Journal of Experimental Psychology:
Learning, Memory, and Cognition
2013, Vol. 39, No. 4, 1037–1053

0278-7393/13/$12.00 DOI: 10.1037/a0031829

The Unexplained Nature of Reading

James S. Adelman and Suzanne J. Marquis
University of Warwick

Maura G. Sabatos-DeVito
University of North Carolina at Chapel Hill

Zachary Estes
Bocconi University

The effects of properties of words on their reading aloud response times (RTs) are 1 major source of evidence about the reading process. The precision with which such RTs could potentially be predicted by word properties is critical to evaluate our understanding of reading but is often underestimated due to contamination from individual differences. We estimated this precision without such contamination individually for 4 people who each read 2,820 words 50 times each. These estimates were compared to the precision achieved by a 31-variable regression model that outperforms current cognitive models on variance-explained criteria. Most (around 2/3) of the meaningful (non-first-phoneme, non-noise) word-level variance remained unexplained by this model. Considerable empirical and theoretical-computational effort has been expended on this area of psychology, but the high level of systematic variance remaining unexplained suggests doubts regarding contemporary accounts of the details of the mechanisms of reading at the level of the word. Future assessment of models can take advantage of the availability of our precise participant-level database.

Keywords: reading, word naming, regression models, megastudies, individual differences

Supplemental materials: http://dx.doi.org/10.1037/a0031829.supp

Reading is a core skill in modern everyday life. For this reason, it is among the oldest and most developed topics of study in experimental psychology. Reading aloud is one task that is commonly used to test theoretical ideas about how people read. Such ideas are turned into computational models of reading aloud that predict response times (RTs) on a word-by-word basis, such as the dual-route cascaded model (DRC; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) and connectionist dual-process plus model (CDP+; Perry, Ziegler, & Zorzi, 2007). Such models are designed to explain the effects of properties of words—such as how long and how common they are—on reading. Indeed, to legitimately claim to understand and explain the processes of reading, we require not only effects but explicit theoretical models of those effects.

One way to assess these models—and hence the corresponding explanations—is to compare these models' predictions with large data-sets containing observed RTs for many words (known as *mega-studies*; e.g., Balota & Spieler, 1998; Balota et al., 2007; Seidenberg & Waters, 1989; Spieler & Balota, 1997; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). The natural way to examine such correspondence is to correlate the observed and predicted RTs across words, and the correspondence is usually summarized with the R^2 statistic (over words).[1]

However, a perfect correlation is not to be expected of any model, because there is some variation between one occasion on which a person reads a word and a different occasion on which they read the same word, typically treated as experimental *noise* by averaging over these occasions. When a model does not make different predictions about these different occasions, it will inevitably fall short of perfect prediction, even if it correctly explains the differences associated with properties of the words. Therefore, to interpret such correlations, we need to know how well a model could possibly predict the response time for any given word. That is, how large a discrepancy between observed and predicted mean item RTs may be written off as due to noise?

Various techniques have been used to produce such estimates of the noise contribution to the mean RT for each word. These techniques rely on examining the variability between the different occasions on which a word has been read, but these occasions involve different participants, not the same participant. So, the estimate of what cannot be explained includes both variability that is due to different occasions and variability that has to do with individual differences. Moreover, some of these individual differences are systematically linked to the properties of words.

This article was published Online First April 8, 2013.

James S. Adelman and Suzanne J. Marquis, Department of Psychology, University of Warwick, Conventry, England; Maura G. Sabatos-DeVito, Department of Psychology, University of North Carolina at Chapel Hill; Zachary Estes, Department of Marketing, Bocconi University, Milan, Italy.

This work was funded by Economic and Social Research Council Grant RES-062-23-0545. We thank Gordon Brown, Anna Cunningham, Christoph Ungemach, and Derrick Watson for helpful comments.

Correspondence concerning this article should be addressed to James S. Adelman, Department of Psychology, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom. E-mail: J.S.Adelman@warwick.ac.uk

[1] Such correlations simultaneously assess essentially all contrasts within the study, including those on which current models differ and those on which they do not.

For instance, if the RT to BREAD is 400 ms with one participant and 600 ms with another participant, the 200-ms discrepancy could be due to any combination of (a) noise,[2] (b) an individual difference in average RT, or (c) an individual difference in the effect of a word-level variable (e.g., length). Analysis techniques that treat individual differences as noise will necessarily overestimate the amount of noise contributing to the mean RT for each word. This overestimation of noise results in an underestimation of the variability that a model should explain, leading to an overestimation of the success of models.

If 10% of the variance in the data were due to noise, and 10% were due to individual differences, then an analysis that treats individual differences as noise would only set a target of 80% variance explained for a model, when, in fact, 90% could in fact be explained. We aim to estimate the percentage of variance that is due to effects of the properties of words (i.e., not noise), even when these effects differ between participants. By doing so, the present research establishes a more stringent and accurate target for the evaluation of our understanding of reading (as instantiated by quantitative models).

Of course, if the available data do not have the same person reading each word repeatedly, there is limited information available to distinguish the sources of variability that are individual differences from those that are noise. By treating individual differences as noise, a lenient criterion is constructed, but this is both more desirable than a too-strict criterion that rejects a correct model and useful if it rejects incorrect models despite its leniency. A model that correctly explains the various effects should have numerical parameters that can be altered to accommodate individual differences in these effects.

Previous approaches to the calculation of R^2 criteria have approached the difficulty of the conflation of individual differences and noise in different ways. All of these involve simplifying assumptions that are summarized in Table 1 and explained below. Our approach does not rely on any of these simplifying assumptions.

One approach (e.g., Seidenberg & Plaut, 1998; Sibley, Kello, & Seidenberg, 2009) has been to use correlations between participants or groups of participants to estimate the non-noise variance. This has the advantage that it does not assume that all relevant effects are known and entered into the analysis. However, such analyses are based on normal distributional assumptions, which are well-known to be incorrect for response times. Moreover, this approach treats individual differences that can be characterized with a difference in intercept and overall slope as non-noise, but other individual differences—such as one participant being very length-sensitive while another being very frequency-sensitive—as noise. Effectively, this approach makes a *general speed* assumption that the only differences between participants are in intercept and slope. Thus, participant–item interactions can only be due to these differences in general speed. In addition, this method uses a data set to stand in for the correct model in estimating how well a correct model will correlate with data. However, the difference between the correct model and data that stands in for it is that the data will contain noise. In effect, the incorrect assumption is made that model predictions contain noise.

More recently, a second version of this approach has been developed by Rey, Courrieu, Schmidt-Weigand, and Jacobs (2009), which uses a *Monte Carlo* method: It simulates several new subgroups of participants by randomly selecting participants from the actual data, from which correlation estimates for the full group can be calculated. By using the Monte Carlo method, normal distributional assumptions are avoided. Even more recently, an adjustment has been made to this method to account for the absence of noise in model predictions (Courrieu, Brand-d'Abrescia, Peereman, Spieler, & Rey, 2011). These are the only changes in assumptions that these developments of the method make.

Another common approach has been to define a set of effects for which models are expected to account and use as the target the R^2 of a regression model with those predictors. For instance, Spieler and Balota (1997) found that a three-factor regression model (log. frequency, orthographic *N,* length) outperformed contemporary computational models (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989) on this criterion. Also, Perry, et al. (2007) compared their model favorably with this criterion. However, a stricter criterion could be set with more predictors, were they known (cf. Adelman & Brown, 2008a; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), but it is likely some predictors will be unknown. Moreover, just because the criterion is met, that does not mean that the model is successful in explaining the effects of interest; it may be capturing other variance. To address this problem, Besner (1999) and Adelman and Brown (2008a) suggested comparing the predicted and observed sizes of the effects, but the method is still limited by the list of known effects. The method operates on item means and so refers only to the average of participants. Consequently, no particular structure of individual differences is assumed—and in particular, not a structure based on intercept and slope alone—except that the participant–item interactions must relate to the known effects. Moreover, it also uses the normal distributional assumptions, which are incorrect for response times. This method does, however, have the advantage that the correlation is calculated on the basis that predictions from lexical variables do not contain noise, producing a correct comparison with computational models.

To avoid the assumption that all relevant variables are known and to explicitly include individual differences, one might wish to construct a linear mixed-effects regression (lmer; cf. Baayen, Davidson, & Bates, 2008) to include adjustments (random intercepts) for words and participants. The per-word adjustments allow for arbitrary unknown effects. To account for individual differences not involving the intercept—such as one participant being particularly sensitive to length—participant slope adjustments (random slopes) of the lexical variables could be included. While these allow for participants differing in more than intercept and a single slope, the adjustments (which are a form of participant–item interaction) are limited to known slopes (i.e., those involving the known effects). The estimate of noise from this approach would be calculated on the basis that predictions from lexical variables do not contain noise, so would be useful for assessing computational models. Typically, such models use the normal distributional assumptions.

[2] Or practice, or priming, or other sequential effects that are treated as noise in a typical experiment comparing some words with some other words. These contributions are discussed in detail in the Discussion.

