Participants shift response deadlines based on list difficulty during reading-aloud megastudies

Michael J. Cortese¹ · Maya M. Khanna² · Robert Kopp¹ · Jonathan B. Santo¹ · Kailey S. Preston² · Tyler Van Zuiden¹

Published online: 16 February 2017
© Psychonomic Society, Inc. 2017

Abstract We tested the list homogeneity effect in reading aloud (e.g., Lupker, Brown, & Colombo, 1997) using a megastudy paradigm. In each of two conditions, we used 25 blocks of 100 trials. In the random condition, words were selected randomly for each block, whereas in the experimental condition, words were blocked by difficulty (e.g., easy words together, etc.), but the order of the blocks was randomized. We predicted that standard factors (e.g., frequency) would be more predictive of reaction times (RTs) in the blocked than in the random condition, because the range of RTs across the experiment would increase in the blocked condition. Indeed, we found that the standard deviations and ranges of RTs were larger in the blocked than in the random condition. In addition, an examination of items at the difficulty extremes (i.e., very easy vs. very difficult) demonstrated a response bias. In regression analyses, a predictor set of seven sublexical, lexical, and semantic variables accounted for 2.8% more RT variance (and 2.6% more zRT variance) in the blocked than in the random condition. These results indicate that response deadlines apply to megastudies of reading aloud, and that the influence of predictors may be underestimated in megastudies when item presentation is randomized. In addition, the CDP++ model accounted for 0.8% more variance in RTs (1.2% in zRTs) in the blocked than in the random condition. Thus, computational models may have more predictive power on item sets blocked by difficulty than on those presented in random order. The results also indicate that models of word processing need to accommodate response criterion shifts.

Keywords Megastudy · Reading aloud · Response deadline

The megastudy, in which performance measures are collected for a large number of stimuli, has recently become a popular method in cognitive psychology. Perhaps the quintessential example of the megastudy is the English Lexicon Project (ELP; Balota et al., 2007), whereby reading-aloud and lexical-decision reaction time (RT) and accuracy measures were collected for over 40,000 English words. According to the Web of Science, the ELP has been cited 722 times in the literature (1,327, according to Google Scholar). Furthermore, the ELP has motivated lexicon projects in other languages, including British English (Keuleers, Lacey, Rastle, & Brysbaert, 2012), Chinese (Sze, Rickard Liow, & Yap, 2014), French (Ferrand et al., 2010), and more. In addition, the technique has been extended to auditory lexical decision (Ernestus & Cutler, 2015), semantic priming (Hutchison et al., 2013), and recognition memory (Cortese, McCarty, & Schock, 2015), to name just a few of these areas.

The megastudy has become a popular method due to its advantages over the traditional factorial design method (see, e.g., Balota et al., 2004). For example, due to the large number of items examined and the fact that continuous variables are not dichotomized, the megastudy offers considerable power to detect relationships among the variables used to test important theoretical and empirical issues. In addition, where factorial designs often compare performance only in extreme conditions, the megastudy allows for assessing performance across a much larger range of variables. Also, through multiple-regression analyses conducted on the item means, the megastudy affords the opportunity to examine the relative influences of a wide array of factors. As Cutler (1981) pointed out many years ago, in a factorial design it is difficult to control for all factors related to performance while varying only

✉ Michael J. Cortese
mcortese@unomaha.edu

¹ University of Nebraska, Omaha, Nebraska, USA

² Creighton University, Omaha, Nebraska, USA

one or two independent variables (for more thorough discussions of these issues, see Balota et al., 2004; Balota, Yap, Hutchison, & Cortese, 2012).

In a typical megastudy, participants respond to thousands of stimuli, sometimes over two long sessions. Thus, a reasonable concern is that the responses are not as reliable as they would be in briefer testing environments. For example, participants might become overly fatigued and/or practiced, and this may lead to noisy data. Fortunately, the evidence indicates that the data obtained from megastudies are very reliable. Balota et al. (2004) tested the reliability of megastudies by examining participant reading-aloud and lexical-decision performance on a large set of monosyllabic words and comparing the reading-aloud and lexical-decision item means for those same words from the ELP. It is important to note that, whereas Balota et al.'s (2004) participants only responded to monosyllabic words, the ELP participants responded to many mono- and multisyllabic words, so the list contexts differed considerably between the studies. Despite this difference, the patterns of results obtained between studies were highly similar. In addition, Keuleers, Diependaele, and Brysbaert (2010) compared performance between the first and last sessions of a lexical-decision megastudy and found little difference in the patterns of results, despite the fact that participants tended to slow down at the end of the study. More recently, in a megastudy of reading aloud conducted by Cortese, Hacker, Schock, and Santo (2015), participants read aloud 585 critical words, each with a different orthographic rime, at either the beginning or the end of a 2,614-word experiment. Cortese and colleagues found that the RTs for the critical items were longer at the end than at the beginning of the experiment, but there was very little change in the effects of the specific variables (e.g., frequency, spelling-to-sound consistency, length, age of acquisition [AoA], etc.), assessed as a function of the location of the critical set. Finally, we note that via the intraclass correlation coefficient, Courrieu and colleagues (e.g., Courrieu, Brand-D'Abrescia, Peereman, Spieler, & Rey, 2011) demonstrated that the data obtained from the ELP are reliable and highly reproducible. Essentially, this method involves repeatedly taking two random samples of N and computing the correlations between the samples.

