Individual differences in reading aloud: A mega-study, item effects, and some models

James S. Adelman[a,*], Maura G. Sabatos-DeVito[b], Suzanne J. Marquis[a], Zachary Estes[c]

[a] *Department of Psychology, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, U.K.*
[b] *Department of Psychology, University of North Carolina at Chapel Hill, Davie Hall, Chapel Hill, NC 27599, U.S.A.*
[c] *Department of Marketing, Bocconi University, Via Roentgen 1, 20136 Milan, Italy*

Abstract

Normal individual differences are rarely considered in the modelling of visual word recognition — with item response time effects and neuropsychological disorders being given more emphasis — but such individual differences can inform and test accounts of the processes of reading. We thus had 100 participants read aloud words selected to assess theoretically important item response time effects on an individual basis. Using two major models of reading aloud — DRC and CDP+ — we estimated numerical parameters to best model each individual's response times to see if this would allow the models to capture the effects, individual differences in them and the correlations among these individual differences. It did not. We therefore created an alternative model, the DRC-FC, which successfully captured more of the correlations among individual differences, by modifying the locus of the frequency effect. Overall, our analyses indicate that (i) even after accounting for individual differences in general speed, several other individual difference in reading remain significant; and (ii) these individual differences provide critical tests of models of reading aloud. The database thus offers a set of important constraints for future modelling of visual word recognition, and is a step towards integrating

[*] Corresponding Author. Telephone: +44 (0) 24 761 50233. Fax: +44 (0) 24 765 24225
Email address: j.s.adelman@warwick.ac.uk (James S. Adelman)

such models with other knowledge about individual differences in reading.

Keywords: visual word recognition, word naming, individual differences, computational modelling, reading aloud

1. Introduction

The identification of individual words is one important component of reading, and one that has been subject to extensive empirical studies, and also theoretical work in the form of computationally explicit implemented models. One of the most extensively modelled tasks is word naming (or reading aloud), in which participants read aloud words or pseudowords presented in isolation. The major empirical phenomena that models have addressed are impaired reading — found in acquired and developmental dyslexia — and *item effects*. Item effects are comparisons between words that differ on some specific dimension, such as length or frequency, usually in terms of response times (RTs).

According to the developers of models such as the dual-route cascaded model (DRC; Coltheart, Curtis, Atkins & Haller, 1993; Coltheart, Rastle, Perry, Langdon & Ziegler, 2001) and the connectionist dual process model (CDP+; Perry, Ziegler & Zorzi, 2007; Zorzi, Houghton & Butterworth, 1998), and to those who compare models to data (e.g., Adelman & Brown, 2008a; Besner, 1999; Reynolds & Besner, 2002, 2004), the goal of such modelling is a complete[1] and detailed account of human visual word recognition, which is indexed by a precise correspondence to the observed data. Whilst our understanding of impairments to reading has informed theories of visual word recognition, the constraints from these data are at a relatively high level. Consequently, the finer details of word recognition processes have been more readily examined using item effects shown by the average of a population of (mostly unimpaired,

[1] Although as the science progresses, models will be in some way incomplete — for instance, these models lack semantic processes — the goal of completeness means that phenomena that the models were not designed to explain can be used to test these posited details.

young adult) undergraduate readers (but see Ziegler, Castel, Pech-Georgel, George, Alario & Perry, 2008, for an exception).

1.1. Average data and individual differences

1.1.1. Importance of individual differences

This concentration on average effects in skilled readers has been an important source of progress, but one could also raise the concern that this focus if it stood alone would be too narrow for models that seek to be a complete explanation of word recognition, because of individual differences in reading. One such concern is practical: Knowledge about word recognition can be applied to educational settings, including reading difficulties. If this knowledge included a theoretical understanding of individual differences in reading, this might be applied in the development of individual reading (and reading-related) education programs, tailored to the cognitive strengths and weaknesses of the individual student. A related concern is that the array of knowledge from developmental studies of individual differences in reading renders the scope of models that do not account for individual differences incomplete.

1.1.2. Misleading nature of average data

Moreover, it is well-known that average performance patterns can differ from the average's constituent patterns of performance (e.g., Brown & Heathcote, 2003; Estes, 1956).

1.1.3. Inferences from individual differences

In any case, patterns of individual differences in performance may simply form additional constraints on models, giving indications as to the correct accounts of the item effects that might not emerge from the average item effects. Such indications may come because of the implications of common loci of effects. For instance, some effects — length and position of irregularity — are attributed by models such as DRC and CDP+ to the left-to-right sequential processing in spelling-sound conversion. If this is the case, these effects should be susceptible to the same causes of individual differences, and those individuals who are particularly susceptible to one effect will also be particularly susceptible to the other, inducing a correlation in the sizes of the effects. In contrast, effects attributed to separate components, such as length and frequency might be expected to show no such relationship, or a negative relationship if there is some trade-off in emphasis between the components.

Using item RT effects nevertheless leaves a complex picture, because these effects are influenced by several parameters, of which at least some should reflect individual differences, and the values these parameters take might be correlated. Depending on these correlations (and other distributional properties), different patterns of correlations among effects might be found. For instance, a dual-route model might have routes whose strengths trade-off against one another, producing negative correlations among effects arising in the two routes, but it could instead have positively correlated route strengths due to general ability or speed, which would result in all effects correlating positively.

As a consequence, without data to constrain the distribution of parameters of a given model for different individuals, it is essentially impossible to give a general characterization of the patterns that model predicts. Future developments of models may allow many or all these parameters to be set by reference to data about other cognitive abilities (and others may be fixed for all participants). Ziegler et al. (2008) used tasks to measure individuals' levels of letter processing, (phonological) word processing and phoneme processing and added noise to the corresponding systems of the DRC to reflect deficits in these processes, with the magnitude of the noise varying parametrically with the magnitude of the deficit; however, this approach is currently incomplete as other parameters are also important for model performance and could vary across individuals.

The data that are available to be used to constrain any simulations are the to-be-explained reading aloud data. At first glance, this might seem to make it too easy for a model to fit the data. To be sure, if we ask a complicated model to adapt to one or two effects that may be summarized with a handful of data points, it may be so easy that any model that complicated could fit the data. If we instead ask a model to fit a *complex* of data, comprising of many effects and their magnitudes, the constraints imposed by the structure of the model may become more important than its flexibility (and this is true in general).

For instance, it may be true that for one set of parameter settings, a model produces some effect A, and for another parameter set, effect B, but there are no parameter sets for which both of A and B occur, whilst both effects are evidenced in the same data set. That the explanations instantiated by the model for these two effects are not *compatible* would be evidence against such a model. In practice, the precise cause of the model's problems will be difficult to isolate when this is the case: Attempts to select parameter sets that reproduce the data will show that the model does not predict one or other of the effects (and which it does not show may vary with slight variations in procedure). Such a pattern need not show that the model cannot predict the effect, but it nevertheless is symptomatic of an inconsistency between model and data. The size of each effect can also be considered a constraint for selecting parameter values, which exacerbates the just-described difficulty. As more and more effects are considered, identifying any particular type of model misspecification becomes even more difficult. The ability to detect problems with a model remains even if parameters are permitted to vary. To say this another way: A demonstration of *incompatibility* for a particular model would be a *falsification* of that particular model[2].