Table 1
Statistical Assumptions That Would Ideally Be Avoided in the Estimation of R^2 Targets for Cognitive Models and Techniques of Estimation

Model/technique	Assumption to be avoided				
	All effects known	Participants differ in only intercept and slope	Participant-item interactions only due to known slopes	Cognitive model predictions are simulated with noise	Noise normally distributed with equal variance
Correlate with other data set (e.g., Seidenberg & Plaut, 1998; Sibley et al., 2009)	✓	x	—	x	x
Squared intraclass correlation by Monte Carlo (Rey et al., 2009)	✓	x	—	x	✓
Intraclass correlation by Monte Carlo (Courrieu et al., 2011)	✓	x	—	✓	✓
Comparison with regression (e.g., Adelman & Brown, 2008a; Spieler & Balota, 1997)	x	✓	x	✓	x
Participant-effect model (by either lmer or ANOVA)	✓	✓	x	✓	x
Present approach	✓	✓	✓	✓	✓

Note. ✓ = avoided; x = not avoided; ANOVA = analysis of variance. Participant-effect model describes estimates based on models with a (random) factor effect of each of participant and item, plus (fixed) word-property effects and their (random) interaction with participants. A dash indicates that this assumption is not meaningful when participants differ in only intercept and slope.

The analogous analysis of variance model has participants and words as factors. The lexical variable effects would be constructed as contrasts within the main effect of word, and these could interact with participants. The same problematic assumptions apply. In particular, the per-participant item effects must be limited to the known effects. This is because in data sets where there is just one observation per participant per item, including arbitrary item effects on a per participant basis (i.e., the complete Participant × Word interaction) would exhaust all the variance in the data set, incorrectly leaving none as noise. If we correctly wish to allow in our statistical model that individual differences may occur in the unknown effects, we must obtain multiple observations of each participant–word combination. In such a data set, the participant–word interaction is not confounded with the trial-to-trial noise.

Additionally, estimating the average over individuals causes problems when testing models because although the magnitude of particular effects may vary between participants, using average data will never require models to account for the range of effect magnitudes shown, only the average magnitude. Moreover, averaging over participants may be misleading as to the underlying functional form (shape) of some effects (e.g.,W. K. Estes, 1956; Heathcote, Brown, & Mewhort, 2002). Therefore, to the extent that models are distinguished by their predictions of functional form of effects, such as the frequency effect (e.g., Adelman & Brown, 2008b), average data are insufficient. On both counts, if possible, it is therefore preferable to analyze participants separately. However, no existing database has sufficient data to support such individual analysis of participants.

With sufficient data from each participant, with replications of each participant–word combination, we can avoid the aforementioned problematic statistical assumptions, and analyze each participant individually. This allows us to construct a statistical model that can accommodate effects that are not known in advance. Moreover, by analyzing them separately, participants may differ arbitrarily, not only in intercept and slope. It is particularly important that we can examine the effect of word and its interaction with participant without restriction to the known effects. By using a Monte Carlo technique appropriately, we can also avoid assuming model predictions have noise and allow that RTs are not normally distributed.

Thus, the present work seeks to address the aforementioned fundamental problems in assessing models (see Table 1) by reporting a mega-study on four individuals who read 2,820 words, 50 times each. By way of comparison, Spieler and Balota's (1997) 30 participants read each of these words once; and Elexicon's naming data (Balota et al., 2007) contain 25 observations for each of 40,481 words, spread over 444 participants (who read 2,530 or 2,531 words each). Notably, neither of these older data sets contains the participant–word repetitions that are necessary to accurately test models of reading aloud. The present study is also the first mega-study to have hand-coding of response times, which is more accurate than the typical voice-key coding (at least, when voice keys are being assessed, hand coding is used as the benchmark for correctness, e.g., Rastle & Davis, 2002). On both the grounds of the precision of individual measurement and the accuracy of coding, these data can be seen as key for the future assessment of cognitive models of visual word recognition. The full database is available in the online supplemental materials.

We use these data to (a) demonstrate the differences between individuals, (b) estimate the proportion of variance in individuals' item mean RTs that is in principle explicable (not noise), and (c) attempt to explain this variance with existing and new factors. We emphasize the variance in item mean RTs (rather than the total variance in individual trials), because this accords with other uses

of mega-study data, and because item-property effects have been the preeminent source of information about the processes of visual word recognition.

Method

Participants

Participants were recruited by an e-mail advertisement to the staff, faculty and postgraduates of the Department of Psychology at the University of Warwick.

D was a 27- to 28-year-old British postdoctoral researcher in psychology at the time of the study. He is right-handed (scoring +20 on the Waterloo handedness questionnaire) and had approximately normal vision (0.04 logMAR in the left eye, 0.14 in the right). He scored 23/80 (corrected for guessing; 1.7 *SD*s above mean in original sample) on the United Kingdom author recognition test (Masterson & Hayes, 2007), and his verbal IQ, according to the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999), was 127 (superior).

A was a 31- to 32-year-old American medical student at the time of the study. He is right-handed (+28) and had vision corrected to normal (0.04, −0.04) by contact lenses or glasses. He scored 13/50 (corrected; 0.6 *SD*s above mean in original sample) on the original U.S. author recognition test (Stanovich & West, 1989), and his WASI verbal IQ was measured as 108 (average).

M was a 24- to 25-year-old British research assistant and doctoral student in psychology at the time of the study. He is right-handed (+32) and had vision corrected to approximately normal (0.02, 0.12) by glasses. He scored 4/80 (corrected; 0.7 *SD*s below mean) on the United Kingdom author recognition test, and his WASI verbal IQ was measured as 128 (superior).

U was a 53-year-old British personal assistant employed in the psychology department at the time of the study. She is right-handed (+22) and had vision corrected to approximately normal (0.06, 0.14) by glasses. She scored 10/80 (corrected; 0.1 *SD*s above mean) on the United Kingdom author recognition test, and her WASI verbal IQ was measured as 121 (superior).

Each received £1000 (ca. U.S. $1,700) for participation.

Apparatus

The experiment was controlled by custom software on computer. Stimuli were presented on a Sony CPD-G200 17-in. (43.18-cm) display at 1024 × 768 pixels. A Plantronics Audio 370 gaming headset with microphone was attached to an Ensoniq 5880 AudioPCI sound card for recording responses.

Design

For each participant, 50 lists were created that each contained a new random ordering of the 2,820 monosyllabic words from Spieler and Balota's (1997) study. Each list was split into two halves for presentation on consecutive sessions of word naming to obtain response times.

Procedure

At the beginning of each of the 100 1-hr sessions, participants tested the audio equipment by reading aloud the word "testing" into the microphone, which was recorded and then played back over the headphones. Once they were confident the equipment was operating correctly, they pressed the space bar on the computer keyboard to proceed. The first trial window began after a delay of a second.

Each trial was allocated a window of 2,300 ms. The onset of the stimulus was jittered to be 100–300 ms (uniformly distributed) after the nominal beginning of the trial. The word was then presented in black-on-white 24-point lowercase Courier font in the center of the screen for 1,500 ms, and the participant's vocal response recorded for this period. No feedback was given. The interval from stimulus offset to next stimulus onset therefore varied between 600 and 1,000 ms.

Participants were permitted breaks on demand. Pressing the space bar instigated a pause in the experiment at the end of the trial window indicated by the display of a blank black screen; the next trial window began 1 s after the next press of the space bar. If a long delay was needed, participants could press the "Q" key to exit the experiment; if so, the remainder of the session began with the testing procedure when they reentered the experiment. Each session was completed on the same day it was started,[3] and participants completed at most one session per day.

Data Coding

Response times and errors were coded by visual and auditory inspection of the waveform, with the assistance of a version of the open-source Audacity software package modified for the purpose by the first author. With the exception of the first session, software provided an estimated response time based on a two-stage voice key applied to a filtered version of the sound wave, which was corrected if necessary by the human coder. The third author coded the data from the first 20 sessions for each participant; the second author coded the remainder.

Given the use of two data coders, the second author additionally coded the first session to evaluate inter-rater reliability. RTs for valid correct trials provided by the two raters correlated $r = .991$ overall, and within each participant, $r = .976$, .995, .990, and .993 for D, A, M, and U, respectively. Given that D's sound waves were subjectively not more difficult to code than the other participants', we attribute the slightly weaker correlation to practice, as his first session was the first coded by both coders.

Results

Of the 564,000 trials, 25,110 (4.45%) were either incorrect or unusable. These are broken down by participant and type in Table 2. Ignoring trials removed for reasons other than error (e.g., stutters, equipment failures), A exhibited a greater proportion of errors than the other participants; this was also true of other language tasks not reported here. As a proportion of the errors made, D showed a relatively high rate of visual errors; this is difficult to interpret: In a tachistoscopic word identification task, D's performance was similar to U's (although D is much more practiced at the task). Summary statistics for the mean correct valid RTs for each participant are given in Table 2.

[3] M did not comply with this instruction on two sessions, once due to equipment failure.

Table 2
Discarded Trials by Participant and Type and Summary Statistics for Response Times

	Type of response					Item mean response times (ms)			
		Incorrect, valid							
Participant	Correct, valid	Phonological	Visual	Other	Invalid	*M*	*SD*	Min.	Max.
D	137,931 (97.82%)	597 (0.42%)	1,281 (0.91%)	93 (0.07%)	1,098 (0.78%)	498.7	35.8	428.0	872.6
A	135,050 (95.78%)	1,719 (1.22%)	1,925 (1.37%)	291 (0.21%)	2,015 (1.43%)	598.8	49.6	486.3	922.5
M	128,104 (90.85%)	442 (0.31%)	570 (0.40%)	113 (0.08%)	11,770 (8.35%)	681.8	46.1	526.4	1,155.0
U	137,805 (97.73%)	323 (0.23%)	386 (0.27%)	72 (0.05%)	2,414 (1.71%)	473.2	37.1	387.3	701.1

Note. Phonological = errors whose most plausible interpretation is that an orthographic segment has been read in a way that would be valid in another word (e.g., a regularization); Visual = errors whose most plausible interpretation is that letters were misperceived; Other = errors that could roughly equally plausibly be interpreted as either visual or phonological, were most plausibly interpreted as mixed (e.g., visual then phonological), or had no clear interpretation; Min. = minimum; Max. = maximum. Invalid trials include equipment failures (a sound card fault affected M), stutters (common for U), absent (or very late) responses (common for M), and nonstandard realization of the first phoneme (exchanges among /t/, /d/ and /θ/ were common for A, due to his accent). Percentages are row-wise.