However, one related issue that has not been examined within the megastudy context is whether or not participants apply a response criterion that restricts the range of RTs when items are presented randomly. In seminal factorial experiments, Lupker, Brown, and Colombo (1997) demonstrated that the effect on RTs of a variable that is related to item difficulty (e.g., word frequency) was smaller when the presentation order was random than when the items were blocked by difficulty. For example, in their Experiment 3, participants read aloud high-frequency (e.g., *post*) and low-frequency (e.g., *pint*) exception words. In one condition the presentation order was random, and in the other condition the words were

blocked by frequency. Lupker and colleagues found that the frequency effect was much larger in the blocked than in the random condition. To explain these results, they proposed a response deadline hypothesis, wherein participants establish a deadline to respond that is based on the average difficulty of the list items. When easy (e.g., high-frequency) and difficult (e.g., low-frequency) items are mixed within a list, the average difficulty is somewhere in the middle. According to the hypothesis, in this context, participants will speed up for the difficult items and slow down for the easy items, relative to their RTs when these items are presented in different blocks. In fact, Lupker et al. found that participants were 25 ms faster to respond to high-frequency words in the blocked than in the random condition, and 24 ms slower to respond to low-frequency words in the blocked than in the random condition. In other words, the effect of frequency was almost 50 ms larger when items were blocked according to difficulty.

This outcome has potential implications for megastudies of reading aloud in which multiple-regression analyses are used to uncover the influences that certain variables have on performance (e.g., Balota et al., 2004). Perhaps typical megastudies, in which random presentation is employed, underestimate the influences of different factors on RTs. If participants are utilizing a response criterion such as the one hypothesized by Lupker and colleagues, then they are speeding up for the more difficult items and slowing down for the easier items in order to meet the perceived average difficulty of the items. The net effect of this process is to restrict the range of RTs, thus reducing statistical power. Another possibility is that the effect of the response criterion only occurs for items at the extreme levels of difficulty (i.e., very easy and very difficult items). In this case, one would not expect the predictive power of standard variables to be affected much over the thousands of items tested. It is important to determine the extent to which a response criterion affects reading-aloud megastudy performance because, if the range of RTs is restricted in the standard setting, there will be less power to detect the influences of variables on performance, and the relationships that exist may be weakened.

In addition, theoretical models of word processing have been evaluated in their ability to account for item-level variance in RTs by extracting these RTs from megastudies (see, e.g., Perry, Ziegler, & Zorzi, 2010). If the range of RTs is being restricted due to the random presentation of items, then the ability for theoretical models to account for the RT variance is also restricted. Furthermore, if response criterion changes were demonstrated in megastudies, then the models should be developed further to account for those changes.

The following experiment was designed to test for the influence of a response deadline in a megastudy of reading aloud. In the experiment, we presented 25 blocks of 100 trials. In the control condition, the words were ordered randomly, and in the experimental condition, the words were blocked by

difficulty and the order of the blocks was randomized. To vary item difficulty, we took 2,500 words from a recent megastudy by Cortese et al. (2015) and rank-ordered them by their mean RT in that study. This resulted in a blocking pattern in which one block of items contained the 100 items with the fastest RTs, another block contained the items with the next fastest 100 items, and so on. We predicted that, among the item means, the range and standard deviation would be larger in the blocked than in the random condition. We also predicted that when we isolated the extreme conditions (i.e., the easiest vs. the most difficult items), we would show faster RTs for the easiest items in the blocked than in the random condition, and slower RTs for the most difficult items in the blocked than in the random condition. In addition, we assessed frequency, length, feedforward onset consistency, feedforward rime consistency, orthographic neighborhood size, AoA, and imageability as predictor variables. We chose to limit the set of variables to these because we had predictor values for each of these variable for all 2,500 words, and we wanted to avoid diffusing the effects of difficulty across a larger set of variables. We predicted that the predictor variables in our statistical model would generally show stronger relationships with RTs in the blocked than in the random condition, because the range of RTs would be greater in the blocked than in the random condition. Finally, assuming that the results were consistent with these predictions, we planned to correlate RTs from the CDP++ model (Perry et al., 2010) with human RTs for each of our conditions. This particular model was selected on the basis of the recent success it has had in accounting for item-level variance in RTs (see Perry et al., 2010). We predicted that the model would account for more variance in the blocked than in the random condition, due to the extended range of RTs associated with the predictor variables in that condition. Furthermore, we assumed that the CDP++ model would also show similar stronger correlations between model

and human RTs in the blocked than in the random condition, to the extent that it was also sensitive to these predictor variables.

Method

Participants

A total of 64 undergraduates (48 females, 16 males) participated in the experiment. Forty of the participants were students from the University of Nebraska at Omaha, and 20 were students from Creighton University. Three of the participants did not provide age or education level information, and six did not report their ethnicity. For the remaining participants, the mean age was 21.4 years (range = 18–36) of age, and the mean education level was 2.48 (where 1 = *freshman* and 4 = *senior*). Forty-two participants identified themselves as White or Caucasian, five identified themselves as Black or African American, four identified themselves as Hispanic or Latino/a, four identified themselves as Asian, two identified themselves as Pacific Islanders, and one other participant identified him or herself as Egyptian.

Materials

We presented 2,500 monosyllabic words in the experiment. The stimulus characteristics for these words can be found in Table 1.

In the blocked condition, 25 lists of 100 items were created based on the mean item RTs from Cortese et al. (2015). Those item means were generated from 60 participants. We rank-ordered the items from the fastest to the slowest RT, and created 100 word lists based on this ordering. For example, the

Table 1 Stimulus characteristics for the words used in the experiment

	Mean	Standard deviation	Range
Orthographic length	4.37	0.90	2 to 8
Frequency	222.12 (Median = 8.63)	1,350.80	.02 to 41,857.12
Rime consistency	.90	.20	.03 to 1.0
Onset consistency	.98	.10	.01 to 1.0
Age of acquisition	4.00	1.01	1.56 to 6.63
Imageability	4.36	1.39	1.20 to 6.90
Orthographic <i>N</i>	6.74	5.12	0 to 25
Phonological <i>N</i>	15.10	9.90	0 to 48
Levenshtein OD	1.55	0.32	1 to 2.85
Levenshtein PD	1.32	0.31	1 to 2.75

Frequency values are per-million estimates taken from Brysbaert and New (2009). OD refers to orthographic distance, and PD refers to phonological distance (Yarkoni, Balota, & Yap, 2008).