Examining individual differences not only allows us to take this approach further but also addresses a problematic assumption implied by ignoring such differences: A valid objection to rejecting a model that does not show effects A and B together is that there may be individual participants who show effect A and some others who show effect B, but none who show both, exactly as the model proposes. Only individual data can address such claims. Moreover, ex-

[2] A particular model here refers to a particular version of a model, not all models that might bear the name.

amining individual data allows more patterns to be added to each effect and its average size: First, the several patterns of compatible effect sizes of each participant, and second, the summary of this by way of the correlations of effects.

In summary, constraints from data on models come from data patterns interpreted as effects and the compatibility of these effects, which we may augment with patterns from individual differences because these give evidence of further (and less misleading) patterns of compatibility that models must address.

1.1.4. Insufficiency of comparisons of good and poor readers

It is thus important to attempt to isolate variation on several dimensions to distinguish different possible mechanisms involved in visual word recognition. Whilst reading success is the ultimate phenomenon of interest, item effects in word naming are side-effects of the processes involved in the visual word recognition component of reading, and these effects constrain the nature of these processes beyond what they achieve to how they achieve it. Various studies have examined how these effects differ between better and poorer readers (e.g., Ashby, Rayner & Clifton, 2005; Kuperman & Van Dyke, 2011) but several factors can lead to better or poorer reading in humans (see, e.g., McClung, O'Donnell & Cunningham, 2012) and in models. Moreover, there might be informative (co)variation in item effects that does not affect reading success per se.

1.2. Previous research

1.2.1. Variability in responses

Ziegler et al. (2008) have successfully used external cognitive predictors to modify parameters controlling noise injected into the DRC model so as to predict type of words that would be susceptible to word naming errors. Moreover, there is a further line of individual differences evidence relating to responses (rather than their latencies) that constrains models: Andrews & Scarratt (1998) found that when several individuals were asked to read the same nonwords, many stimuli had multiple pronunciations. Pritchard, Coltheart, Palethorpe & Castles (2012) recently compared DRC and CDP+ on a new data set showing similar results. Zevin & Seidenberg (2006) obtained comparable levels of variation in PDP simulations by using different (randomly selected) sequences of words in the training trials for different individuals. Thus, they attributed the whole effect to *unsystematic* individual differences. However, this seems unlikely to be the correct explanation. The magnitude of difference shown is similar to that which can be explained by *systematic* individual differences introduced by reading instruction (Thompson, Connelly, Fletcher-Flinn & Hodson, 2009). Moreover, there are likely additional differences due to cognitive ability (McClung et al., 2012; Kuperman & Van Dyke, 2011); and there may be variability in responding that occurs within individuals (in none of these studies did a participant read the same stimulus twice) rather than being due to individual differences.

1.2.2. Individual differences in item effects and external predictors

Some other studies have directly addressed the RT differences in word recognition tasks. One such line of research seeks to link measures of orthographic knowledge or experience to lexical decision performance; there is evidence (Chateau & Jared, 2000; Sears, Siakaluk, Chow & Buchanan, 2008) that print exposure predicts the strength or quality of orthographic-lexical processes in lexical decision as assessed by neighbourhood size and frequency item effects. That the neighbourhood size and frequency effects are both predicted by the same variable suggests of a common locus related to the strength or quality of orthographic-lexical processes. In turn, this suggests these effects should be positively correlated.

A related line of research addresses a similar issue, but with form priming effects: Better spellers show inhibitory priming effects where poorer spellers show facilitatory ones, consistent with an account in terms of precision of lexical representations (Andrews & Hersch, 2010). Vocabulary has also been linked to word length effects (Butler & Hains, 1979), and frequency effects in both first and second languages (Diependale, Lemhöfer & Brysbaert, 2013); and more generally Yap, Balota, Sibley & Ratcliff (2012) found that higher vocabulary was associated with more rapid responding, and lesser sensitivity to item effects.

1.2.3. Individual differences in item effects and relations among them

More similarly to our present approach, Yap et al.'s (2012) analysis of word naming and lexical decision latencies from the English Lexicon Project (ELP: Balota, Yap, Cortese, Hutchison, Kessler, Loftis, Neely, Nelson, Simpson & Treiman, 2007) also included an analysis of the correlations of item effects (or rather their principal components). One important question they addressed was whether two distinct processes — linked to larger (word-sized) and smaller (grapheme-sized) units — might be seen to trade-off across individuals. They found no such evidence, but their measure of the process with smaller units was based on effects not necessarily due to the smaller unit process in implemented models such as the DRC (Coltheart et al., 2001). Moreover, the ELP followed the mega-study approach of seeking to obtain good estimates of average performance of essentially the whole of English, with individual differences analyses as a secondary goal (see also Balota, Cortese, Sergent-Marshall, Spieler & Yap, 2004; Balota, Yap, Hutchison & Cortese, 2012). Some of the ways in which the ELP was designed to address the primary goal and makes it useful for other types of analysis did however mean that the individual differences analyses Yap et al. could perform did not permit a direct comparison with model simulations. First, participants were not assessed on the same words as one another, which weakens the comparability of effect magnitude estimates from different individuals. Second, most words in the experiment were polysyllabic because most English words are polysyllabic; whilst this makes the sample more representative, many predictors and most models are only applicable to monosyllables. This limitation on the scope of the models meant that Yap et al. could not directly address these models. Third, assessment was based on principal components of item variables, not directly upon theoretically interesting item effects because lexical variables underlying item effects are intercorrelated in the language. Such a procedure might, however, subsume effects with different causes in a single principal component[3]. Finally, adjustments for general speed could not be based on individual participant reliabilities, because participants did not read the same word more than once. Our observation that existing databases were not optimal for the kind of individual differences modelling that interested us motivated the collection of new data regarding major item effects for later modelling.

1.3. Major item effects

Certain item effects are so ubiquitous as to be taken as basic effects that a model must incorporate, whilst others are important because they highlight contrasts between theoretical positions. The list of variables that have been claimed to affect word naming response times is extensive, so this discussion focusses on those that have received a modelling response and whose source in a given model would be important for understanding individual differences (and so were included in the experiment described below; for a longer list, see Adelman, 2012). We will argue that individual differences among these readers can also be informative about the mechanisms of reading. Such individual differences will be seen in theoretically interesting item effects, and not only overall performance.