The General Speed Assumption Is Wrong

The present study was motivated by our supposition that individual differences between participants would invalidate approaches that treat participants as replications of one another, in the sense that a single common underlying difficulty factor is responsible for the differences between items in RTs for all participants. While there are clear overall differences in participants' RTs, such approaches can adapt to differences in general speed. We therefore first compared the participants using a simple regression model to examine individual differences in effect magnitude, with some of the most important variables in word naming, as follows:

First phoneme. First phoneme was dummy-coded with 38 levels from the CELEX (Baayen, Piepenbrock, & Gulikers, 1995) transcription.

Exception costs. An exception cost is defined as the effect of having a pronunciation that breaks spelling-sound rules (being an exception word not a regular word); for present purposes, spelling-sound rules were taken from the DRC (Coltheart et al., 2001), and an exception was identified if these rules produced a pronunciation that differed from that in the model's vocabulary. Exception costs were calculated separately for each of the first three letter positions (the earliest irregular position, if more than one), because costs are greater if the exception comes about early in the word (the position of irregularity effect, e.g., Rastle & Coltheart, 1999). For instance, PINT is exceptional in second position (I) where the regular pronunciation would be as in HINT.

Orthographic and phonographic neighborhood sizes. Orthographic neighbors are words formed by replacing a single letter. Phonological neighbors are words formed by replacing a single phoneme. Phonographic neighbors are words that are simultaneously orthographic and phonological neighbors (Adelman & Brown, 2007; Peereman & Content, 1997). For instance, STOKE and SPOKE are phonographic neighbors, but STOKE and STORE are only orthographic neighbors without being phonological neighbors (and hence are not phonographic neighbors). Neighborhood size (*N*) is the count of number of neighbors of a specific type. Phonographic *N* is reported, and orthographic *N* is reported as the number of orthographic neighbors that are not phonographic neighbors, so that each neighbor contributes to only one coefficient.

Frequency. The natural logarithm of the written frequency from CELEX (plus one, as usual, to avoid taking the logarithm of zero, e.g., Balota et al., 2004) was also used.

Separate slopes were fitted for exception and regular words to code the Frequency × Regularity interaction (e.g., Seidenberg, Waters, Barnes, & Tanenhaus, 1984).

Orthographic length. Number of letters was entered with separate slopes for exception and regular words, coding the interaction observed in Spieler and Balota's (1997) data by Adelman (2005).

The regression coefficients (except for first phoneme) are illustrated in Figure 1, and interaction tests for the difference between these data and Spieler and Balota's (1997), and for the difference among our four participants are given in Table 3. Our participants showed significantly stronger frequency effects, weaker length effects and stronger position 2 exception effects compared to Spieler and Balota's, as well as showing inhibitory neighborhood effects instead of facilitatory ones.

Turning to look at individual participants, D and U show similar patterns to one another and the SB97 data, except for neighborhood effects. In general, A shows stronger effects than the other participants. While M's exception costs are similar to the other participants', he does not show inhibition from length, and his frequency effects are weaker.

Neighborhood effects are typically facilitatory in naming (Adelman & Brown, 2007; Andrews, 1989, 1992; Balota et al., 2004), so it is surprising that they are absent and inhibitory in these data. This may result from either the conditions of the experiment (e.g., the trials are not self-paced, or the long-term nature of the experiment) or properties of the participants (i.e., individual differences).[4] Otherwise, the results appear typical.

Concentrating on this set of well-known effects, however, biases us toward finding the similarities between individuals: These effects are well-known because they were demonstrated in group studies of word naming. Such studies are not usually published if the by-subjects statistic (t_1 or F_1) is not significant, which is more likely if individuals differ in the size of the effect, because these statistics weigh the average size of the effect against such vari-

[4] There is reason to believe that more experienced readers are less sensitive to neighborhood effects (Sears, Siakaluk, Chow, & Buchanan, 2008).

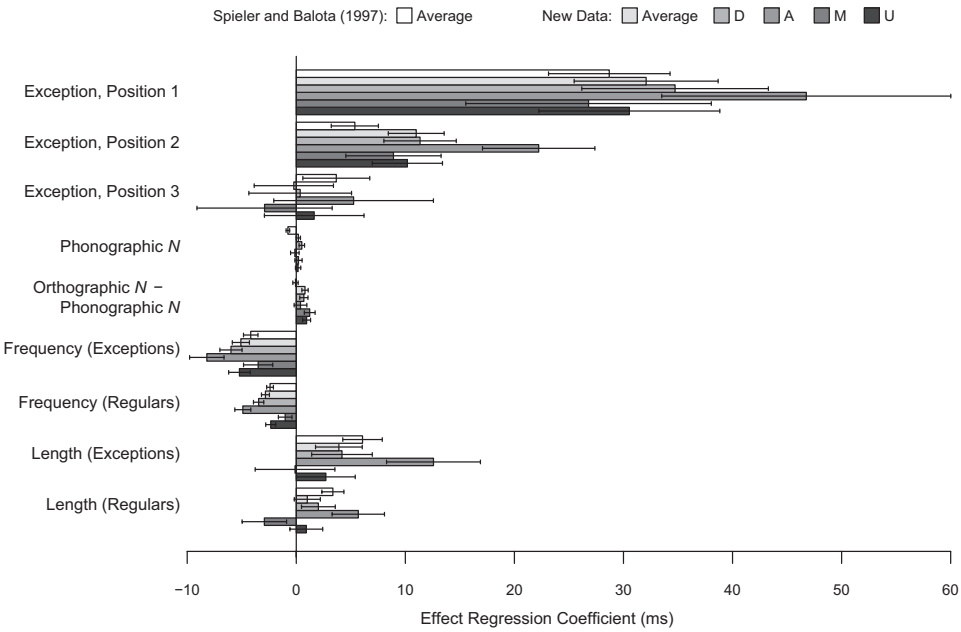

Figure 1. Effect magnitudes (ms) for some key effects compared across the four participants and Spieler and Balota's (1997) data. Error bars represent 95% central confidence intervals for the regression coefficients.

ability. A more direct way to examine the similarity of participants is the correlation between their item mean response times. The raw correlations are presented in the top half of Table 4, alongside some previous mega-studies.

First, we note that two of our four participants show negative correlations with SB97 and the other previous mega-studies, despite the strong qualitative similarities in most of the standard effects given in Figure 1. Examination of the same calculations using the residuals after first phoneme given in the bottom half of Table 4—as this main effect is not usually considered to be caused by the reading process of interest—reveals that the first phoneme is the source of the serious inconsistency. This may reflect the different measurement method for response times (hand coding vs. voice key) used in the two studies, or it may reflect differences in the articulatory influences on response times between our participants and the average participant. Of our participants, A is clearly the most similar to the average undergraduate from the past studies.

Even so, the correlations between participants were moderate, but far from perfect. Of course, the imperfect correlations that are exhibited consist of both the actual underlying differences between participants, and the variation in each participant's response to each word. The structure of the present data set, with 50 replications of each word for each participant, allows us to calculate the split-half (Sessions 1, 2, 5, 6, . . . vs. Sessions 3, 4, 7, 8, . . .) reliability and thereby estimate the reliability of each person's data with the Spearman-Brown correction. These reliability estimates were D: .869, A: .823, M: .681, U: .861 for the raw values and D: .623, A: .816, M: .245, U: .492 for the residuals after first phoneme; overall, these were generally good.

Given these estimates of reliability and the observed correlations, we can estimate the underlying correlation (adjusting for the

Table 3
Comparison of Data Sets on Participants on Major Word Naming Effects Using the Interaction Between Data Set/participant and the Regression Slope for Each Effect

	SB97 vs. new data		D vs. A vs. M vs. U	
Effect	$F(1, 2664)$	p	$F(3, 7990)$	p
Exception, Position 1	1.565	.211	3.557	**.014**
Exception, Position 2	6.245	**.012**	5.269	**.001**
Exception, Position 3	0.154	.695	2.550	.054
Phonographic *N*	75.690	**<.0001**	3.412	**.017**
Orthographic *N* − Phonographic *N*	28.164	**<.0001**	3.344	**.018**
Frequency (Exceptions)	5.406	**.020**	19.476	**<.0001**
Frequency (Regulars)	3.912	**.048**	38.778	**<.0001**
Length (Exceptions)	7.236	**.007**	16.648	**<.0001**
Length (Regulars)	11.669	**.001**	21.329	**<.0001**

Note. Boldface type indicates significant at the .05 level.

