter RTs and greater accuracy, there is no evidence of a speed–accuracy trade-off in the data.

Analyses of critical set location

HLM multilevel analyses were conducted separately for raw RTs, standardized RTs, and accuracy rates for the CN items in which specific predictors were tested as a function of critical set location. The first set of analyses tested length, frequency, feedforward rime consistency, and orthographic *N* as predictor variables, and the second set of analyses also included AoA and imageability in the set of predictors. The results for raw RTs and standardized RTs were very similar, and so only the results on standardized RTs are reported. In addition, reducing the number of variables, as was done in the first round of analyses, did not produce a qualitatively different pattern of effects than the second round that included more variables, and so we only report the results of the second round of analyses.

Results for length, frequency, feedforward rime consistency, feedforward onset consistency, orthographic neighbourhood size, AoA, and imageability as predictor variables

Standardized reaction times (see Table 2). The intraclass correlation for the unconditional model indicated that 43.63% of the variability was at Level 2, and that this variability between subjects was statistically significant, $\chi^2(59) = 1446.43$, $p < .05$. A second model included length, frequency, rime consistency, onset consistency, neighbourhood size, AoA, and imageability as predictor variables. This model showed a significant main effect of length, frequency, rime consistency, neighbourhood size, AoA, and imageability, but not onset consistency. Only the effect of length varied between subjects. Length and AoA were positively associated with RT, and frequency, rime consistency, orthographic *N*, and imageability were negatively associated with RT. Including these variables provided a significantly better prediction of accuracy than using the intercept alone, $\chi^2(35) = 8987.24$, $p < .05$, reducing prediction error by 1.94%.

A final model assessed the effects of critical set location on each predictor variable. Adding critical set location to the effect of length, frequency, rime consistency, onset consistency, orthographic *N*, AoA, and imageability did not improve the estimation of any of these variables and did not reduce the prediction error.

Accuracy (see Table 3). The intraclass correlation for the unconditional model indicated that 1.36% of the variability was at Level 2, and that this variability between subjects was statistically significant, $\chi^2(59) = 363.45$, $p < .05$. A second model included length, frequency, rime consistency, onset consistency, orthographic *N*, AoA, and imageability as predictor variables. This model showed a significant main effect of length, onset consistency, rime consistency, and AoA but not frequency, orthographic *N*, or imageability. The effects of length, frequency, onset consistency, orthographic *N*, and imageability did not vary between subjects, but the effects of rime consistency and AoA did. Longer words, those with more consistent onsets,

Table 4. *Correlation matrix of variables assessed in stepwise regression analyses*

Variable	*1.*	*2.*	*3.*	*4.*	*5.*	*6.*	*7.*
1. Orthographic length		−.280**	.003	−.022	−.633**	.300**	−.023
2. Log frequency			−.121**	−.044*	.190**	−.708**	−.063**
3. Feedforward rime consistency				.017	−.008	.065**	.036
4. Feedforward onset consistency					.107**	−.018	.055**
5. Orthographic neighbourhood size						−.221**	.063**
6. AoA							−.345**
7. Imageability							

Note: $N = 2540$. AoA = age of acquisition.
$^{*}p < .05$. $^{**}p < .01$.

and later acquired words were responded to less accurately, while words with more consistent rimes were responded to more accurately. Including these variables provided a significantly better prediction than using the intercept alone, $\chi^2(35) = -369.10$, $p < .05$, reducing prediction error by 1.99%.

Adding critical set location to the effect of length, frequency, rime consistency, onset consistency, orthographic *N*, AoA, and imageability did not improve the estimation of any of these variables and did not reduce the prediction error.

Stepwise regression analyses on item means for standardized RTs and accuracy

As we found very little evidence for a change as a function of critical set location, it is possible that these null results occurred because the specific predictors were weakly related to performance overall. To address this possibility, we conducted stepwise regression analyses on item means employing most of the standard variables previously examined in megastudies of this type (see, e.g., Cortese & Khanna, 2007). Analyses were performed on 2528 items for which predictor variable values were available. Initial phoneme characteristics were assessed in Step 1, sublexical and lexical variables were assessed in Step 2, and AoA and imageability were assessed in Step 3. The patterns were quite similar between conditions, and so we also analysed the combined data (i.e., from all 60 participants). The correlation matrix for these predictors appears in Table 4. The results of the regression analyses are presented in Figures 2a and 2b.