1.3.1. Frequency

Response times are shorter for more frequent words than less frequent words (e.g., Forster & Chambers, 1973). In models that posit parallel activation of individual units for each word (*localist* models), the difference may reflect a lower required threshold of activation for frequent words (e.g., Morton,

[3] Correlations among effect sizes across people are suggestive of cognitive locus in visual word recognition. But the correlations used in the principal components analysis are correlations among lexical variables across words, wich are suggestive of locus in language evolution, not locus in visual word recognition.

1964), or a biasing input that inhibits activation of infrequent words (e.g., Coltheart et al., 2001). In models that learn parallel connections without such individualised units (*distributed representation* models), more relevant weight adjustments occur for frequent words (e.g., Seidenberg & McClelland, 1989). Other models suggest that more frequent words have higher priority in a sequential search of known lexical items (e.g., Forster, 2012). In terms of individual differences, effects that correlate positively with the size of the frequency effect might be related to a process sensitive to details of individual words; such a process is a lexical process or route. Those that correlate negatively might be related to another process that complements and trades off with the lexical process.

1.3.2. Length and lexicality

Nonwords and to a lesser extent low frequency words are read more slowly if they are long (have many letters) than if they are short (e.g., Weekes, 1997). Dual-route theorists have interpreted this as due to a left-to-right nonlexical process (composing a pronunciation on the basis of the components of its spelling) that is slower to complete for longer stimuli, and is more influential for stimuli with an absent or weak lexical route contribution to pronunciation (Coltheart et al., 2001). Alternative interpretations include articulation preparation being sensitive to utterance length and reduced quality of the input code for longer words due to information compression (Chang, Furber & Welbourne, 2012); these mechanisms affect different stages of the word naming process, and so are not mutually exclusive. Nonwords are also read slower than words, because of their lack of a stored representation. The size of this advantage for words will depend on the relative efficiency and strength of the two routes: A stronger lexical route and weaker nonlexical route should lead to a greater lexicality effect in favour of words. Whether individual differences in the length effect can be modelled as differences in the efficiency of spelling-sound translation (in a manner compatible with other effects) will be informative as to its cause.

1.3.3. Neighbourhood size

The neighbourhood of a word is the set of words that are similar to it — its neighbours — and the neighbourhood size (often denoted N, following Coltheart, Davelaar, Jonasson & Besner, 1977) is the count of these neighbours. A variety of types of neighbourhood size have been considered, most commonly the orthographic (or Coltheart's) N, where a neighbour is defined as a word that can be created by replacing one letter with another in the same position (e.g., BOG is a neighbour of DOG). On balance, words with more neighbours are read faster (Andrews, 1997; Mathey, 2001) albeit with several qualifications. The effect is somewhat surprising insofar as identifying a word uniquely should be harder if it has many neighbours with which it could be confused. This effect arises in models because similarly spelled words are usually similarly pronounced thus providing partial support for the correct pronunciation, although Andrews (1992) offered an alternative explanation in terms of top-down support for letter identities. Adelman & Brown (2007) conducted analyses of megastudies that suggested the effect was limited to phonologically similar orthographic neighbours (the *phonographic* neighbours, Peereman & Content, 1997), consistent with the phonologically-based interpretation. Indeed, the number of phonological neighbors (i.e., ignoring orthography) appears to have a unique influence (Mulatti, Reynolds & Besner, 2006; Yates, 2005), although this may require orthographic overlap (Grainger, Muneaux, Farioli & Ziegler, 2005). Given the strong correlations among these variables, we did not select items to manipulate them orthogonally for the present study. The extent to which neighbors are activated is one way in which individual differences in reading skill have been found to manifest in other paradigms (Andrews & Hersch, 2010; Sears et al., 2008).

1.3.4. Consistency

The kind of association of similar phonological forms for similar orthographic forms described above could be considered in more general terms as spelling-sound consistency. Most commonly, such consistency has been considered empirically in terms of rime consistency, the extent to which words of one syllable that have the same bodies (the end of the orthographic word from the orthographic vowel onwards) have the same rimes (the end of the phonological word from the vowel onwards). For instance, MINT and LINT rhyme and so are consistent, or *friends*, but PINT is inconsistent with them, and is an *enemy* of both. Words that have many friends and few enemies are read quicker (e.g., Jared, McRae & Seidenberg, 1990). This may be explained through partial support from similar words, although it has alternatively been proposed that body-sized units may be used in a nonlexical conversion process (e.g., Coltheart, 1980, 2012) of a type that is usually implicated in effects of regularity. The relationship of individual differences in consistency effects to those in neighbourhood size and regularity effects ought to constrain its interpretation.

1.3.5. Regularity and Position of Irregularity

Such regularity effects give an alternative or additional reason that a word like PINT might be read slowly: If there is a nonlexical conversion that translates spelling to sound using (grapheme-sized) sublexical rules such as I → /ɪ/, then this mechanism will produce the incorrect pronunciation, which would conflict with or fail to support the correct (known) pronunciation. Words that follow such rules, known as *regular words*, are read more quickly than *exception words* that do not. Moreover, exception words whose departure from the rule is to the left in the word exhibit a greater slowing relative to their controls than do those whose irregularity is further to the right (Rastle & Coltheart, 1999). One major debate in the reading aloud literature has surrounded whether such effects can be subsumed under processes sensitive to various forms of consistency (e.g., Zorzi, 2000) or require a rule-based approach.

The relationship between regularity and consistency effects is important to ascertain their interpretations, and the position of irregularity effect's relationship to length effects gives a test of their supposed common locus.

1.4. Models

Models that have been implemented to make quantified predictions about word naming are the dual-route cascaded model, the connectionist dual-process model and a variety of backpropogation models.

1.4.1. Dual-route cascaded model

The DRC (Coltheart et al., 1993, 2001) as a dual-route theory combines two routes: a lexical route that uses knowledge of specific words, retrieving a stored pronunciation, and a nonlexical route that parses stimuli into units smaller than the word to determine a pronunciation. Its lexical route is based on an extension of the interactive-activation and competition model (McClelland & Rumelhart, 1981). Its nonlexical route uses a set of classical rules for converting graphemes (letters or short sequences of letters that represent individual phonemes) into phonemes; these are applied in a temporally left-to-right fashion. The nonlexical rule route is responsible for length effects, the regularity effect, and the regularity effect's sensitivity to position. To the extent that the strength or speed of this route might vary, we would expect the magnitude of these three effects to be controlled by this strength/speed and so correlate positively with another.

Frequency effects have their locus in the lexical route, as do neighbourhood effects (Reynolds & Besner, 2002), so naïvely we would infer these should positively correlate. Lexicality effects relate to the advantage conferred by the lexical route to words, and so should also be in this group, although this may not be the only influence on lexicality effects.

The situation with consistency effects is less clear. Whilst some consistency effects (those of Jared, 1997) have been explained as nonlexical route confounds, Coltheart (2012) has suggested that parameter modifications might make the model predict consistency effects observed without such confounds via the lexical route in the same manner as neighbourhood effects. Exploring parameter sets to match individual participants will discover whether this is feasible.