“easiest” list consisted of the 100 items with the fastest RTs, the next “easiest” list consisted of the 100 items with the next fastest RTs, and so on. The stimulus characteristics for each of the 25 lists can be found in Table 2.

For the random condition, items were again presented in lists of 100 items, but the membership of items within a list was created randomly, with no regard to the word characteristics or the RTs produced for the words in Cortese et al. (2015). Thirty-six other words were used as practice stimuli. For the blocked condition, we created three practice lists of 12 words each. These words were drawn from Cortese et al. (2015) and varied in their average RT. Of these 36 words, the 12 with the fastest RTs were placed in one list, those with the next fastest RTs were placed in a second list, and those words with the slowest RTs were placed in a third list. The same practice stimuli were used in the random condition, but the three blocks of 12 words were all drawn randomly from a single list of 36 items.

Predictors

Word length *Word length* referred to the number of letters a word contains.

Subtitle frequency This measure refers to the frequency with which a word appears in the written text of film subtitles (Brysbaert & New, 2009). For our analyses, we used the log of the word frequency per million words, taken from Brysbaert and New (2009).

Feedforward consistency *Feedforward consistency* refers to orthographic-to-phonological consistency (see Balota et al., 2004). We used the values from Kessler et al. (2008), which were based on the Zeno frequency norms (Zeno, Ivens, Millard, & Duvvuri, 1995). We analyzed both feedforward rime consistency and feedforward onset consistency (the orthographic-to-phonological consistency of the initial letter or letter cluster preceding the vowel).

Table 2 Stimulus characteristics (with standard deviations in parentheses) for the words used in the experiment as a function of difficulty level, with Level 1 being the least difficult or fastest in Cortese et al. (2015)

Difficulty	Frequency	Length	AoA	N	FFR	FFO	Imageability
1	64.4 (2,514.5)	3.8 (0.9)	3.1 (0.7)	8.4 (4.6)	.92 (.16)	.99 (.02)	4.5 (1.6)
2	44.7 (2,180.5)	3.9 (0.9)	3.2 (0.7)	9.1 (5.3)	.94 (.15)	.98 (.09)	4.7 (1.5)
3	33.8 (1,759.9)	4.0 (0.9)	3.5 (0.9)	9.6 (5.3)	.92 (.20)	.99 (.02)	4.3 (1.5)
4	31.9 (4,199.8)	3.9 (0.8)	3.4 (0.8)	9.4 (5.2)	.93 (.15)	.98 (.09)	4.5 (1.5)
5	18.9 (1,481.2)	4.1 (1.0)	3.5 (0.8)	8.4 (5.3)	.91 (.18)	.98 (.05)	4.5 (1.4)
6	18.5 (795.1)	4.2 (0.9)	3.6 (0.9)	8.7 (5.3)	.92 (.22)	.99 (.03)	4.8 (1.4)
7	17.9 (553.0)	4.2 (0.9)	3.7 (0.9)	8.0 (5.2)	.91 (.20)	.98 (.09)	4.5 (1.4)
8	10.9 (660.2)	4.2 (0.8)	3.8 (0.9)	7.9 (5.1)	.91 (.19)	.99 (.02)	4.5 (1.3)
9	12.5 (367.2)	4.3 (0.8)	3.8 (0.8)	7.4 (5.0)	.95 (.11)	.99 (.04)	4.5 (1.4)
10	16.9 (1,189.1)	4.2 (0.8)	3.7 (0.8)	8.8 (5.1)	.90 (.20)	.97 (.13)	4.8 (1.4)
11	7.4 (319.8)	4.1 (0.7)	4.0 (1.0)	7.8 (5.0)	.91 (.20)	.99 (.02)	4.4 (1.3)
12	8.8 (259.6)	4.3 (0.8)	3.9 (0.9)	7.3 (4.8)	.95 (.13)	.99 (.03)	4.4 (1.4)
13	10.2 (1,074.0)	4.4 (0.8)	4.1 (1.0)	6.3 (4.8)	.88 (.23)	.98 (.11)	4.4 (1.4)
14	10.9 (733.8)	4.4 (0.7)	4.0 (0.9)	6.3 (5.1)	.93 (.16)	.98 (.08)	4.4 (1.5)
15	7.3 (1,730.0)	4.4 (0.8)	3.9 (0.9)	6.9 (5.5)	.91 (.21)	.96 (.14)	4.5 (1.4)
16	6.1 (1,357.2)	4.4 (0.8)	4.2 (0.9)	6.7 (4.7)	.89 (.23)	.96 (.13)	4.2 (1.4)
17	4.7 (590.8)	4.5 (0.9)	4.2 (1.0)	6.8 (4.9)	.88 (.21)	.98 (.08)	4.3 (1.3)
18	4.3 (215.6)	4.4 (0.9)	4.2 (0.9)	5.7 (4.2)	.89 (.20)	.98 (.12)	4.3 (1.3)
19	4.7 (98.7)	4.6 (0.8)	4.3 (0.9)	5.6 (4.5)	.91 (.20)	.98 (.09)	4.2 (1.3)
20	3.3 (333.5)	4.6 (0.8)	4.4 (0.9)	5.5 (4.3)	.91 (.20)	.97 (.08)	4.2 (1.2)
21	3.9 (885.2)	4.7 (0.8)	4.5 (1.0)	4.5 (4.2)	.93 (.15)	.96 (.16)	4.0 (1.2)
22	2.5 (169.1)	4.8 (0.9)	4.5 (1.0)	4.1 (4.1)	.88 (.24)	.98 (.08)	4.4 (1.3)
23	2.7 (281.6)	5.0 (0.8)	4.6 (0.9)	3.8 (3.8)	.87 (.24)	.99 (.06)	4.1 (1.3)
24	1.9 (149.6)	5.0 (0.9)	4.9 (1.0)	3.7 (3.7)	.87 (.24)	.97 (.14)	3.9 (1.3)
25	1.4 (74.0)	5.2 (0.9)	5.0 (1.0)	2.4 (3.2)	.78 (.30)	.91 (.25)	3.9 (1.4)