Table 4
Correlations Between Item Mean RTs of Each Participant and Previous Mega-Studies and Partial Correlations After First Phoneme Between Item Mean RTs of Each Participant and Previous Mega-Studies

Variable	D	A	M	U	SB97	EL	EL LDT	BL LDT
				Raw correlations				
D	—							
A	0.556	—						
M	0.433	0.432	—					
U	0.531	0.596	0.569	—				
SB97	0.122	0.162	−0.144	−0.009	—			
EL	0.032	0.132	−0.198	−0.001	0.562	—		
EL LDT	0.188	0.245	0.056	0.135	0.302	0.429	—	
BL LDT	0.220	0.263	0.058	0.153	0.281	0.394	0.677	—
				Correlations calculated using the residuals after first phoneme				
D	—							
A	0.348	—						
M	0.234	0.222	—					
U	0.392	0.388	0.268	—				
SB97	0.265	0.353	0.122	0.286	—			
EL	0.337	0.419	0.130	0.323	0.506	—		
EL LDT	0.315	0.318	0.077	0.226	0.363	0.494	—	
BL LDT	0.332	0.338	0.075	0.236	0.364	0.466	0.677	—

Note. RT = response time; SB97 = Spieler and Balota (1997) item means; EL = Elexicon Project (Balota et al., 2007); BL = British Lexicon Project; LDT = task is lexical decision, not naming.

fact that observed correlations correlate both the signal and the noise—attenuation—due to imperfect reliability) using the standard method from classical test theory. For the raw values, these were D versus A: .657, D versus M: .563, D versus U: .614, A versus M: .577, A versus U: .708, M versus U: .743. For the residuals after first phoneme, these were D versus A: .488, D versus M: .599, D versus U: .708, A versus M: .797, A versus U: .612, M versus U: .772. While these values are large, they are far from 1, suggesting that substantial individual differences exist.

We further tested the correlations in the upper part of Table 4 with the reliabilities above using Kristof's (1973) method[5] to see whether participant differences could be accounted for by noise plus a linear relationship (i.e., the intercept and a single overall slope). As might be expected from the low disattenuated correlations, such a relationship was heavily violated for every pair of participants (all $t > 14$, $p < .0001$). When the same procedure was applied to the correlations and reliabilities calculated from the residuals after first phoneme in the lower part of Table 4, four of the six pairs showed a strong violation (those showing the lower disattenuated correlations; $t > 3.5$, $p < .001$), and the other two did not ($t < 0$; only positive values give evidence to reject a null hypothesis.). That is, participant differences could not be accounted for by noise plus a linear relationship, as the models used in many analyses assume.

Overall, these analyses show that there are stable individual differences between participants that go beyond general speed in terms of a different average or a simple multiplier (which might correspond to a different cycles-to-milliseconds conversion in a simulated model). While it would be possible to perform analyses on these data that do not assume a general speed relationship between participants, and thus calculate better targets to apply to the average data, this would still underestimate the strength of constraint that these data can place on models. Given that the item average data from each of the individual participants are reasonably reliable, we could in principle ask how well a model fares in explaining the systematic properties of the item data of each participant. This would require that a model is adjusted to the particular data set in some way to account for the maximum possible variance, as is true of any attempt to rule out models using R^2 targets (Adelman & Brown, 2008a). The comparison of course also requires that we have an appropriate estimate of how much of the variance in the item means is systematically due to item properties (as averaging reduces but does not eliminate noise and other irrelevant influences).

How Much Variance Would a Correct Model Explain?

We therefore now turn to our main question: If a model gave the correct mean RT for each word for each participant, how much of the observed variance *in item mean RT* would that model explain (i.e., how high would its R^2 be) for each participant? This will not be 100% because there is trial-to-trial noise in the data. Put otherwise, the question is how much misprediction by a model should we tolerate as due to noise? The answer to this question is critical for evaluating models of reading.

To avoid the assumptions of equality of variance and normal distributions for each participant–word combination, we used Monte Carlo simulations where each simulated RT is drawn from the observed RTs (adjusted for the main effect of session) for that participant–word combination (replicated 1,001 times). Such a Monte Carlo bootstrap procedure assumes that the observed distribution is a good approximation to the true distribution. We then examined how well a sample of the size of our data set would be captured by a perfect model that predicted the true underlying participant–word mean RTs. This was done in terms of the error

[5] Alternative forms were constructed using split halves from odd and even double-sessions (i.e., Sessions 1, 2, 5, 6, etc. vs. Sessions 3, 4, 7, 8).

sum-of-squares, because its distribution (unlike the correlation coefficient's or R^2's) does not depend on the quality of our estimate of the true values of the underlying mean RTs (which property is known as being a pivotal statistic). For ease of comparison, we then used the (fixed) total sum-of-squares estimate to compute R^2 distributions for each participant giving the expected R^2 of a correct model.[6] Table 5 presents these target values (in the second row) and presents them alongside a baseline measure (in the first row) of the variance explained by the first phoneme (as previously described). While first phoneme is often the most important predictor, it is usually only treated as a control variable (rather than a variable of interest) because it is presumed to reflect articulatory-mechanical (and possibly voice-key-acoustic) speech factors that are not related to the reading process of interest. We consider only the explicable variance after the main effect of first phoneme to unambiguously be of interest for the computational modeling of item-level variance in word recognition, so this amount-above-baseline measure is also presented in the table.

How Does Our Present Knowledge Fare?

The data that we collected produce highly reliable item means, and therefore it should be possible to explain these using the properties of the words. As the targets in the second row of Table 5 indicate, up to ca. 90% (depending on the participant) in each participant's data would be explained by a model correctly generating the mean for each word for each participant, because we have averaged over the noise (and sequential effects) at the trial-to-trial level, although much of this is due to first phoneme. We ask here whether the factors that we already know affect visual word recognition account for this variance.

The early criticisms of computational models' abilities to account for item-level variance from Spieler and Balota (1997) made use of a three-variable regression (log. frequency, length, orthographic *N*) that has been used as a standard target by some modelers. For instance, Perry et al. (2007) give their CDP+ model a check mark as successful on an item-variance criterion because it performs similarly in R^2 to such a regression. Table 5 includes the R^2 values for such a regression (after first phoneme). The results show that these three variables account for rather little of the variance in the data, even when the proportion is rescaled to exclude first phoneme and noise variance: Less than one sixth of the variance we would hope to be explained is due to these three variables. As such, these data show this to be a very weak test of a model's ability to account for item-level variance.

The weakness of this criterion can be seen as not particularly surprising because a wider range of variables is known to affect word recognition, and we have previously used such a wider range for assessing models. In particular, Adelman and Brown (2008a) used a regression model for Spieler and Balota's (1997) mean item response times, which included 17 orthographic and phonological variables and interactions to assess the performance of the dual-route cascaded (DRC) model (Coltheart et al., 2001), which is among the most influential models of visual word recognition. This regression model serves as a useful benchmark because it can explain 98.21% of the variance in DRC predictions; that is, essentially every variable important to the DRC is included. These variables are as follows:

First phoneme. As previously described.

Exception costs. Exception costs were used as previously defined, but separate costs were estimated for each of five positions, crossed with two types of irregularity, those affecting one-letter graphemes and those affecting two-letter graphemes (see Andrews, Woollams, & Bond, 2005). In addition, for the latter case an extra cost was estimated for the case where more than one phoneme was irregular.

Whammy costs. A *whammy* is defined as a spelling that follows spelling-sound rules in a way that changes as read left-to-right (Rastle & Coltheart, 1999). For instance, CASH has a whammy in third position because adding the H to CAS changes the pronunciation of the third phoneme (rather than adding a fourth phoneme). Costs were estimated for each of five positions in which whammies occurred.

Frequency. We used polynomial terms up to cubic in log. CELEX frequency. Separate coefficients were again used for exception and regular words.

Orthographic and phonemic length. Number of letters and number of phonemes were both included with separate slopes for exception and regular words.

Orthographic, phonological and phonographic neighborhood sizes. All three neighborhood sizes were included.

Feedforward rime consistency ratio. The feedforward rime consistency ratio is defined as the ratio of number of friends to (number of friends plus number of enemies), where friends are words that are spelled as though they rhyme and do, and enemies are words that are spelled as though they rhyme but do not. For instance, FORK is a friend of PORK, but WORK is an enemy to both. We use both the type ratio, where each word counts once regardless of its frequency, and the log. tokens ratio where each word is weighted by the log. of its frequency.

Interactions with frequency. Multiplicative interactions of the frequency terms with phonographic *N,* and with type feedforward rime consistency ratio were included.

We used this regression for the present data. The results for the additional effects in this regression (i.e., those not illustrated in Figure 1 and discussed above), were as follows: The cost of a second irregularity was significant for D, A and U, but not present for M. A whammy in position one was inhibitory for all participants; a whammy in position two was inhibitory for A; in contrast, later whammies tended to be facilitatory, significantly so for positions three, four and five for D, four for M, and three for U. Phonological length was facilitatory; this was significant only for exceptions for D, A and U, but for both regulars and exceptions for M. In addition, the phonological length effect removed the surprising facilitation from orthographic length for M (making it null).

D showed no consistency effects. A showed a facilitatory consistency effect by the types measure (only if it was entered alone, but the log-tokens measure was not significant alone) and no interaction with frequency. M showed a facilitatory consistency effect by either measure (but neither explained unique variance) and no interaction with frequency. U showed a facilitatory consis-

[6] We also calculated how low an R^2 would suggest rejection of some fixed model in a null-hypothesis significance test at the 5% level by taking the 5th percentile of the simulated distribution; these values were D: 87.28%; A: 90.90%; M: 68.39%; U: 86.27%.