In any case, these can only be tentative predictions because (i) verbal analyses may not accurately reflect the properties of the mechanisms in the model; (ii) simulations with arbitrary parameter values need not reflect the behavior of the model in other areas of the parameter space; and (iii) exhaustive analysis of the parameter space is computationally infeasible.

1.4.2. Connectionist dual-process model

The CDP+ (Perry et al., 2007) uses a similar set of lexical components to the DRC, but its nonlexical components are based on a two-layer connectionist learning procedure (in constrast to DRC's symbolic rules), as well as having slightly different phonological representations and decision rule. The connectionist learning procedure implies that this system uses information about how often different phonemes associate with particular graphemes, rather than associating individual rules to graphemes. At the level of assigning effects to components, this more graded nonlexical route is the source of consistency effects. As such, consistency effects clearly should be more associated with length and exception effects than lexical frequency effects. Otherwise, our (tentative) expectations regarding the CDP+ would be quite similar to those for the DRC.

1.4.3. Backpropagation ("PDP") models

A variety of models have used the connectionst backpropagation learning rule to learn connection weights in a multilayer network with hidden units to model various aspects of visual word recognition (e.g., Plaut, McClelland, Seidenberg & Patterson, 1996). These are commonly known as parallel-distributed processing — or PDP — models (although this is arguably a broader term), and are conceptualised within a 'triangle' framework described by Seidenberg & McClelland (1989). This framework posits three explicit representations of stimuli — orthographic, phonological, semantic — that are each linked to the other by way of distributed processing (in hidden units), diagramatically forming a triangle. Stimuli are read aloud by influences coming from the two routes linking orthographic and phonological representations: the direct route (via only hidden units) and the semantically-mediated route.

Coltheart (2012, Table 1.1) lists six variants of this type of model that have been used to model various effects. Coltheart points out that these models have each been developed to model a relatively circumscribed set of effects that does not always subsume those of its predecessors, and, moreover, that differences between a model and its predecessors are not always justified by reference to a better explanation of the particular phenomena the later model explains. Thus there is no current model that represents the sum knowledge of the best model of its class (to which its predecessors are an approximation). This contrasts with the approach of DRC and CDP+ modelers, whose later models are presented as replacements of previous models as the state of the art, justified by improvements in overall compatibility with the data.

This difference reflects a difference in modeling goals: For the most part, PDP modelers have argued that modeling is and should be used to elucidate general principles (Seidenberg & Plaut, 2006) and to show that these can lead to a small number of empirically observed phenomena of current key interest. For this reason, compatibility is only evidenced between models similar at the level of general principles, with lower-level detail free to vary for convenience. In contrast, such lower-level detail is considered to be of theoretical interest by other modelers (including Coltheart et al., 2001; Perry et al., 2007), because such details affect the predictions the models make, and so presumably are critical to understanding the precise behavior of humans across a variety of situations and phenomena. Thus (in the absence of an interacting factor that influences a parameter; e.g., Rastle & Coltheart, 1999, Exp. 2) compatibility is sought within a single model with no changes. For PDP modellers, the explanation of a phenomenon is created by the modeler and necessarily precedes and motivates the model, which validates the explanation if it produces the phenomenon. For other modellers (in this context, Coltheart and colleagues and Perry and colleagues), analysis and examination of a model may give the explanation for a phenomenon (whilst examination of phenomena may also suggest mechanisms a model requires); such an explanation is validated by the validation of the model by producing several phenomena.

The current approach to individual differences relies upon having a single model — except for the possibility of parameter changes — for different effects to see what the detailed mechanisms of that model imply about individual differences. This is only a coherent goal (i) if the model is an account of the effects in which individual differences are being examined, and (ii) if the model is in fact a hypothesis about a complex of mechanisms. This has not been the case for the PDP models listed by Coltheart (2012), and so it makes little sense to investigate them in this manner[4]. Thus, we could not simultaneously model individual differences in these effects with these PDP models because no single PDP model explains all these effects: each phenomenon has its own explanation within the broader framework, rather than them sharing mechanisms in a single computational model. Our modelling, therefore, does not include a PDP model.

1.5. *The present study*

We provide the first study designed to examine individual differences in item effects in a way that can directly assess models. We sought to apply to word naming the correlational logic of individual differences and models described above (§1.1.3), mindful that verbal analysis of models is simplistic compared to the actual function of the model processes. We collected new data to examine individual differences in the major item effects — frequency (e.g., Forster & Chambers, 1973), irregularity (e.g., Rastle & Coltheart, 1999; Seidenberg, Waters, Barnes & Tanenhaus, 1984), consistency (e.g., Glushko, 1979; Jared et al., 1990), length (e.g., Weekes, 1997), and neighbourhood size (e.g., Andrews, 1992), and the correlations between them. We also examined the ability of two existing implemented models of reading aloud — the DRC (Coltheart

[4] One recent backpropagation model has gone some way towards including key effects addressed by no previous such model (Chang et al., 2012), but it does not simulate response times (it produces error scores that do not relate directly to processing time).

et al., 2001) and CDP+ (Perry et al., 2007) models — to account for these differences in simulations. To the extent that the models cannot account for the data, our results would identify problems with these models and suggest how they might be corrected.

2. EXPERIMENT

2.1. Method

2.1.1. Participants

Participants ($N = 100$) were 32 staff and 54 students of the University of Warwick, supplemented by 14 others who responded to local advertisements. Ages ranged from 17 to 55 years (mean 28 years, median 23 years). All were tested as part of a larger study and were paid (prorated over the whole study) approximately £10 for their participation in this experiment (ca. U.S. $15). A questionnaire was administered to gather information about the participants' language, academic and work history and to screen for any potential reading or learning disabilities. All participants were native English speakers and all but two reported normal or corrected-to-normal vision (the others reporting essentially monocular vision with normal corrected acuity).

2.1.2. Stimuli

Details follow of the 592 monosyllabic words and 93 nonwords that were used to measure key marker effects in visual word recognition including: frequency, length (and lexicality), neighbourhood size, position of irregularity (and regularity), and consistency (and regularity). For the first three effects, new stimulus sets were selected using CELEX (Baayen, Piepenbrock & Gulikers, 1995). For the remaining two effects, stimulus sets were taken from Rastle and Coltheart (1999, Exp. 1) and Jared (2002, Exp. 1). Due to programming error, the 26 stimuli (of which 18 were exception words) that were in both Rastle and Coltheart's and Jared's experiments were repeated, and data from the second trial with a given word in any session overwrote that for the first.

Table 1: Mean stimulus characteristics (standard deviations in parentheses) for words selected to test frequency effects in naming.