Frequency values are median values of the per million estimates taken from Brysbaert and New (2009). AoA = age of acquisition, N = orthographic neighborhood size, FFO = feedforward onset consistency, FFR = feedforward rime consistency.

For words in our experiment that did not have a consistency value in the Kessler et al. (2008) database, we calculated their values using the Zeno et al. frequency norms.

Orthographic neighborhood size (i.e., Coltheart's N: Coltheart, Davelaar, Jonasson, & Besner, 1977) *N* refers to the number of words that can be constructed from a target word by changing a single letter while maintaining the identity and position of the word's other letters. For example, *ball* has the neighbors *call*, *bill*, *bale*, and so forth.

Age of acquisition (AoA) AoA represents an estimate of the age at which a word was first learned. In our analyses, we used the Cortese and Khanna (2008) norms.

Imageability *Imageability* refers to the ease or difficulty one has in generating a mental image of the referent of a word. In our analyses, we employed the values obtained by Cortese and Fugett (2004).

Equipment

The stimuli appeared on a 19-in. computer monitor that was controlled by a microcomputer running the E-Prime Professional software (Schneider, Eschman, & Zuccolotto, 2002).

Procedure

Participants read aloud 25 blocks of 100 words in a single session that lasted about 2 h. In the random condition, the assignment of words to each block was random, and in the blocked condition, words appeared as a predetermined set based on our blocking procedure. In this condition, the blocks were ordered randomly, and the words within a block were presented in random order, but the words within each block were always the same 100 words. Prior to the experiment, participants were instructed to name each word as quickly and accurately as possible. After the instructions were read, participants performed three blocks of 12 practice trials. In the blocked condition, the blocks varied by difficulty. The easiest block occurred first, and the most difficult block occurred third. In the random condition, the order of words was random. On each trial, a fixation mark (+) appeared in the center of the screen for 300 ms and was followed by a blank screen for 200 ms. The to-be-named word followed and remained in view until the voice key had registered the onset of the acoustic signal. Accuracy was then coded by a researcher as correct, incorrect, or noise. A 700-ms blank screen separated trials.

Results

The data from two participants were discarded because their error rates exceeded 20%. Before the major analyses, the RT data were screened for outliers and voice key errors. For each participant, an initial screening removed voice key errors and RTs less than 300 ms and greater than 1,250 ms (1.5% in the blocked condition, and 1.2% in the random condition). Then, RTs beyond three standard deviations from the participant's mean were eliminated (1.5% in the blocked condition, and 1.6% in the random condition). In all, this eliminated 2.9% of the RT data. Finally, the RT analyses were conducted on correct responses only.

Table 3 provides the mean RTs, the ranges and standard deviations of the RTs, and the mean accuracies for items, as a function of stimulus presentation condition.

We note that participants were generally faster and slightly less accurate in the blocked than in the random condition. However, although participants in the blocked condition were generally faster than those in the random condition, they were slower at the higher end of list difficulty than those in the randomized condition. We also note that although participants were slightly less accurate in the blocked than in the randomized condition, they were less accurate at the extreme difficulty levels, at which their RTs were either faster or slower than the randomized condition, indicating that their faster RT was not simply a function of a speed–accuracy trade-off. Finally, it is important to note that, consistent with our prediction, the ranges and standard deviations among the item mean RTs were larger in the blocked than in the random condition.

Analytic strategy

One set of analyses was conducted using a two-level multilevel modeling framework (in HLM, ver. 7.20; Raudenbush, Bryk, Cheong, & Congdon, 2000) with sets of words (Level 1) nested within each participant (Level 2). Multilevel modeling is a variation of traditional multiple regression that serves to address the violation of independence inherent with nested data, as in the present study. Given that combinations of words were presented to participants in either random or blocked fashion, we could not assume that the variability across sets

Table 3 Mean reaction times (RTs) in milliseconds, range, standard deviation, and accuracy percentages as a function of list context

	Random	Blocked
Mean RT	515.0	508.7
Item RT range	446.9 to 664.5	423.2 to 678.0
Item RT standard deviation	26.3	35.0
Item accuracy	97.4	96.5
Item accuracy range	64.5 to 1.0	58.1 to 1.0

was random. Multilevel modeling allowed us to nest the various combinations of words within each individual to account for the proportion of differences within participants (i.e., across sets), as compared to the variability that exists between participants (as a function of condition). Multilevel modeling analyses were conducted separately for RTs and accuracy rates, for the items in which our specific predictors were tested as a function of presentation condition. Multilevel modeling analyses began first with an unconditional model (containing only the dependent variable), so that the proportion of variability between participants could be compared to the variability within participants (based on the intraclass correlation), thus providing a justification for the analytic approach. This unconditional model also served as the baseline model to which subsequent models were compared. The models were built up by adding predictors and testing for significant improvements (using a χ^2 test) and proportional reductions in prediction error (interpreted as a pseudo- R^2).