There are two important things to note. First, initial phoneme characteristics accounted for much less variance than is typical. Specifically, when collapsed across all 60 subjects, initial phoneme characteristics accounted for 11.9% of the variance in standardized RTs. In contrast, Balota et al. (2004) accounted for 35.0% of the variance in RTs. However, this outcome is not problematic for our purposes because the variance attributable to the initial phoneme was similar between beginning and end conditions, and since the same items occurred in each condition, initial phoneme was perfectly controlled. Furthermore, the lack of variance accounted for by the initial phoneme left the variables of interest to fill the void (see Brysbaert & Cortese, 2011). Fortunately, the predictor variables examined in this experiment produced strong effects. Comparing the effects to Cortese and Khanna (2007), one can see that, for every variable except *N*, the standardized betas were larger in our study. Thus, these variables are not showing a change across beginning and end conditions despite accounting overall for a relatively large amount of variance. Thus, there appears to be sufficient power to observe changes in these variables as a function of critical set location, if they existed. Secondly, when collapsed across all subjects, we did observe a small but significant effect of imageability on RTs. This outcome differs from results of Cortese and Khanna (2007) who

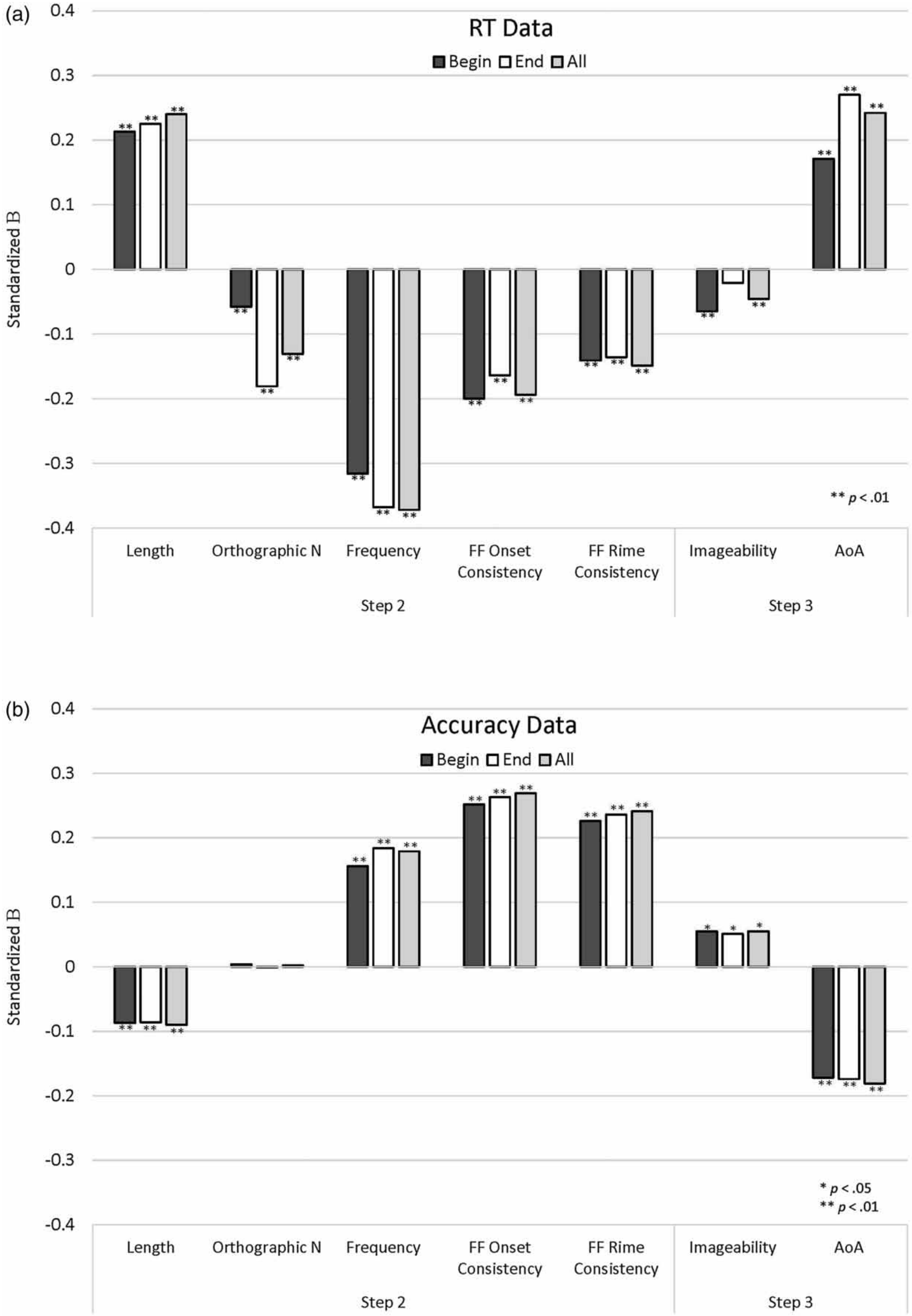

Figure 2. *Standardized beta regression coefficients for the beginning, end, and combined (i.e., all) data. (a) Coefficients for item reaction time (RT) analyses; (b) coefficients for item accuracy analyses (n = 2524).*

found that imageability did not predict RTs when AoA was controlled. More research will need to be done to determine the extent to which imageability affects reading-aloud RTs of monosyllabic words. We note that Connell and Lynott (2014) reported fairly strong effects of imageability on reading-aloud RTs when imageability is specified along visual and auditory dimensions. Finding that imageability affects reading-aloud performance is consistent with models that assume interactivity between semantic and orthographic/phonological levels.