	1 ppm		3 ppm		9 ppm		27 ppm		81 ppm		$F(4,80)$	Sig.
CELEX												
Freq.	20.24	(3.16)	67.24	(15.70)	187.53	(44.28)	546.62	(45.02)	1586.19	(318.39)		
$\log_e$ Freq.	3.00	(0.16)	4.19	(0.21)	5.21	(0.22)	6.30	(0.08)	7.35	(0.18)		
SUBTLEX												
Freq.	2.10	(2.62)	7.20	(8.38)	13.06	(10.24)	68.31	(123.44)	106.30	(106.99)		
$\log_e$ Freq.	0.15	(1.15)	1.27	(1.30)	2.24	(0.89)	3.37	(1.22)	4.30	(0.85)		
Orth. *N*	10.85	(4.85)	11.10	(6.28)	11.95	(4.59)	11.00	(6.55)	11.57	(4.86)	0.275	*ns*
PhGr. *N*	8.10	(3.40)	8.57	(5.95)	10.00	(4.31)	8.57	(5.81)	7.67	(4.42)	1.184	*ns*
Consist.	0.93	(0.10)	0.93	(0.17)	0.96	(0.07)	0.94	(0.17)	0.92	(0.17)	0.272	*ns*

Note — *F*-values are for a repeated-measures ANOVA (on the basis of matched quintuples) predicting the lexical statistic from the frequency band. Freq. stands for frequency. CELEX frequencies are based on raw values (per 17.9 million); SUBTLEX frequences are per million. ppm stands for parts (i.e. tokens) per million. Orth. *N* stands for orthographic neighbourhood size. PhGr. *N* stands for phonographic neighbourhood size. Consist.: Rime consistency calculated on a types basis (friends per friends plus enemies).

Frequency stimuli. 21 quintuplets of words were selected from five CELEX frequency bands, defined by the ranges 15–26, 51–107, 132–308, 479–657, and 1087–2649 (raw frequency; per 17.9 million words), each designed to be 3 times (or 1.1 $\log_e$-units) more frequent than the last. Similar differences between bands were shown with the SUBTLEX frequency norms (Brysbaert & New, 2009). Within each quintuplet, words were matched on orthographic length and onset, and as closely as possible on orthographic *N* and rime consistency. All words were regular according to DRC rules. Summary statistics are given in Table 1, and the items are listed in Appendix A.

Neighbourhood stimuli. 42 pairs of words were selected with one member of each pair low on neighbourhood size (both orthographic and phonographic) and the other high on neighbourhood size. Items in each pair were matched on onset and orthographic length and approximately matched on total CELEX frequency. All were regular according to DRC rules. After data collection, SUBTLEX frequencies were examined, which revealed that two stimuli (BITCH

Table 2: Mean stimulus characteristics (standard deviations in parentheses) for words selected to test neighbourhood effects in naming.

	Low *N*		High *N*		$F(1,39)$	Sig.
Orthographic *N*	4.85	(1.96)	15.68	(3.97)		
Phonographic *N*	3.68	(1.72)	12.65	(3.25)		
Phonological *N*	19.75	(13.09)	29.43	(9.75)		
Rime Consistency	.96	(.14)	.93	(.14)	1.12	*ns*
CELEX						
Frequency	745.3	(1833.2)	787.4	(1980.8)	0.01	*ns*
$\log_e$ Frequency	4.93	(1.84)	4.99	(1.78)	0.02	*ns*
SUBTLEX						
Frequency	35.7	(96.1)	42.1	(129.3)	0.12	*ns*
$\log_e$ Frequency	1.66	(1.95)	2.02	(1.77)	2.48	*ns*

Note — *F*-values are for a repeated-measures ANOVA (on the basis of matched pairs) predicting the lexical statistic from the neighbourhood condition. CELEX frequencies are based on raw values (per 17.9 million); SUBTLEX frequences are per million. Rime consistency was calculated on a types basis (friends per friends plus enemies). *N* stands for neighbourhood size.

and NOPE) were of substantially higher spoken frequency than the stimuli to which they were paired (BEARD and NOUN); these pairs were excluded from analysis.

Summary statistics (for the 40 analysed pairs) are given in Table 2, and the items are listed in Appendix B.

Length stimuli. 31 triplets of three, four, and five-letter monosyllabic words were selected. In each triplet, words were matched on onset and length and approximately matched on CELEX frequency and rime consistency; SUBTLEX frequency was also not significantly associated with length. Neighbourhood sizes could not be matched. All stimuli were regular according to DRC rules. Summary statistics are given in Table 3 and items are listed in Appendix C. 93 nonwords whose regular pronunciation was monosyllabic were chosen to match the word stimuli in length and phonemic onset (when pronounced regularly).

Table 3: Mean stimulus characteristics (standard deviations in parentheses) for words selected to test length effects in naming.

	3 letters	4 letters	5 letters	$F(2,60)$	Sig.
CELEX					
Frequency	629.66 (634.79)	396.61 (515.93)	414.32 (522.42)	0.658	*ns*
$\log_e$ Frequency	5.27 (1.33)	5.29 (1.27)	5.33 (1.27)	0.498	*ns*
SUBTLEX					
Frequency	37.19 (64.76)	24.99 (37.18)	40.38 (85.99)	0.798	*ns*
$\log_e$ Frequency	1.60 (2.54)	1.66 (2.17)	1.74 (2.30)	1.365	*ns*
Rime Consistency	0.96 (0.05)	0.95 (0.13)	0.98 (0.06)	0.792	*ns*
Orthographic N	16.32 (4.32)	12.03 (4.25)	4.13 (1.77)	103.230	$p < .001$
Phonographic N	13.52 (4.07)	9.36 (4.03)	3.61 (1.65)	67.146	$p < .001$

Note — F-values are for a repeated-measures ANOVA (on the basis of matched triples) predicting the lexical statistic from the length (as a factor). CELEX frequencies are based on raw values (per 17.9 million); SUBTLEX frequences are per million. Rime consistency was calculated on a types basis (friends per friends plus enemies). N stands for neighbourhood size.

Consistency stimuli. These were taken from Jared (2002, Exp. 1). The stimuli were 160 low-frequency monosyllabic words, divided into 80 inconsistent words (40 exception words, 40 regular words) and 80 consistent words (all regular according to DRC rules). Consistency was defined on the basis of word bodies. The 80 regular-consistent words were divided into four groups of 20 and matched to the 40 exception-inconsistent and 40 regular-inconsistent words for length, initial letter and phoneme, frequency, mean summed frequency of friends and mean bigram frequency.

Position of irregularity stimuli. These were selected from Rastle and Coltheart (1999, Exp. 1). They included 88 monosyllabic, three- to six-letter exception words from the CELEX database whose divergence from the DRC's grapheme-phoneme conversion rules occurred at the first letter position (20), second letter position (39), or third letter position (29). Each exception word was matched to one of 88 regular words on number of letters and initial phoneme, and all stimuli were low frequency.