Again, the dependent variables were RT and accuracy. We conducted an initial analysis on RTs to determine whether a response criterion was operating at the extreme levels of difficulty within our dataset. For this analysis, only the RTs for the 100 words at the least difficult level and the 100 words at the most difficult level were analyzed. For this analysis, the first level was analyzed within participants, and consisted of 12,400 (200 words $\times$ 62 participants) monosyllabic words. The second level consisted of the 62 participants who read aloud each word. Our model building began with an unconditional model without any predictors to demonstrate the proportions of variance both within (Level 1) and between (Level 2) individuals. An analogous analysis was conducted on accuracy rates. One of the advantages of multilevel modeling is that the effects at lower levels (significant or not) can vary significantly at higher levels (as measured using a chi-square test), so variability can be accounted for. At Level 1 (differences between types of words), hypothesis testing began with level of difficulty. At Level 2 (between-group differences), the group variable of whether the stimuli were presented randomly or in a blocked fashion was included in the model. The Level-2 model also tested the interaction between group and level of difficulty.

For the analyses on the complete set of 2,500 words, the first level was performed within participants and consisted of 155,000 (the 2,500 words $\times$ 62 participants) monosyllabic words. The second level consisted of the between-participants analyses of the 62 participants who named each word. Our model building began with an unconditional model without any predictors, to demonstrate the proportions of variance within the individual (Level 1) and between (Level 2) individuals. At Level 1 (differences between types of words), hypothesis testing began with the initial-phoneme variables added as a block. Then length, log frequency, feedforward rime consistency, feedforward onset consistency, and

orthographic neighborhood size were added to the model. In the next step, AoA and imageability were assessed. At Level 2 (between-group differences), the group variable of whether the stimuli were presented randomly or in a blocked fashion was included in the model. The Level-2 model also tested the interaction between group and each of our (sub)lexical and semantic predictor variables. All variables were grand-mean centered and entered at random (meaning that we assumed that there would be between-participants variability in the Level-1 predictors).

Finally, we conducted stepwise multiple-regression analyses on the item mean RTs, standardized item mean RTs (i.e., zRT), and accuracy levels collapsed across participants. These analyses included initial-phoneme characteristics in the first step, length, log frequency, feedforward rime consistency, feedforward onset consistency, and orthographic neighborhood size in the second step, and AoA and imageability in the final step. These results indicated that the relationships between a standard set of predictor variables and RTs (and zRT s) were generally stronger in the blocked than in the random condition.

Analyses of presentation condition (random vs. blocked)

Multilevel analyses were conducted separately for RTs and accuracy rates, for the items in which specific predictors were tested as a function of presentation condition. In the first set of analyses, we examined the two extreme levels of difficulty, to test the prediction that response biases would be present at these extreme levels. If so, we would expect a crossover interaction between presentation condition and level of difficulty. The item mean RTs and accuracy for these conditions are presented in Table 4.

In subsequent analyses, we focused on length, log frequency, feedforward rime consistency, feedforward onset consistency, N , AoA, and imageability.

Results for the extreme levels of difficulty

Reaction times (see Table 4) The intraclass correlation for the unconditional model indicated that 64.09% of the variability in RTs was at the between-participants level (inversely, 35.91% of the variability was at the within-participants level),

Table 4 Reaction times in milliseconds (with accuracy rates in parentheses) for the least difficult and most difficult conditions, as a function of the list context

Difficulty	Random	Blocked
Lower	483.4 (99.1)	462.2 (98.4)
Higher	563.6 (91.1)	583.3 (89.3)
Difficulty effect	80.2 (8.0)	121.1 (9.1)

and that this variability in RTs between participants was statistically significant [$\chi^2(61) = 80,789.62, p < .001$]. A second model included level of difficulty. This model showed a significant main effect of level of difficulty ($b = 49.65, SE = 2.99, t(61) = 16.63, p < .001$, which improved the model's prediction of RTs [$\Delta\chi^2(2) = 1,579,029.52, p < .001$] over the intercept alone, with a proportional reduction in prediction error (PRPE = 11.89%). Thus, more difficult words took longer to read aloud than less difficult words. We observed no main effect of presentation condition ($b = -0.42, SE = 15.78, t(60) = 0.03, p = .979$, but there was a significant interaction between level of difficulty and presentation condition ($b = 20.94, SE = 5.35, t(60) = 3.92, p < .001$, that improved the modeling of the effect of difficulty on RTs [$\Delta\chi^2(1) = 175.56, p < .001$] and reduced prediction error by 19.77%. In other words, the difference in RTs as a function of difficulty was significantly larger in the blocked than in the random condition.

Accuracy (see Table 4) The intraclass correlation for the unconditional model indicated that 2.01% of the variability in accuracy rates occurred between participants (again, conversely, 97.99% of the variability was at the within-participants level), and that this variability in accuracy rates between participants was statistically significant [$\chi^2(61) = 3,202.35, p < .001$]. A second model included level of difficulty. This model showed a significant main effect of level of difficulty, reducing accuracy ($b = -0.04, SE < 0.001, t(61) = 8.42, p < .001$, which improved the model's prediction of accuracy [$\chi^2(2) = 107,176.78, p < .001$] relative to the unconditional model. Neither the main effect of group nor the group by level of difficulty interaction was significant, both $ps > .25$.