Table 5
Per-Participant R^2 Values (%) for Hypothetical Correct Model, Various Regression Equations, and Computational Models

Type of model	Total R^2				Amount above baseline				As % of target			
	D	A	M	U	D	A	M	U	D	A	M	U
Baseline: First phoneme alone	53.30	40.62	55.01	61.02	←subtracted from *baseline* (e.g., 87.96 − 53.30 = 34.66)							
Target from hypothetical correct model	87.96	91.63	69.81	87.17	34.66	51.01	14.80	26.15	← divided by *target* (e.g., 5.41/34.66 = 0.1560)			
Regression models												
log. frequency, length, *N*	58.71	46.95	56.07	63.03	5.41	6.43	1.06	2.02	15.60	12.61	7.16	7.72
Adelman & Brown's (2008a) 17 predictors	61.78	53.36	59.58	67.65	8.48	12.74	4.57	6.62	24.46	24.98	30.87	23.32
All 31 predictors	65.61	56.41	60.08	68.80	12.31	15.79	5.07	7.78	35.52	30.95	34.26	29.75
Computational models												
DRC	57.08	46.15	55.41	63.07	3.78	5.53	0.40	2.05	10.91	10.84	2.70	7.84
CDP+	57.25	47.43	55.25	62.97	3.95	6.81	0.24	1.95	11.40	13.35	1.62	7.46
PMSP Simulation 1	55.53	44.86	56.05	63.49	2.23	4.24	1.04	2.47	6.43	8.31	7.03	9.45
PMSP Simulation 4	57.31	46.08	55.43	62.94	4.01	5.46	0.42	1.92	11.57	10.70	2.84	7.34

Note. DRC = dual-route cascaded model (Coltheart et al., 2001); CDP+ = connectionist dual-process plus model (Perry, Ziegler, & Zorzi, 2007); PMSP = Plaut et al. (1996).

tency effect by either measure (but only the tokens measure explained unique variance) and an interaction with frequency, such that facilitation was greater for low frequency words.

The R^2 values for these regressions are given in Table 5. While these variables make some progress toward the target, these values are well shy of explaining all the variance that can be attributed to word properties: The overall R^2 values are 17–44% short of the target, which amounts to 69–76% of the variance due to neither first phoneme nor noise remaining unexplained.

What Have We Missed?

Given the disappointing level of variance for which the variables in our regression equation could account, we sought to identify further relevant measures.

Accessibility measures. We examined whether another accessibility measure could improve the regression model's R^2, using five additional corpora from which both word frequency (WF) and contextual diversity (CD; the number of contexts/documents in which a word appears; Adelman, Brown, & Quesada, 2006) counts were available or calculable. We replaced the polynomial (cubic) regression in log WF in Adelman and Brown's (2008a) regression equation with restricted cubic splines with four knots[7] (at {.05, .35, .65, .95} quantiles) in log. WF or log. CD (1 was added before the logarithm was taken), improving the reflection of the data particularly in the tails (which is the usual reason to prefer cubic splines). The resulting R^2 values with each corpus count for each participant are presented in Table 6. M and U had shown the weaker frequency effects, and there was no evidence here that this was due to the frequency count being suboptimal: The count originally used accounts for the most variance. By contrast, the effects for D and A may have been underestimated, as the CELEX count performs relatively poorly in explaining their RTs, with TASA CD being best for D, and SUBTL WF best for A.

Vowel length. Examination of the most under- and over-predicted RTs in U's data suggested that those that were slower than expected tended to have longer vowels (e.g., *lose*, *chief*, and *cease*), while those that were quicker than expected tended to have shorter vowels (e.g., *his*, *soot*, and *drag*). We categorized the vowel in each word as short, long or diphthongal according to CELEX. Adding this factor only modestly improved the variance accounted for by D: 0.06%, A: 0.63%, M: 0.02%, and U: 0.08% (*ns* for M). Moreover, the ordering of the types differed between participants.

Visual confusability. Examination of the words with slower than expected responses in D's data suggested many came from words that had orthographic neighbors formed by replacing an *h* with a *b* or vice versa. Given that many of D's erroneous responses were substitutions involving this letter pair, the effect appears to be due to visual confusion. To quantify this effect we used the confusion probabilities (at distance) observed by Bouma (1971) to calculate a notional probability (product of the probability of the individual confusions) of confusing the word with its nearest neighbor (the word with the highest confusion probability). Words with high probability confusion neighbors were read more slowly by D (increase in R^2 of 0.14%) and U (0.09%), but not A (0.00%) and M (0.02%). We also calculated the notional probability of correctly identifying all the letters in the word. Words whose letters were not confusable were read faster by D (0.29%) and M (0.09%), but more slowly by A (0.12%), with no effect (0.02%) on U. This is consistent with the neighborhood effects: A relies on generalization from neighbors to aid naming, while the others discriminate the word precisely, in line with the lexical quality hypothesis (e.g., Andrews, 2008).

Other neighborhood variables (OLD20 & PLD20). We also calculated Orthographic Levenshtein Distance 20-nearest (OLD20; Yarkoni, Balota, & Yap, 2008) and the analogous phonological measure (PLD20) from CELEX. These newer neighborhood measures are based on the average Levenshtein distance of the 20 nearest neighbors; compared to the traditional neighborhood size, these measures take into account more types of neighbors (as

[7] Using 7 knots increased the R^2 without modifying the qualitative patterns we report, but with the introduction of nonmonotonicities in the predictions that were highly uninterpretable.

Table 6
Per-Participant R^2 Values (%) for Regressions Using a Variety of Measures of Frequency and Contextual Diversity

Type of model	Accessibility transformation	Source	D	A	M	U
Regression models (Adelman & Brown's, 2008a, 17 predictors)	Polynomial (log.)	CELEX	61.78	53.36	59.58	**67.65**
	Cubic spline (log.)	CELEX	61.86	53.52	**59.58**	67.64
	Cubic spline (log.)	KF WF	61.17	52.77	59.33	66.77
	Cubic spline (log.)	KF CD	61.48	52.68	59.30	66.67
	Cubic spline (log.)	TASA WF	63.00	54.65	59.45	67.41
	Cubic spline (log.)	TASA CD	**63.22**	54.60	59.46	67.34
	Cubic spline (log.)	BNC WF	62.46	53.93	59.43	67.31
	Cubic spline (log.)	BNC CD	62.83	53.91	59.40	67.27
	Cubic spline (log.)	USENET WF	62.44	54.19	59.42	67.10
	Cubic spline (log.)	USENET CD	62.64	53.95	59.44	67.06
	Cubic spline (log.)	SUBTL WF	63.11	**54.80**	59.42	67.21
	Cubic spline (log.)	SUBTL CD	63.20	54.67	59.44	67.16

Note. WF = word frequency; CD = contextual diversity. Corpora used are CELEX (Baayen, Piepenbrock, & Gulikers, 1995); the Brown corpus (KF; Kuçera & Francis, 1967); the Touchstone Applied Science Associates (TASA) corpus used by Zeno, Ivens, Millard, & Duvvuri (1995); the British National Corpus (BNC; British National Corpus Consortium, 2000); a USENET corpus (Shaoul & Westbury, 2009); and a film and television subtitle corpus (SUBTL; Brysbaert & New, 2009). Among the 17-predictor regression models with the various accessibility measures from the corpora, the highest R^2 (best measure) is highlighted in bold.

they permit deletions and insertions) and more distant neighbors (not only the very nearest). OLD20 and PLD20 together accounted for an additional D: 0.47%, A: 0.09%, M: 0.03%, U: 0.16% of the variance, which was significant for D and U such that more isolated words were read more quickly. Again, this is contrary to the usual effect of neighborhoods.

Feedback and onset consistency measures. In addition to the feedforward (sound implied by spelling) rime consistency that is most heavily relied upon, we also examined feedback (spelling implied by sound) measures, and those based on the onsets of words, following Balota et al. (2004) in using the log. tokens measure for this purpose. The three additional measures further increased the R^2 by D: 0.52%, A: 0.18%, M: 0.07%, U: 0.41%. In terms of the individual coefficients, feedfoward onset consistency was facilitatory for all participants, feedback onset consistency was significant for U only (in the facilitatory direction) and feedback rime consistency was not significant for any participant.

Imageability. Semantics is an underrepresented domain both in models of visual word recognition and among variables suitable for regression analyses. For most semantic variables, relevant values are not available for the majority of words in the present study; this means they cannot sensibly be used for overall R^2 assessments for the present data. The imageability norms of Cortese and Fugett (2004) are an exception, with only 206 words missing (for which we used the mean). The imageability effect was significant only for A (0.15% of the variance).

Number of semantic features. Pexman, Lupker, and Hino (2002) found evidence that number of semantic features affected response times in naming and lexical decision such that words with more features were read faster. However, in the present data, using Harm's (2002) feature lists (derived from WordNet; Miller, 1995), D and U showed a significant effect in the opposite direction: Words with fewer features were given responses sooner (D: 0.19%, A: 0.00%, M: 0.05%, U: 0.09%). To investigate whether this discrepancy was due to the response time data differing or the feature norms differing, we conducted the analogous analysis with the SB97 response time data and these feature counts (and CELEX frequencies). These data also showed the inhibitory effect of number of features. Moreover, the zero-order correlations of RT and number of features was positive for each of our participants. Given that Pexman et al. used results of a feature listing task to count features, it seems likely that the discrepant results here reflect the salience or accessibility of features (over and above the number of features) being important in their counts and for naming RTs.

Familiarity and age of acquisition. Familiarity and age of acquisition variables have been offered as alternatives or additions to the more objective corpus-based measures of experience of language. First, we entered familiarity from Balota, Pilotti, and Cortese's (2001) norms, which accounted for some significant additional variance (D: 0.42%; A: 0.22%; M: 0.07%; U: 0.16%). Then we added age of acquisition from Cortese and Khanna's (2008) norms, which accounted for no additional variance (the only value numerically above zero being U: 0.02%). Age of acquisition did have a significant inhibitory effect for three participants if entered into the regression before familiarity (D: 0.12%; A: 0.14%; M: 0.01%; U: 0.14%).

Emotion-related variables: Arousal and valence. We collected arousal and valence ratings as described in the Appendix. More arousing words were read more quickly (D: 0.09%; A: 0.03%; M: 0.00%; U: 0.04%), this effect being significant only for D. In line with our previous findings (Z. Estes & Adelman, 2008a, 2008b), we used a binary split of valence at the midpoint of the scale. An effect such that negative words were read more quickly was significant for D (0.06%) and U (0.07%), which was an effect in the opposite direction to previous research (the same pattern was weaker with the raw valence scores), while no significant effect was shown by A or M.