2.1.3. Apparatus

The stimuli were presented on a Sony CPD-G200 17" monitor driven by an NVIDIA GeForce 7025 graphics card. Responses were collected via a Plantronics Audio 370 gaming headset with microphone attached to an Ensoniq 5880 AudioPCI sound card. To code response (voice onset) times, the third author conducted a visual and auditory inspection of wave forms on the basis of an estimate output by custom voice key software written by the first author for the analysis of Adelman, Marquis, Sabatos-DeVito & Estes's (2013) data. Error responses, those where no single onset could be identified, and those where the initial phoneme was realised in a non-standard manner (e.g., /θ/ for /t/) were excluded; for words, all dictionary-listed variants were accepted as correct, and for nonwords, any pronunciation that appeared to be reasonably constructed by rule or analogy from the properties of English words was accepted as correct.

2.1.4. Procedure

Participants attended three separate sessions, each on a different day. During each session, all 711 stimuli[5] were presented in newly randomised order. Each stimulus was presented on the computer screen for 1500 ms; the intertrial interval varied between 600 and 1000 ms. Participants wore a gaming headset with microphone, which recorded vocal responses. Participants were instructed to read the stimuli aloud as quickly and clearly as possible as they appeared. They were warned that some stimuli would be unfamiliar, and in those cases, they should take their best guess.

2.2. Results

Our empirical analyses of the data covered the analysis of the replication of the major item effects; the replicability of these effects over sessions; a com-

[5] This number is made up of 105 (5 × 21) frequency stimuli; 84 (2 × 42) neighbourhood stimuli; 93 (3 × 31) word length stimuli and 93 nonword length stimuli; 160 Jared stimuli; and 196 (2 × 88) Rastle and Coltheart stimuli. The number 711 thus includes the 26 stimuli that were common to the Rastle and Coltheart and Jared stimuli and were erroneously repeated.

parison with a past mega-study; estimation of effects for individuals and the effects' (test-retest) reliabilities; observation of the effect distributions; and observation of the correlations of the item effects.

2.2.1. Replication of standard effects

As a check on our manipulations, we compared the conditions for each of them using by-subjects and by-items ANOVAs. The condition means associated with these comparisons are in the first numerical column in Table 4, and a summary of the pattern of significance is in Table 5. Two-tailed p-values are reported in the text except when F is very small or $p < .001$ (disambiguated by the F-value).

Frequency effect. A frequency effect, such that more common words are read more quickly, has been observed repeatedly in word naming (e.g., Forster & Chambers, 1973), though there are some alternative suggestions for the underlying causative factor (e.g., Adelman, Brown & Quesada, 2006; Cortese & Khanna, 2008).

Such an effect of frequency was significant by subjects, $F_1(4, 396) = 33.59$, and by items, $F_2(4, 80) = 11.16$.[6]

The bulk of the effect was carried by the linear trend (in log. frequency), $F_1(1, 396) = 123.74$, $F_2(1, 80) = 40.60$, with the departure from linearity being significant only by subjects, $F_1(3, 396) = 3.54$, $p = .015$, $F_2(3, 80) = 1.34$, $p = .266$.

Overall, the frequency effect replicated previous studies.

Neighbourhood size. Greater orthographic neighbourhood size has often been associated with shorter naming latencies, especially in English (see Andrews, 1997; Mathey, 2001, for reviews), although other neighbourhood measures

[6] The 1 and 3 ppm bands clearly did not differ from one another, $F < 0.04$ in both analyses, and nor did the 9 and 27 ppm bands, $F < 0.7$. The difference between the 27 and 81 ppm bands was significant by subjects, $F_1(1, 396) = 7.35$, $p = .007$, but not by items $F_2(1, 80) = 2.55$, $p = .114$, but the difference between 9 and 81 ppm was significant in both analyses, $F_1(1, 396) = 12.19$, $F_2(1, 80) = 4.08$, $p = .047$.

	new data	SB97	DRC	CDP+	DRC-FC
Frequency					
1 ppm	552	469	555	554	553
3 ppm	552	460	551	552	550
9 ppm	540	457	547	550	549
27 ppm	538	456	543	548	548
81 ppm	532	453	540	548	549
Neighbourhood size					
few	544	464	545	548	547
many	542	463	546	548	547
Word length					
3 letters	542	454	544	551	549
4 letters	551	458	549	553	551
5 letters	550	466	552	555	553
Nonword length					
3 letters	587	n/a	593	593	597
4 letters	609	n/a	604	602	601
5 letters	631	n/a	622	616	612
Jared (2002) stimuli					
Exception (E>F)	561	479	558	550	557
Controls	544	462	546	547	545
Exception (F>E)	548	473	556	547	552
Controls	534	464	542	543	543
Inconsistent (E>F)	562	468	550	549	548
Controls	545	463	549	549	548
Inconsistent (F>E)	527	470	532	534	532
Controls	532	470	533	532	532
Rastle & Coltheart (1999) stimuli					
Exception Pos. 1	617	500	591	581	586
Controls	546	464	552	555	553
Exception Pos. 2	563	472	557	549	553
Controls	548	462	548	549	549
Exception Pos. 3	553	490	559	549	556
Controls	552	484	553	552	552

Table 4: Condition means (ms) for data and simulations. SB97 = Spieler & Balota (1997). DRC = Dual-route cascaded model. CDP+ = Connectionist dual process model. DRC-FC = DRC with frequency-weighted connections.

have been proposed (e.g., Adelman & Brown, 2007; Yarkoni, Balota & Yap, 2008).

The slight advantage for words with dense neighbourhoods was only significant with a one-tailed correction by subjects, $F_1(1, 99) = 2.91$, $p = .091$, and not at all by items, $F_2(1, 39) = 0.74$, $p = .394$.

The neighbourhood size effect interacted with age, $F_1(1, 98) = 4.31$, $p = .041$, $F_2(1, 39) = 5.69$, $p = .022$, with less facilitation for older participants,

		new data	SB97	DRC	CDP+	DRC-FC
Effects						
New stimulus sets						
Frequency		✔	✔	✔	✔	✔
Neighbourhood size		s	**0**	-SI	-S	S
Word Length		✔	✔	✔	✔	✔
Nonword Length		✔	NW	✔	✔	✔
Lexicality		✔	NW	✔	✔	✔
Length × Lexicality		✔	NW	✔	✔	✔
Jared (2002) stimuli						
Exception (E>F)		✔	✔	✔	S	✔
Exception (F>E)		Si	✔	✔	S	✔
Inconsistent (E>F)		✔	**0**	S	S	**0**
Inconsistent (F>E)		-S	**0**	-S	S	**0**
Rastle & Coltheart (1999) stimuli						
Exception Pos. 1		✔	✔	✔	✔	✔
Exception Pos. 2		✔	✔	✔	**0**	✔
Exception Pos. 3		**0**	**0**	✔	-Si	Si
Position × Regularity		✔	✔	✔	✔	✔
Correlations	Dir.					
Exception $\sim$ Pos. Irr.	+ve	✔	ID	✔	✔	✔
Frequency $\sim$ Neigh.	+ve	✔	ID	✘	**0**	✔
Lexicality $\sim$ Neigh.	-ve	✔	ID	**0**	**0**	**0**
Lexicality $\sim$ Word Len.	+ve	✔	ID	✔	✔	✔
Lexicality $\sim$ Exc. (J.)	+ve	✔	ID	**0**	✘	✔
Frequency $\sim$ Nonword Len.	+ve	✔	ID	✔	✘	✔