Results for length, log frequency, feedforward rime consistency, feedforward onset consistency, and neighborhood size as predictor variables

Reaction times (see Table 5)

At Level 1 (differences between types of words), hypothesis testing began with the initial-phoneme variables added as a block. Then length, log frequency, feedforward rime consistency, feedforward onset consistency, and orthographic neighborhood size were added to the model. In the next step, AoA and imageability were assessed. As can be seen in Table 5, all of the predictor variables except imageability were significantly related to RTs. Including these variables provided a significantly better prediction than did using the intercept alone [$\chi^2(35) = 12,099.33, p < .001$], reducing prediction error by 5.63%. Most important were the tests that examined the interaction between presentation condition and each of our (sub)lexical and semantic predictor variables. In this case, we found that although the interaction between frequency and presentation style was only marginally significant, $t(60)$

Table 5 Effects of length, frequency, rime consistency, onset consistency, neighborhood size, AoA, and imageability on reaction times

	<i>b</i> (<i>SE</i>)	<i>t</i> (<i>df</i>)	PRPE	$\Delta\chi^2(1)$
Intercept	522.22 (12.63)	41.36 (60)*		
Random/blocked	-2.40 (19.70)	0.12 (60)	0.00%	0.04
Length	4.17 (0.82)	5.07 (60)*		
Random/blocked	1.05 (1.61)	0.65 (60)	0.00%	2.57
Frequency	-5.77 (0.70)	8.26 (60)*		
Random/blocked	-2.08 (1.08)	1.93 (60)	7.62%	7.86*
Rime consistency	-13.61 (2.49)	5.46 (60)*		
Random/blocked	-4.62 (3.01)	1.53 (60)	3.62%	4.48*
Onset consistency	-16.10 (4.07)	3.95 (60)*		
Random/blocked	-7.52 (5.32)	1.41 (60)	7.00%	2.91
Neighborhood size	-0.76 (0.10)	7.40 (60)*		
Random/blocked	-0.39 (0.15)	2.54 (60)*	14.66%	12.25*
AoA	4.90 (0.72)	6.77 (60)*		
Random/blocked	2.31 (1.16)	1.99 (60)*	6.58%	9.79*
Imageability	-0.04 (0.23)	0.17 (60)		
Random/blocked	0.40 (0.32)	1.26 (60)	11.24%	1.17

* $p < .05$

$= -1.93, p = .06$, including this term did improve the fit of the model [$\chi^2(1) = 7.86, p < .01$; PRPE = 7.62%]. The same was true for AoA, $t(60) = 1.99, p = .051$ [$\chi^2(1) = 9.79, p < .01$; PRPE = 6.58%]. Orthographic *N* interacted with presentation style and improved the fit of the model, $t(60) = -2.54, p = .014$ [$\chi^2(1) = 12.25, p < .001$; PRPE = 14.66%]. Although the effect of rime consistency was not significant, $t(60) = -1.53, p = .13$, it also improved the model [$\chi^2(1) = 4.48, p < .05$; PRPE = 3.62%]. No other interactions approached significance, all $ps > .13$.

Accuracy As can be seen in Table 6, all of the predictor variables except length were significantly related to accuracy. Including these variables provided a significantly better prediction than did using the intercept alone [$\chi^2(35) = 1,217.70, p < .001$], reducing prediction error by 1.21%. In the accuracy data, we found no significant interactions between presentation group and any of the (sub)lexical or semantic predictor variables, all $ps > .14$ (PRPEs = 0.00%).

Stepwise regression analyses on item means for RTs, zRTs, and accuracy

The correlation matrix for the predictor variables used in this experiment is presented in Table 7. The results of the regression analyses are presented in Tables 8 and 9.

The results also demonstrate that the relationships between the (sub)lexical and semantic predictor set and RTs are stronger in the blocked than in the random condition. The (sub)lexical + semantic predictor set account for 2.8% more

Table 6 Effects of length, frequency, rime consistency, onset consistency, neighborhood size, AoA, and imageability on accuracy

	<i>b</i> (<i>SE</i>)	<i>t</i> (<i>df</i>)	PRPE	$\Delta\chi^2(1)$
Intercept	0.89 (0.02)	50.21 (60)*		
Random/blocked	−0.02 (0.03)	0.58 (60)	0.00%	0.91
Length	−0.00 (0.00)	1.61 (60)		
Random/blocked	0.00 (0.00)	1.52 (60)	0.00%	3.36
Frequency	0.00 (0.00)	4.49 (60)*		
Random/blocked	−0.00 (0.00)	−0.08 (60)	0.00%	0.01
Rime consistency	0.04 (0.01)	5.28 (60)*		
Random/blocked	−0.00 (0.01)	0.03 (60)	0.00%	0
Onset consistency	0.04 (0.01)	4.08 (60)*		
Random/blocked	0.01 (0.01)	0.57 (60)	0.00%	0.79
Neighborhood size	0.00 (0.00)	2.51 (60)*		
Random/blocked	0.00 (0.00)	0.17 (60)	0.00%	0.04
AoA	−0.00 (0.00)	2.32 (60)*		
Random/blocked	−0.00 (0.00)	1.47 (60)	0.00%	3.96*
Imageability	0.00 (0.00)	3.19 (60)*		
Random/blocked	0.00 (0.00)	0.05 (60)	0.00%	0.00

* $p < .05$

variance in raw RTs, and 2.6% more variance in zRTs, in the blocked than in the random condition. For accuracy, the predictor set accounted for a small amount of variance overall. In general, all of the (sub)lexical and semantic variables were significantly related to RTs, except length. Imageability was significantly related to accuracy in the random but not the blocked condition, and AoA was marginally related to accuracy in the random condition, and significantly related in the blocked condition.