First phoneme plosivity interaction with second position exception cost. To achieve the release of acoustic energy of a plosive first phoneme, the second phoneme must also be executed. On the hypothesis that the first phoneme is executed as soon as possible, the response time is determined by the resolution of the first phoneme if it is not plosive but by the resolution of the second

phoneme if the first phoneme is plosive. As such, any cost of second position exceptions should be exacerbated when the first phoneme is plosive; that is, there should be an interaction between whether the first phoneme is plosive and whether the position of irregularity is second, and this pattern has been observed (Cortese, 1998; Kawamoto, Kello, Jones, & Bame, 1998). This pattern was also present for D (0.08%), A (0.14%), and M (0.12%) but not U (0.00%) in our data.

Summary of Influence of Additional Factors

As the values for this final model in Table 5 show, these extra variables went some way to explain the remaining explicable variance in the mean item RTs for each participant. Although these improvements were rather small (a total of less than 4%) in terms of overall R^2, of the relevant (i.e., non-noise, non-first-phoneme) variance, together these variables explained up to an extra 11% (for D). Nevertheless, of this relevant variance, 65–70% remained unexplained after all 31 variables.

Which Are the Most Important Factors?

First phoneme by far and away predicted the most variance of a single factor in each of the participants' data, as it has done in previous mega-studies of naming but not lexical decision (e.g., Balota et al., 2004; Ferrand et al., 2011). Adding variance after first phoneme, accessibility (best choice of WF or CD, cubic splines) contributed the most for D (first phoneme plus accessibility: 58.70%), A (57.05%), and U (62.97%). For M, the single factor that (after first phoneme) increased variance accounted for most was PLD20 (first phoneme plus PLD20: 56.69%). However, for both M and U, if whammies and irregularities were combined to form a single position of spelling-sound abnormality variable (with six levels), this accounted for slightly more variance (M: 56.77%; U: 63.33%).

The Effects of Practice

Analogous to the problem we described with individual differences in the older mega-studies, we can only examine the variation in effects with practice (or indeed, any session-to-session change) for those effects that are known; for the intercept and the nine variables of our simpler regression model (from Figure 1), the variation over session is illustrated in Figure 2. There is very little evidence that practice increased the efficiency of word processing in the mean RTs, and the practice trends in the individual effects—mostly involving a slight weakening of the effects—are quite small relative to the variability. Weakening of the effects might be a form of response time homogenization (cf. Taylor & Lupker, 2001); that is, as participants become familiar with the structure of the stimulus set, their decision rule incorporates an expectation of the ideal RT; as the influence of this ideal becomes greater, RTs become more similar and item effects become smaller. There is some suggestion in the data that the inhibitory neighborhood effects are due to practice, possibly reflecting an adjustment to the precise visual properties of the experimental display. That it is due to practice does not diminish the implied requirement on models that the possibility of an inhibitory neighborhood size effect be explained. Similar changes appeared in first phoneme effects over the duration of the experiment.

Such practice effects are interactions between a word property and a session, and so the calculation of mean item RTs averages over such effects, meaning that they cannot directly contribute to the R^2 estimates for mean item RTs. While these practice effects are theoretically important and will account for variance at the trial level, they are not relevant to assessing models at the item level.

Computational Models

Finally, we calculated R^2 values for four computational models, the DRC (Coltheart et al., 2001), CDP+ (Perry et al., 2007), and Plaut et al.'s (1996) Simulations 1 and 4[8] (using cycles—a time-taken measure—for DRC and CDP+ but error scores for Plaut et al.'s, 1996, models; all simulated with their standard parameters), which are shown in Table 5. These are similar to the values for the three-variable regression model, and worse than those for the 17- and 31-variable regression models.

Discussion

We collected data that offer the most precise assessment of reading aloud at the level of the individual and, hence, the most stringent benchmark for modeling of reading. The structure of our data—with repetitions of each word for each participant—allowed us to perform analyses without problematic assumptions about individual differences (see Table 1). Removing these problematic assumptions made our estimate of the noise in the data more accurate than was previously possible. This more accurate noise estimate sets an accurate target R^2 for models to achieve. Using our regressions, we were able to explain between 56% and 69% of the variance in item means for each of our participants. Much of this explained variance was due to first phoneme—a factor usually only treated as a control variable—and some of the unexplained variance could (in the context of item means) be considered inexplicable noise. Of the variance due to neither noise nor first phoneme, our 31-variable regressions accounted for between 30 and 35%.

This was better than the R^2 achieved by DRC, CDP+, and PDP models (2–11% of this non-noise, non-first-phoneme variance), though in principle these models might be improved by adjustments to individual differences in the parameters (Adelman, Sabatos-DeVito, Marquis, & Estes, 2011). In fact, they performed similarly to a three-variable (log. frequency, length, neighborhood size) regression model that has often been used as a model comparison. However, this three-variable regression obtained only 7–16% of the variance that could not be attributed to first phoneme or noise, suggesting this is a very weak target for cognitive modeling endeavors.

Idiosyncrasies and Individual Differences

Analyzing individual participants will reveal their idiosyncrasies, and so typical patterns of performance will often not emerge. The well above average reading or reading experience of three of our participants (D, M, and U) appears to be the source of one set of differences from average performance; in contrast, A was more

[8] Neither Simulation 2 nor Simulation 3 was the best of the four Plaut et al. models for any participant.

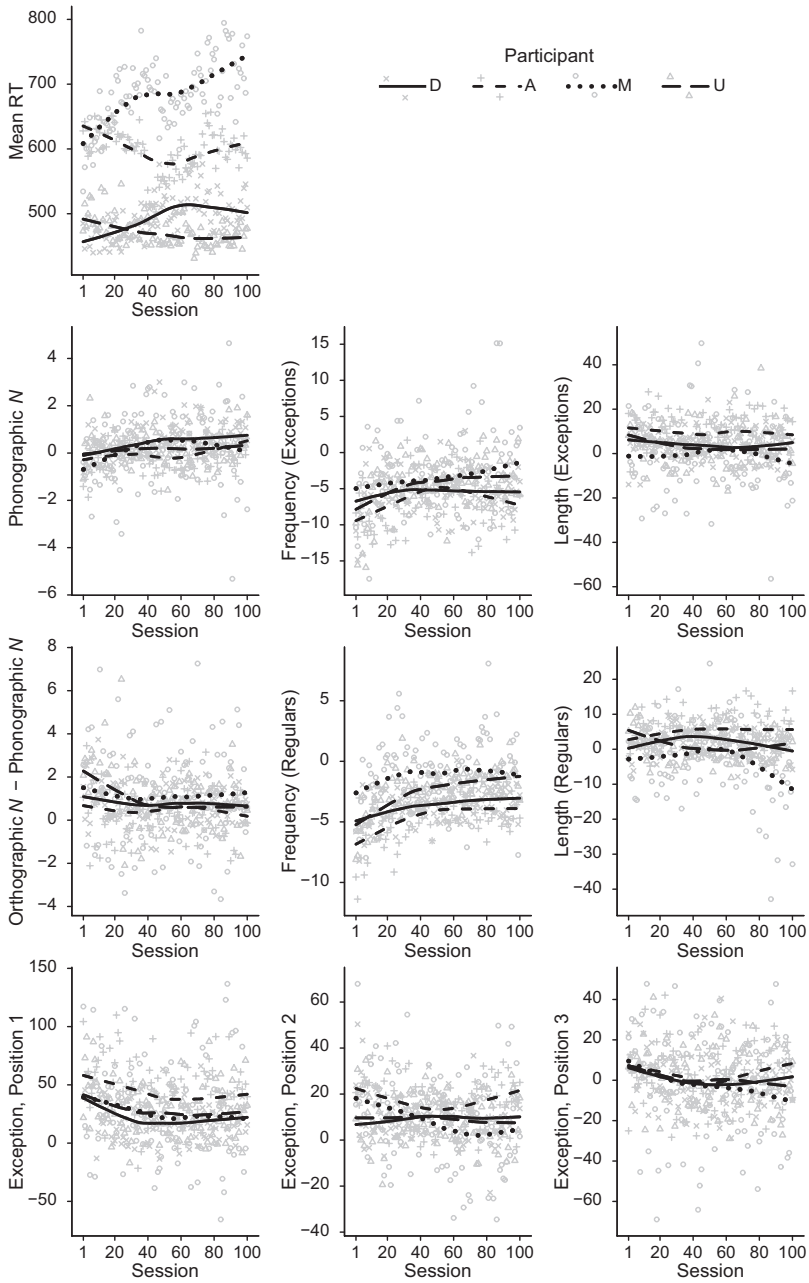

Figure 2. Per session response time (RT) and effect magnitudes (ms). Symbols represent estimates from models with estimates of effect of first phoneme shared over all sessions. Lines represent the loess ($\alpha = .667$) smooth of these points to illustrate trends.

similar to the typical undergraduate participant. These three participants showed both atypical (inhibitory) neighborhood effects (both in terms of *N* and the OLD20 and PLD20 measures) and no imageability effect. We have observed an association between vocabulary and neighborhood size effects in naming (Adelman, Sabatos-DeVito, & Marquis, YEAR), and there is also an association between reading skill and imageability effects (e.g., Strain & Herdman, 1999). In effect, better readers are more able to rely upon precise formal identification of individual words (i.e., lexical quality, e.g., Andrews & Hersch, 2010; Perfetti, 1992), which might explain why our three participants showed atypical effects.