Summary and conclusions

We designed an experiment to provide a strong test of context effects in standard megastudies of reading aloud. One group of subjects read aloud a critical set of items at the beginning of the experiment, and another group of subjects responded to the same set of items at the end of the experiment. We found some evidence for general fatigue, in that items that occurred at the end of the experiment were associated with longer RTs than items located at the beginning of the experiment. However, there were no systematic changes in frequency, rime consistency, onset consistency, word length, orthographic *N*, AoA, and imageability as function of the location of the critical set.

One can take away four important messages from our experiment. First, although there is evidence for general fatigue, fatigue did not interact with any of the most powerful predictors of reading-aloud RT. Second, neither criterion changes nor changes in processing pathways appear to be operating within this context. Third, there appears to be little need for concern for the rime priming effect reported by Seidenberg et al. (1984). Fourth, the general concerns raised by Sibley et al. (2009) do not seem to be justified (also see Balota et al., 2012). Therefore, *z*-score-transformed RTs from megastudies provide a reliable measure of performance that can be used to assess computational models of word recognition.

REFERENCES

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1037–1053. doi:10.1037/a0031829

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. doi:10.1037/0096-3445.133.2.283

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2906 monosyllabic words. *Memory & Cognition*, *29*, 639–647. doi:10.3758/BF03200465

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: Large scale analyses of lexical processes. In Adelman (Ed.), *Visual Word Recognition Vol. 1: Models and Methods, Orthography and Phonology* (pp. 90–115). Hove, UK: Psychology Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. doi:10.3758/BF03193014

Baluch, B., & Besner, D. (1991). Visual word recognition: Evidence for strategic control of lexical and nonlexical routines in oral reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 644–652. doi:10.1037/0278-7393.17.4.644

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of familiarity and age of acquisition survive better word frequency norms? *The Quarterly Journal of Experimental Psychology*, *64*, 545–559. doi:10.1080/17470218.2010.503374

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and imporved word frequency measure for American English. *Behavior Research Methods*, *41* (4), 977–990. doi:10.3758/BRM.41.4.977

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

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256. doi:10.1037//0033-295X.108.1.204

Connell, L., & Lynott, D. (2014). I see/hear what you mean: semantic activation in visual word recognition depends on perceptual attention. *Journal of Experimental Psychology: General*. doi:10.1037/a0034626

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2342 words. *The Quarterly Journal of Experimental Psychology*, *60*(8), 1072–1082. doi:10.1080/174701071315467

Cortese, M. J., Khanna, M. M., & Hacker, S. D. (2010). Recognition memory for 2578 monosyllabic words. *Memory*, *18*(6), 595–609. doi:10.1080/09658211.2010.493892

Courrieu, P., Brand-D'Abrescia, M., Peereman, R., Spieler, D., & Rey, A. (2011). Validated intraclass correlation statistics to test item performance models. *Behavior Research Methods*, *43*, 37–55. doi:10.3758/s13428-010-0020-5

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*, 777–799. doi:10.1037//0033-2909.125.6.777

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, *29*(6), 687–715. doi:10.1016/0749-596X(90)90044-Z

Kessler, B., Treiman, R., & Mullennix, J. (2008). Feedback consistency effects in single-word reading. In E. L. Grigorenko & A. J. Naples (Eds.), *Single-word reading: Behavioral and biological perspectives* (pp. 159–174). Mahwah, NJ: Erlbaum.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, *1*, 174. doi:10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2011). The British lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. doi:10.3758/s13428-011-0118-4

Lorch, Jr., R. F., & Myers, J. L. (1990). Regression analyses of reapeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 149–157. doi: 10.1037/0278-7393.16.1.149

Lupker, S. J., Brown, P., & Colombo, L. (1997). Strategic control in a naming task: Changing routes or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 570–590. doi:10.1037/0278-7393.23.3.570

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*(2), 273–315. doi:10.1037/0033-295X.114.2.273

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115. doi:10.1037/0033-295X.103.1.56

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2000). *HLM5: Hierarchical linear and nonlinear modeling* [Computer Program]. Chicago: Scientific Software International.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime user's guide.* Pittsburgh: Psychology Software Tools Inc.

Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition?. *Journal of Verbal Learning and Verbal Behavior*, *23*(3), 383–404. doi:10.1016/S0022-5371(84)90270-6

Sibley, D. E., Kello, C. T., & Seidenberg, M. S. (2009). *Error, error everywhere: A look at megastudies of word reading*. Proceedings of the Annual Meeting of the Cognitive Science Society. Amsterdam, The Netherlands.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, *8*(6), 411–416. doi:10.1111/j.1467-9280.1997.tb00453.x

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*, 502–529. doi:10.1016/j.jml.2009.02.001

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science.

Zevin, J. D., & Balota, D. A. (2000). Priming and attentional control of lexical and sublexical pathways during naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 121–135. doi: 10.1037//0278-7393.26.1.121