Analysis of CDP++ model

We also predicted that, to the extent to which models of word processing are sensitive to the variables associated with RTs, they should account for more variance in the blocked than in the random condition. We tested this prediction with the CDP++ model (Perry et al., 2010). We

Table 8 Standardized regression coefficients, percentage reductions in prediction error (PRPE), and R^2 , for RTs and zRTs

Predictor variable	Random	PRPE	Blocked	PRPE
RT				
Step 1				
Initial phoneme				
R^2	.094		.084	
Step 2				
Length	.160**	1.80%	.158**	1.80%
Neighborhood size	−.164**	2.10%	−.179**	2.50%
Subtitle frequency	−.339**	13.310%	−.359**	17.24%
FF onset consistency	−.080**	0.70%	−.092**	1.00%
FF rime consistency	−.112**	1.80%	−.121**	2.10%
R^2	.329		.345	
Step 3				
Imageability	.005	0.00%	.011	0.00%
AoA	.199**	2.00%	.217**	2.50%
R^2	.347		.365	
zRT				
Step 1				
Initial phoneme				
R^2	.095		.100	
Step 2				
Length	.181**	2.30%	.179**	2.40%
Neighborhood size	−.166**	2.20%	−.178**	2.60%
Subtitle frequency	−.358**	15.10%	−.370**	16.50%
FF onset consistency	−.064**	0.50%	−.086**	0.90%
FF rime consistency	−.098**	1.40%	−.122**	2.30%
R^2	.355		.382	
Step 3				
Imageability	−.003	0.00%	.010	0.00%
AoA	.196**	2.10%	.229**	3.00%
R^2	.374		.405	

** $p < .01$

chose this model because prior research had demonstrated that it typically accounts for substantially more variance in RTs than do other contemporary models, and thus

Table 7 Correlation matrix of variables assessed in the stepwise regression analyses

	1.	2.	3.	4.	5.	6.	7.
1. Orthographic length		−.276**	.008	−.027	−.633**	.295**	−.017
2. Log frequency			−.122**	−.059**	.186**	−.706**	−.063**
3. Feedforward rime consistency				.012	−.011	.068**	.032
4. Feedforward onset consistency					.106**	.006	.055**
5. Orthographic neighborhood size						−.217**	.059**
6. AoA							−.341**
7. Imageability							

 $N = 2,500$, ** $p < .01$

Table 9 Standardized regression coefficients, percentage reductions in prediction error (PRPE), and R^2 , for accuracy

Predictor variable	Random	PRPE	Blocked	PRPE
Step 1				
Initial phoneme				
R^2	.005		.008	
Step 2				
Length	-.043	0.10%	.003	0.00%
Neighborhood size	.057*	0.20%	.060*	0.20%
Subtitle frequency	.153**	2.20%	.163**	2.70%
FF onset consistency	.106**	0.40%	.115**	1.40%
FF rime consistency	.215**	4.70%	.174**	3.30%
R^2	.084		.071	
Step 3				
Imageability	.044	0.20%	.034	0.10%
AoA	-.068**	0.10%	-.116**	0.50%
R^2	.090		.081	

* $p < .05$, ** $p < .01$

would provide a good test for our prediction (see Perry et al., 2010). We obtained RTs from the model for the 2,500 words in the experiment, and then eliminated RTs associated with incorrect codes. We also eliminated words that are considered irregular in British but not in American English, and vice versa (e.g., *vase*, *aunt*). This left us with RTs for 2,393 words. For this set of RTs, the CDP++ model accounted for 12.0% and 13.0% of the variance in RTs and zRTs, respectively, in the blocked condition, and 10.6% and 11.8% of the variance in RTs and zRTs in the random condition. When we entered the initial-phoneme characteristics in Step 1 and the model RTs in Step 2, the CDP++ model accounted for an additional 11.5% and 12.8% of the variance in RTs and zRTs, respectively, in the blocked condition, and 10.9% and 11.6% of the variance in RTs and zRTs in the random condition. Obviously, the overall change in the amount of variance accounted for in the blocked versus the random condition is greater than the change in the overall amount of variance accounted for by the CDP++ model across the two conditions. We think that this outcome can best be explained in terms of the individual factors that show a change across the two conditions and how sensitive the model is to these factors. Specifically, the results demonstrated that the blocked condition showed larger effects for frequency, AoA, orthographic neighborhood size, and feedforward consistency. Because the CDP++ model is sensitive to frequency and consistency, but not necessarily to N (see Perry, Ziegler, & Zorzi, 2007), and probably not to AoA (see Cortese & Schock, 2013), this modest change in its predictive power across conditions is not particularly surprising.

Discussion

We have shown strong evidence that the list homogenization effect identified by Lupker et al. (1997) applies to megastudies of reading aloud and affects the predictive power of the CDP++ model. First, across the item means, the range and standard deviation were larger in the blocked than in the random condition. This outcome confirmed the hypothesis that randomized presentation of stimuli restricts the range of RTs relative to blocking items by difficulty. Second, when comparing the extreme levels of difficulty across random versus blocked presentation, the level-of-difficulty effect was much larger in the blocked than in the random condition. Third, sublexical, lexical, and semantic variables generally showed stronger relationships to RTs in the blocked than in the random condition. More specifically, orthographic neighborhood size, frequency, feedforward rime consistency, and AoA were more strongly related to performance in the blocked than in the random condition. Since virtually all megastudies of reading aloud have employed random presentation, their power to uncover the relationships that variables have with RTs likely has been reduced. Finally, we demonstrated that the CDP++ model accounted for slightly more variance in RTs in the blocked than in the random condition.

The results of this experiment lead to questions about how best to represent item mean RTs in megastudies of reading aloud, as well as what type of presentation schedule should be utilized in future megastudies of reading aloud. Our view is that blocking items by difficulty may be preferable to randomized presentation, especially when a new variable is being examined. On the basis of our results, blocking should increase the power to uncover new relationships, as well as strengthen already-known relationships between predictor variables and RTs. However, because we utilized a somewhat crude method of equating difficulty with RT, it is not clear that our method of blocking stimuli on the basis of item RT means from another study (Cortese et al., 2015) was optimal. In fact, looking at Fig. 1, one can see that, although the overall range of RTs is extended in the blocked condition, there is

Fig. 1 Reading-aloud reaction times (RTs) for the random and blocked conditions as a function of list difficulty.

considerably more variability among the items as a function of the level of difficulty, as we defined it.