Table 5: Patterns of significance for effects and correlations of interest. ✔ indicates an effect was significant on a two-tailed test ($p < .05$) in the expected direction or a correlation was significant in the indicated direction (Dir.); where both by-subjects and by-items analyses were used, ✔ means significant in both analyses. A negative (**-**) sign indicates an effect was in the unexpected direction. S and I mean significant by-subjects and by-items, respectively. s and i mean $.05 < p < .1$ in by-subject and by-item analyses, respectively (or equivalently $p < .05$ one-tailed). +ve = positive. -ve = negative. **0** indicates an effect or correlation was not significant. ✘ indicates a correlation was significant in the opposite direction to that indicated. SB97 = Spieler and Balota (1997). NW = effect not estimated due to absence of nonwords in experiment. ID = effect not estimated due to non-publication of individual trial data.

consistent with the finding of Spieler & Balota (2000).

Length and lexicality. Nonwords are typically read aloud more slowly than words, and longer stimuli are typically read more slowly than shorter stimuli; for monosyllabic stimuli it has been argued the effect is only present for nonwords (Weekes, 1997), but analyses of mega-studies with monosyllabic words do find a small effect (e.g., Adelman & Brown, 2007, 2008a; Balota et al., 2004).

Words. The effect of length such that longer words were read more slowly was significant by subjects, $F_1(2, 198) = 13.04$, and by items $F_2(2, 60) = 6.64$, $p = .002$.

Partialing the confounded variable of orthographic neighbourhood size from item analyses gave adjusted mean RTs of 542, 551 and 549 ms for three, four, and five letter words, respectively. After the (facilitatory) effect of

orthographic neighbourhood size, $F_2(1, 59) = 5.97$, $p = .018$, was partialled as a preceding step, the effect of length remained significant $F_2(2, 59) = 3.61$, $p = .033$. Orthographic neighbourhood size was not, however, significant, $F_2(1, 59) = 0.10$, when added as the last step.

In the uncovaried analyses, the linear trend was strong, $F_1(1, 198) = 19.35$, $F_2(1, 60) = 9.06$, $p = .004$, but augmented by the quadratic trend, $F_1(1, 198) = 6.72$ $p = .010$, $F_2(1, 60) = 4.21$ $p = .045$. The covaried analysis had a weaker linear trend $F_2(1, 59) = 3.05$, $p = .086$, with a quadratic trend in evidence, $F_2(1, 59) = 4.16$, $p = .045$.

Overall, the length effect for words was in the expected direction and non-linear.

Nonwords. The effect of nonword length was significant by subjects, $F_1(2, 198) = 130.99$, and by items $F_2(2, 60) = 15.10$.

Partialing the confounded variable of orthographic neighbourhood size from item analyses gave adjusted mean RTs of 590, 610 and 629 ms for three, four, and five letter words, respectively. After the (facilitatory) effect of orthographic neighbourhood size, $F_2(1, 59) = 15.59$, was partialled, the effect of length remained significant $F_2(2, 59) = 7.68$, $p = .001$. Orthographic neighbourhood size did not, however, explain unique variance, $F_2(1, 59) = 1.50$, $p = .226$, when added as the last step.

The effect was wholly due to the linear trend: $F_1(1, 198) = 261.88$, $F_2(1, 60) = 30.19$, covaried $F_2(1, 59) = 16.59$, with no role for the quadratic trend: $F_1(1, 198) = 0.10$, $F_2(1, 60) < 0.01$, covaried $F_2(1, 59) = 0.01$.

There was a clear, linear, effect of length for the nonwords.

Lexicality. The words were read more quickly (548 ms on average) than the nonwords (611 ms on average), $F_1(1, 495) = 966.57$, $F_2(1, 150) = 236.23$; and the interaction with length was significant $F_1(2, 495) = 29.31$, $F_2(2, 150) = 7.30$.

The lexicality effect was replicated, as was its interaction with length.

Regularity and Consistency. Words whose mapping from spelling-to-sound is unusual are read more slowly than those whose mapping is typical or common (e.g. Glushko, 1979; Seidenberg et al., 1984). Such an effect is indexed in terms of following rules (regular words) or not following rules (exception or irregular words), namely *regularity*, or in terms of a graded construct involving the number of matching exemplars (friends) and mismatching exemplars (enemies), namely *consistency*; both regularity and consistency appear to have independent effects, though the relative frequencies of friends and enemies is important for both, with the cost being reliant on the enemies being higher frequency (Jared, 2002).

Regularity. For Jared's (2002) stimuli, the exception words whose enemies were higher frequency than their friends were read more slowly than their matched control words, $F_1(1, 99) = 56.37$, $F_2(1, 19) = 10.92$, $p = .004$, as were the exception words whose enemies were lower frequency then their friends, though the by-items analysis was only significant one-tailed, $F_1(1, 99) = 43.78$, $F_2(1, 19) = 3.43$, $p = .080$ (unadjusted for direction).

The regularity effect was therefore replicated, though the effect was less dependent on the frequencies of friends and enemies than in previous studies.

Consistency. For Jared's (2002) stimuli, the regular-inconsistent words whose enemies were higher frequency than their friends were read more slowly than their matched control words, $F_1(1, 99) = 79.37$, $F_2(1, 19) = 11.97$, $p = .003$. The opposite was true for the regular-inconsistent words whose friends were higher frequencies than their enemies: Responses to these words were faster than to their matched controls, though only signficantly so in the by-subjects analysis, $F_1(1, 99) = 5.70$, $p = .019$, $F_2(1, 19) = 0.91$.

This pattern of a strong consistency effect only for those words whose enemies were higher frequency than their friends replicated the previously observed data.

Position of irregularity. The latency cost for exception words is greater for those whose irregularities are toward the beginning of the word (Coltheart & Rastle, 1994; Rastle & Coltheart, 1999), the effect being strong for position 1 irregularities, moderate for position 2 irregularities and for position 3 absent in smaller studies (Rastle & Coltheart, 1999; Roberts, Rastle, Coltheart & Besner, 2003), or present but weak in mega-study analyses (Adelman & Brown, 2007).

Words with an irregular grapheme-phoneme correspondence in the first letter produced longer RTs than their matched controls, $F_1(1, 99) = 437.80$, $F_2(1, 19) = 18.89$. Words whose irregularity was in the second letter also showed an exception cost relative to their controls, $F_1(1, 99) = 75.35$, $F_2(1, 38) = 9.55$, $p = .004$. No such effect was observed for the third position irregulars whose RTs were equivalent to their controls', both $F < 0.5$.