One participant (M) showed further idiosyncrasies. He showed a relatively weak sensitivity to frequency, and he did not show inhibition from orthographic length. In our initial analyses, his orthographic length effect appeared to be facilitatory. However, this was in fact attributable to phonological length when this variable was added to the regression. That is, he showed a facilitatory phonological length effect. Such an effect was also pres-

ent after the orthographic length effect was partialled in Spieler and Balota's (1997) data (Adelman & Brown, 2008a) and at least for exception words for D, A, and U. That is, there are countermanding influences of orthographic and phonological length, which have probably contributed to the difficulty of finding unique variance from length effects for words in smaller studies (cf. Weekes, 1997). The orthographic influence may be letter-by-letter processing in at least one route (as in the DRC and CDP+), or it may be some other form of limited capacity processing, such as visual competition (simple division of processing resources over letters, e.g., Adelman, 2011) or a limited accuracy in the coding of longer visual sequences (Chang, Furber, & Welbourne, 2012). The phonological influence might relate to reduced competition in lexical selection at a phonological level for longer words, less overlapping representations at a phonological level for longer words, or—given the relationship between orthographic and phonological length—this may reflect some aspect of processing multiletter graphemes that is not captured by the whammying variable.

The idiosyncrasies that these participants show do nothing to invalidate our results. First, that participant A's performance was similar to previous studies indicates that nothing about our method was so unusual as to make the task different from other studies. Second, case studies have long since (e.g., Plaut & Shallice, 1993) been sources of data regarding visual word recognition; that we have no reason to believe the present participants have suffered neurological damage does not excuse theories from explaining these individuals' performance. Ultimately, models will come to quantitatively explain both what differs between individuals and how individuals come to differ; however, to do so, models must incorporate sensitivity to the same factors to which individuals are sensitive.

What About the Effects of Practice?

The main effect of practice (i.e., one affecting all words equally) is assumed to be within the grasp of models (by use of the average intercept over all sessions) by our adjustment for session mean. However, one might expect that certain variables' influences—particularly frequency's—would vary with practice.

At first glance, it might seem that our data are critically contaminated with practice effects, but in fact our data are less contaminated by practice than other studies for two reasons. First, reading aloud is an overlearned task—used heavily in the early stages of reading instruction—that is relatively immune to task-specific learning. As illustrated in the first panel of Figure 2, there is no overall trend for response times to become shorter with practice. Second, variability due to practice also exists in other databases because individuals differ in how practiced they are at reading. For instance, in Spieler and Balota's (1997) data set, they averaged together 31 observations of each word, each average containing one observation from each participant. In each of our four data sets, for each word we have averaged together 50 observations from the same participants at different points in time. On average, between one trial with a particular word and some arbitrary[9] other trial with that word, on average around 47,000 experimental trials—or slightly fewer than the number of words in *Slaughterhouse-Five*—intervened. In comparison to the differences in reading practice between one arbitrarily chosen undergraduate and another, this one-novel's-length difference seems quite minimal.

To be clear, by choosing to form each database from a single individual, we have chosen to move from having databases in which variability is contributed by trait individual differences and potentially large interindividual differences in practice to having databases (one per subject) in which variability is contributed by smaller intraindividual changes in practice over the course of the experiment but not by trait individual differences. Understanding such individual differences is, of course, important, but individual differences are a distinct problem to the one considered here, and so need studies of a different design to the present one (cf. Adelman et al., 2011).

As described above, we can only examine the variation in effects with practice for those effects that are known, as we have done in Figure 2. The theoretical implications of the observed changes—such as whether the changes are stimulus-specific or task-general (Dutilh, Krypotos, & Wagenmakers, 2011)—are beyond the scope of the present article, and the data are not optimal to examine this question, because there are no unpracticed baseline items.

However, they do cause our noise estimate to be an overestimate, because a replication could not contain all early or all late trials (whereas our simulated replications could). We therefore removed the session variation by further adjusting the observed RTs by interactions of the nine predictor variables with session number. The revised R^2 targets were higher by 0.01% for each participant. While these practice effects are theoretically important and will account for variance at the trial level, when models are assessed at the level of item means, the additional stability provided by trends in practice is negligible.

What About Sequential or Priming Effects?

From the viewpoint of a statistical criterion that averages RTs over different occurrences of a word, any differences between different occurrences that do not systematically recur are treated as noise; this includes effects that can be attributed to properties of the preceding trial or trials—including priming and changes in pathway control (cf. Reynolds & Besner, 2008)—when the order of trials is randomized. Moreover, even in a statistical analysis that incorporates such sequential effects, their inclusion does nothing to increase or decrease the estimated error of the observed mean item RTs as estimates of the underlying mean item RTs (although it reduces the error in the individual trial predictions). Model analyses can be used to attempt to remove these influences, but estimates based on conforming to a particular statistical model may distort the data in ways that it would be undesirable for cognitive models to attempt to mimic.[10]

[9] This calculation is for an *arbitrary* other trial, not the next trial, since we are treating pairs of sessions (with a single participant) in our experiment as analogous to pairs of participants in a more typical experiment (who do not come in order from least to most experienced): Our pairs of the same participant are more similar than a pair of different participants. The number of trials between subsequent trials with the same word might be relevant to arguments about priming: The average number of trials intervening between one trial with a word, and the next is, of course, 2,819.

This statistical point about sequential effects in a standard additive statistical model (where the effects are calculated relative to the mean over the data set) does not detract from the possibility that mechanistically sequential effects (which should be calculated relative to some other baseline) do ultimately explain some of the item-level variance, because they affect different items in different ways, that is, contribute a difference to the average. For instance, if items are responded to faster when the preceding item shares a contextual usage with them—a contextual priming effect—then items that occur in many contexts would show an advantage (i.e., this is one possible explanation of the contextual diversity effect shown by Adelman et al., 2006). This priming effect would show up in both the item mean and trial-by-trial analyses; to the extent that an effect of this type systematically shows up in the item means, it is included in our estimates of an explicable effect. Only those priming (or other sequential) effects that associate with some items more than others affect the item averages and so are considered to be open to a model that predicts these items means (and therefore are included in our target R^2s). While these kinds of effects are not responsible for the shortfall in R^2, these data do provide a unique resource for examining sequential and priming effects over long trial sequences in reading.

Word Naming and Reading

Word naming is but one task that is used to assess the processes of reading. Other tasks that are specifically focused on word-level variables include lexical decisions and perceptual identification. Higher level (e.g., grammatical) processing is often studied with eye-tracking measures. Of course, our results need not generalize perfectly to these other tasks. For instance, first phoneme is usually only found to be important for tasks involving spoken production, and accessibility measures like frequency have a much stronger influence on lexical decision. Ideally, analogous data sets will become available for these other tasks and measures.

Conclusion

Overall, the assessments reported here suggest that psychologists' current empirical and theoretical understanding of word naming is lacking. Our present understanding of the relevant factors reveals less than half of what is systematically due to word properties in word naming response times. Candidate areas for improvement include representations of semantics, visual similarity and articulatory duration, as well as the inclusion of morphological information. However, it is unlikely that these account for the majority of the remaining variance. In our opinion, it is likely that one or more major factors is currently overlooked. Testing such new ideas and finding new effects will be one of the major onward uses of these data. Neither practice nor sequential effects should be expected to improve predictions of item mean RTs, only individual trial RTs, and therefore would not substantially improve models on the criterion we use here. Moreover, attempts to model response times directly have been approaching a target that is simply far too low to be useful. That is, even if a model gave a detailed, quantitatively accurate account of how currently known word-level factors affect response times, the amount of variance systematically due to words but remaining unexplained would be so large as to render to the account uncompelling in our view. We hope the present data will provide some of both the impetus and the information needed to approach these unexplained properties of reading aloud; they give the most precise picture available of reading aloud by individual participants. These data and the criteria that they establish are critical for evaluating models of reading.

[10] In particular, the use of the best linear unbiased predictor for the random intercept associated with each word in a mixed effects model is misleading because these values reduce overall error by introducing bias at the item level, despite the impression the name gives: Such predictors are only unbiased on average over all words; this usage of the term *unbiased* is considered appropriate because they apply to random factors, which are not intended to be examined individually (because their identity is not considered part of the replicable experimental design by virtue of their designation as random).

References

Adelman, J. S. (2005). Regularity and length effects in word naming: A test of the dual route cascaded model. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty-seventh annual conference of the Cognitive Science Society* (pp. 57–61). Alpha, NJ: Sheridan Printing.

Adelman, J. S. (2011). Letters and time in retinotopic space. *Psychological Review, 118,* 570–582. doi:10.1037/a0024811

Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review, 14,* 455–459. doi:10.3758/BF03194088

Adelman, J. S., & Brown, G. D. A. (2008a). Methods of testing and diagnosing models: Single and dual route cascaded models of word naming. *Journal of Memory and Language, 59,* 524–544. doi:10.1016/j.jml.2007.11.008

Adelman, J. S., & Brown, G. D. A. (2008b). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review, 115,* 214–227. doi:10.1037/0033-295X.115.1.214

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17,* 814–823. doi:10.1111/j.1467-9280.2006.01787.x

Adelman, J. S., Sabatos-DeVito, M. G., & Marquis, S. J. (2010). [Individual differences in word naming: External predictors of item effects]. Unpublished raw data.

Adelman, J. S., Sabatos-DeVito, M. G., Marquis, S. J., & Estes, Z. (2011). *Individual differences in word naming: A mega-study, item effects, and dual-route cascaded models*. Manuscript submitted for publication.