One might prefer a function that was smoother, like the one seen for random presentation, but with the extended range seen in our blocked condition. One may ask why the data in the blocked condition appear so variable, list by list. One possibility is that, although RTs may be strongly correlated with item difficulty, they may also be influenced by other factors not related to difficulty. Although we do not think that these factors would differentially affect the conditions of our experiment, they might add variability to the RTs that is unrelated to item difficulty. For example, the RT measure is affected by voice key artifacts that may have more to do with the sensitivity of the measurement device than the difficulty of the items (Rastle & Davis, 2002). In addition, one would expect accuracy also to be related to item difficulty, yet the correlation between RT and accuracy is moderate at best ($r = -.34$ for items in the randomized condition, and $r = -.41$ for items in the blocked condition). Therefore, using only RTs to represent item difficulty may not be the best way to organize the stimuli in the blocked condition. One possibility would be to define level of difficulty after voice key factors are removed, and also to take accuracy levels into consideration.

In addition, we note that some of the list-to-list variability in RTs associated with the blocked condition may be associated with the randomization of block order. Thus, a particular block of words would be preceded by a block of very easy words for one participant, and by a block of more difficult words for another participant. The only way for the participant to determine the list difficulty would be to proceed through the list. Perhaps one would want to modify the experiment to include information about the difficulty of the items in any upcoming block of stimuli. For example, one might say something like “On the basis of previous research, the following list of words are typically perceived to be very easy/easy/moderately easy/moderately difficult/difficult/very difficult.” However, anything that we suggest about blocking procedures at this point would be purely speculative. More research will need to be done to identify a more optimal method of blocking that would lead to a smoother function between item difficulty and RT.

Finally, it is important to note that although the overall R^2 differences in the blocked versus the random condition (2.8% in RTs, 2.6% in zRTs) may, on the surface, seem somewhat small, they are actually quite substantial. As a comparison, consider that adding AoA in the final step of a regression model, as Cortese and Khanna (2008) did, increased R^2 by less than 0.5% above and beyond the other predictors. In our experiment, while the overall difference in R^2 between the two conditions may seem relatively small (2.8% for raw RTs, 2.6% for zRTs), this difference is larger than that of any of the single predictor variables when that variable is added to the regression equation (performed on overall RTs) in a step after all of

the other variables have been assessed. More specifically, word frequency accounted for 2.0% of the change in R^2 , and all of the other predictors accounted for less than that.

In sum, the experiment reported in this article indicates that participants reading aloud words in megastudies utilize a response criterion based on the perceived difficulty of the words in a list. When the words follow a standard random-presentation schedule, this criterion functions to restrict the range of RTs and reduces the strengths of the relationships between item-level variables, such as frequency and AoA, and RT. Alternatively, blocking items together on the basis of average RT increases the range of RTs and strengthens the power of these relationships. In addition, using blocked presentation strengthens the predictive relationship between the CDP++ model RTs and human RTs. More research will need to be done to establish the optimal blocking technique.

References

- 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, 283–316. doi:10.1037/0096-3445.133.2.283
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39, 445–459. doi:10.3758/BF03193014
- Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition* (Models and methods, orthography and phonology, Vol. 1, pp. 90–115). Hove: Psychology Press.
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 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: Erlbaum.
- Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, 36, 384–387. doi:10.3758/BF03195585
- Cortese, M. J., Hacker, S., Schock, J., & Santo, J. B. (2015). Is reading aloud performance in megastudies systematically influenced by the list context? *Quarterly Journal of Experimental Psychology*, 68, 1711–1722. doi:10.1080/17470218.2014.974624
- Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods*, 40, 791–794. doi:10.3758/BRM.40.3.791
- Cortese, M. J., McCarty, D. P., & Schock, J. (2015). A mega recognition memory study of 2,897 disyllabic words. *Quarterly Journal of Experimental Psychology*, 68, 1489–1501. doi:10.1080/17470218.2014.945096
- Cortese, M. J., & Schock, J. (2013). Imageability and age of acquisition effects in disyllabic word recognition. *Quarterly Journal of Experimental Psychology*, 66, 946–972. doi:10.1080/17470218.2012.722660

- 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
- Cutler, A. (1981). Making up materials is a confounded nuisance: Or will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10, 65–70.
- Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology*, 68, 1469–1488. doi:10.1080/17470218.2014.984730
- Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42, 488–496. doi:10.3758/BRM.42.2.488
- Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E., Tse, C., & Buchanon, E. (2013). The semantic priming project. *Behavior Research Methods*, 45, 1099–1114. doi:10.3758/s13428-012-0304-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: Erlbaum.
- Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies. A lexical decision study on 14000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Language Sciences*, 1, 174. doi:10.3389/fpsyg.2010.00174
- Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, 44, 287–304. doi:10.3758/s13428-011-0118-4
- 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, 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, 273–315. doi:10.1037/0033-295X.114.2.273
- Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large scale modelling of reading aloud with the connectionist dual process (CDP++) model. *Cognitive Psychology*, 61, 106–151. doi:10.1016/j.cogpsych.2010.04.001
- Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance*, 55, 307–314. doi:10.1037/0096-1523.28.2.307
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2000). *HLM5: Hierarchical linear and nonlinear modeling [Computer program]*. Chicago: Scientific Software.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh: Psychology Software Tools.
- Sze, W. P., Rickard Liow, S. J., & Yap, M. J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2, 500 Chinese characters. *Behavior Research Methods*, 46, 263–273. doi:10.3758/s13428-013-0355-9
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15, 971–979. doi:10.3758/PBR.15.5.971
- Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster: Touchstone Applied Science.