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 802–814. doi:10.1037/0278-7393.15.5.802

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 234–254. doi:10.1037/0278-7393.18.2.234

Andrews, S. (2008). Lexical expertise and reading skill. *Psychology of Learning and Motivation, 49,* 249–281. doi:10.1016/S0079-7421(08)00007-8

Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General, 139,* 299–318. doi:10.1037/a0018366

Andrews, S., Woollams, A., & Bond, R. (2005). Spelling-sound typicality only affects words with digraphs: Further qualifications to the generality

of the regularity effect on word naming. *Journal of Memory and Language, 53,* 567–593. doi:10.1016/j.jml.2005.04.002

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

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., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,983 monosyllabic words. *Memory & Cognition, 29,* 639–647. doi:10.3758/BF03200465

Balota, D. A., & Spieler, D. H. (1998). The utility of item-level analyses in model evaluation: A reply to Seidenberg and Plaut. *Psychological Science, 9,* 238–240. doi:10.1111/1467-9280.00047

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

Besner, D. (1999). Basic processes in reading: Multiple routines in localist and connectionist models. In R. M. Klein & P. McMullen (Eds.), *Converging methods for understanding reading and dyslexia* (pp. 413–458). Cambridge, MA: MIT Press.

Bouma, H. (1971). Visual recognition of isolated lower-case letters. *Vision Research, 11,* 459–474. doi:10.1016/0042-6989(71)90087-3

British National Corpus Consortium. (2000). *British National Corpus* (World ed.) [CD-ROM]. Oxford, England: Humanities Computing Unit, University of Oxford.

Brysbaert, M., & New, B. (2009). Moving beyond Kuçera 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: Instruments & Computers, 41,* 977–990. doi:10.3758/BRM.41.4.977

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. doi:10.1177/1745691610393980

Chang, Y.-N., Furber, S., & Welbourne, S. (2012). "Serial" effects in parallel models of reading. *Cognitive Psychology, 64,* 267–291. doi:10.1016/j.cogpsych.2012.01.002

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

Cortese, M. J. (1998). Revisiting serial position effects in reading. *Journal of Memory and Language, 39,* 652–665. doi:10.1006/jmla.1998.2603

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., & 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

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

Dutilh, G., Krypotos, A.-M., & Wagenmakers, E.-J. (2011). Task-related versus stimulus-specific practice. *Experimental Psychology, 58,* 434–442. doi:10.1027/1618-3169/a000111

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53,* 134–140. doi:10.1037/h0045156

Estes, Z., & Adelman, J. S. (2008a). Automatic vigilance for negative words in lexical decision and naming: Comment on Larsen, Mercer, and Balota (2006). *Emotion, 8,* 441–444. doi:10.1037/1528-3542.8.4.441

Estes, Z., & Adelman, J. S. (2008b). Automatic vigilance for negative words is categorical and general. *Emotion, 8,* 453–457. doi:10.1037/a0012887

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., . . . Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from CHRONOLEX. *Frontiers in Psychology, 2,* 306. doi:10.3389/fpsyg.2011.00306

Harm, M. W. (2002). *Building large scale distributed semantic feature sets with WordNet* (CNBC Tech. rep. PDP.CNS.02.01). Retrieved from http://www.cnbc.cmu.edu/~mharm/research/tools/semantics_tr2002/index.html

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7,* 185–207. doi:10.3758/BF03212979

Kawamoto, A. H., Kello, C. T., Jones, R., & Bame, K. (1998). Initial phoneme versus whole-word criterion to initiate pronunciation: Evidence based on response latency and initial phoneme duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 862–885. doi:10.1037/0278-7393.24.4.862

Kristof, W. (1973). Testing a linear relation between true scores of two measures. *Psychometrika, 38,* 101–111. doi:10.1007/BF02291178

Kuçera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Masterson, J., & Hayes, M. (2007). Development and data for UK versions of an author and title recognition test for adults. *Journal of Research in Reading, 20,* 212–219.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38,* 39–41. doi:10.1145/219717.219748

Peereman, R., & Content, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory and Language, 37,* 382–410. doi:10.1006/jmla.1997.2516

Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145–174). Hillsdale, NJ: Erlbaum.

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

Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review, 9,* 542–549. doi:10.3758/BF03196311

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

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377–500. doi:10.1080/02643299308253469

Rastle, K., & Coltheart, M. (1999). Serial and strategic effects in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 482–503. doi:10.1037/0096-1523.25.2.482

Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 307–314. doi:10.1037/0096-1523.28.2.307

Rey, A., Courrieu, P., Schmidt-Weigand, F., & Jacobs, A. M. (2009). Item performance in visual word recognition. *Psychonomic Bulletin & Review, 16,* 600–608. doi:10.3758/PBR.16.3.600

Reynolds, M., & Besner, D. (2008). Contextual effects on reading aloud: Evidence for pathway control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 50–64. doi:10.1037/0278-7393.34.1.50

Sears, C. R., Siakaluk, P. D., Chow, V., & Buchanan, L. (2008). Is there an effect of print exposure on the word frequency effect and the neighborhood size effect? *Journal of Psycholinguistic Research, 37,* 269–291. doi:10.1007/s10936-008-9071-5

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568. doi:10.1037/0033-295X.96.4.523

Seidenberg, M. S., & Plaut, D. C. (1998). Evaluating word-reading models at the item level: Matching the grain of theory and data. *Psychological Science, 9,* 234–237. doi:10.1111/1467-9280.00046

Seidenberg, M. S., & Waters, G. S. (1989). Word recognition and naming: A mega study. *Bulletin of the Psychonomic Society, 27,* 489.

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,* 383–404. doi:10.1016/S0022-5371(84)90270-6

Shaoul, C., & Westbury, C. (2009). *A USENET corpus (2005–2009)*. Edmonton, Alberta, Canada: University of Alberta. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html

Sibley, D. E., Kello, C. T., & Seidenberg, M. S. (2009). Error, error everywhere: A look at megastudies of word reading. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the thirty-first annual meeting of the Cognitive Science Society* (pp. 1036–1041). Austin, TX: Cognitive Science Society.

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

Stanovich, K., & West, R. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24,* 402–433. doi:10.2307/747605

Strain, E., & Herdman, C. M. (1999). Imageability effects in word naming: An individual differences analysis. *Canadian Journal of Experimental Psychology, 53,* 347–359. doi:10.1037/h0087322

Taylor, T. E., & Lupker, S. J. (2001). Sequential effects in naming: A time-criterion account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 117–138. doi:10.1037/0278-7393.27.1.117

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General, 124,* 107–136. doi:10.1037/0096-3445.124.2.107

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence: WASI*. San Antonio, TX: Psychological Corporation.

Weekes, B. S. (1997). Differential effects of number of letters on word and nonword naming latency. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 50A,* 439–456.

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, NY: Touchstone Applied Science.

Appendix

Collection of Arousal and Valence Norms

Ratings were collected via Amazon Mechanical Turk (AMT), which is an online crowd-sourcing platform that allows registered Requesters (e.g., researchers) to post tasks for completion by registered Workers (e.g., research participants). AMT currently has approximately 500,000 Workers in 190 countries around the world. It is ideal for the collection of large-scale data sets in which participants provide many relatively simple ratings. AMT has been used extensively by psychological researchers in recent years, and its use has been validated (e.g., Buhrmester, Kwang, & Gosling, 2011). These ratings are available in the Supplemental Materials.

Participants

Two hundred sixty-four distinct participants contributed to our ratings. All participants were registered by Amazon to be in the United States and completed a demographic questionnaire that included a first language question; only those participants aged over 18 years and responding "English" to this question were permitted to continue. Participants whose ratings either were invariant or appeared random were excluded and replaced, so that each of the 2,820 words was rated by exactly 40 approved raters. Participants were paid \$0.05 (5 U.S. cents) for every 10 words that they rated. As described below, participants rated different numbers of words, at their discretion. Thus, payments ranged from \$0.05 to \$14.10, with an average payment of \$2.14 (SD = \$3.31).

Stimuli and Design

The 2,820 stimuli from the reading aloud experiment were split into 282 sets of 10 words. Each participant was permitted to rate as many or as few of these sets (without repetition) as they wished, up to a total of 40 participants per set. Participants completed between one and 282 sets (i.e., rated between 10 and 2,820 words), with an average of 43 sets (430 words) per participant. On each presentation of a set, the order of the words within that set was randomized.

Procedure

Participants who wished to take part in our study were instructed to complete a demographic questionnaire, which contained questions about gender, age, religion, income and first language, without being instructed as to the qualification criteria, which were at least 18 years of age and first language English. Those meeting these criteria were permitted to select to complete a task corresponding to a set of 10 words. On this first set, extended instructions were given, including examples of words of positive and negative valence, and high and low arousal. Participants were instructed to use the whole scale rather than just two or three different values, and to select the "don't know" option if they were unfamiliar with the given word. Depending on the browser software settings used by the participant, this first set was identified as either the first time the relevant page was visited (using the cookie mechanism) or when the page had not been reached using the automatic get-next-task feature of AMT. Otherwise, a brief reminder of the instructions was given that appeared with every set; in this case, an option was provided to review the longer instructions.

Each word from the list appeared in bold lowercase print above the center of a 7-point arousal scale and a 7-point valence scale, each implemented using radio buttons. A *don't know this word* option appeared above each of these scales. Once an option had been selected for both scales, a button below the scales was activated that allowed participants to move to the next word. Once an option had been selected for both scales for all 10 words in the set, a button below the final scale was activated to allow participants to submit their ratings to the AMT server. This then either returned the participant to the list of available tasks, or—if the participant selected the appropriate option in the AMT interface—to the next set of 10 words, which was selected randomly without replacement.

Received January 24, 2012
Revision received November 12, 2012
Accepted November 13, 2012 ■