The interaction between position and regularity was significant, $F_1(2, 495) = 256.73$, $F_2(2, 85) = 15.41$.

This replicated the pattern typically found in standard-sized studies of the position of irregularity effect.

2.2.2. *Effects by Session*

Given the use of multiple sessions, we examined the effect of session. Overall, responses were faster in session 1 (544 ms) than session 2 (556 ms) than session 3 (562 ms), $F_1(2, 198) = 10.06$, $F_2(2, 1368) = 409.01$.

Frequency. The effect of frequency replicated in every session: session 1, with RTs of 535, 539, 529, 521 and 518 ms, $F_1(4,396) = 19.33$, $F_2(4,80) = 8.87$; session 2, with RTs of 554, 552, 541, 539 and 532 ms, $F_1(4,396) = 11.56$, $F_2(4,80) = 6.69$; session 3, with RTs of 563, 561, 546, 550 and 545 ms, $F_1(4,396) = 10.39$, $F_2(4,80) = 6.27$. Moreover, the effect did not interact with session $F_1(8,1188) = 1.31$, $p = .236$, $F_2(8,160) = 0.97$.

Neighbourhood size. The weak neighbourhood effect was not significant in any individual session (session 1: 532 vs. 529 ms.; session 2: 549 ms vs. 547 ms; session 3: 554 vs. 550 ms). It also did not interact with session, both $F < 0.5$.

Word Length. The effect of word length was present in each session, though it was not significant by items in the second session: session 1, RTs of 527, 525 and 536 ms (for 3, 4 and 5 letter words, respectively), $F_1(2,198) = 7.53$, $F_2(2,60) = 6.55$, $p = .003$; session 2, RTs of 545, 549 and 552 ms, $F_1(2,198) = 4.00$, $p = .020$, $F_2(2,60) = 1.93$, $p = .153$; session 3, RTs of 551, 560 and 558 ms, $F_1(2,198) = 4.98$, $p = .008$, $F_2(2,60) = 3.72$, $p = .030$. There was no interaction with session, both $F < 0.6$.

Nonword Length. The effect of nonword length was present in all sessions: session 1, with RTs of 577, 602 and 623 ms, $F_1(2,198) = 84.18$, $F_2(2,60) = 12.88$; session 2, with RTs of 583, 600 and 622 ms, $F_1(2,198) = 70.21$, $F_2(2,60) = 14.86$; session 3, with RTs of 592, 608 and 618 ms, $F_1(2,198) = 38.91$, $F_2(2,60) = 8.56$. This effect did interact by session, $F_1(4,396) = 6.28$, $F_2(4,120) = 4.77$, $p < .001$, because the longer nonwords were immune to the overall slowing by session.

Lexicality. The effect of lexicality and its interaction with length were repeated in every session: session 1 (534 and 606 ms), $F_1(1,495) = 859.95$, $F_2(1,150) = 236.96$ for lexicality and, $F_1(2,495) = 23.79$, $F_2(2,150) = 6.56$, $p = .002$ for the interaction; session 2 (551 and 608 ms), $F_1(1,495) = 569.01$, $F_2(1,150) = 192.30$ for lexicality and, $F_1(2,495) = 21.64$, $F_2(2,150) = 6.99$, $p = .001$ for the interaction; and session 3 (558 and 610 ms), $F_1(1,495) = 507.74$, $F_2(1,150) =$

181.95 for lexicality and, $F_1(2, 495) = 7.87$, $F_2(2, 150) = 3.36$, $p = .037$ for the interaction.

The lexicality effect clearly interacted with session, $F_1(2, 198) = 32.29$, $F_2(2, 120) = 7.86$. The lexicality by length by session interaction was only significant by subjects, $F_1(4, 396) = 3.25$, $p = .012$, $F_2(4, 240) = 2.18$, $p = .072$. Again, both these effects were driven by the immunity of the longer nonwords to the slowing showed by other conditions.

Regularity and Consistency. Regularity. The cost for the high-frequency enemies exception words was consistent across sessions: session 1 (545 and 529 ms), $F_1(1, 99) = 26.99$, $F_2(1, 19) = 10.12$, $p = .005$; session 2 (560 and 545 ms), $F_1(1, 99) = 19.10$, $F_2(1, 19) = 8.34$, $p = .009$; session 3 (568 and 552 ms), $F_1(1, 99) = 15.88$, $F_2(1, 19) = 10.10$, $p = .005$. This effect did not interact with session, both $F < .2$.

The cost for the lower-frequency enemies exception words was less consistent over sessions: session 1, (536 and 520 ms), $F_1(1, 99) = 30.02$, $F_2(1, 19) = 6.22$, $p = .022$; session 2, (548 and 536 ms), $F_1(1, 99) = 11.88$, $F_2(1, 19) = 3.11$, $p = .094$; session 3, (552 and 542 ms), $F_1(1, 99) = 7.75$, $F_2(1, 19) = 0.86$, $p = .365$. This effect did not, however, interact with session, both $F < .8$.

Consistency. The regular-inconsistent words whose enemies were higher frequency than their friends were read more slowly than their matched control words across all sessions: session 1 (548 and 533 ms), $F_1(1, 99) = 26.32$, $F_2(1, 19) = 4.95$, $p = .038$; session 2 (560 and 528 ms), $F_1(1, 99) = 17.40$, $F_2(1, 19) = 8.67$, $p = .008$; session 3 (570 and 540 ms), $F_1(1, 99) = 25.58$, $F_2(1, 19) = 14.66$, $p = .001$. There was no interaction with session, both $F < .2$.

Those whose friends were higher frequency than their enemies did not show this difference in a consistent manner: session 1 (514 and 521 ms), $F_1(1, 99) = 5.15$, $p = .025$, $F_2(1, 19) = 2.17$, $p = .157$; session 2 (530 and 528 ms), both $F < 0.09$; session 3 (534 and 540 ms), $F_1(1, 99) = 3.95$, $p = .050$, $F_2(1, 19) = 0.92$. The interaction with session was not significant, $F_1(2, 198) = 1.93$, $p = .197$, $F_2(2, 38) = 1.74$, $p = .189$.

Position of irregularity. The first position exception cost was present in all sessions: session 1 (608 and 535 ms), $F_1(1, 99) = 238.47$, $F_2(1, 19) = 18.43$; session 2 (610 and 544 ms), $F_1(1, 99) = 259.47$, $F_2(1, 19) = 19.17$; session 3 (611 and 551 ms), $F_1(1, 99) = 202.90$, $F_2(1, 19) = 18.68$. An interaction with session was significant by items (and nearly so by subjects), $F_1(2, 198) = 3.01$, $p = .052$, $F(2, 30) = 4.05$, $p = .025$, driven by the immunity of the exception words to the slowing shown by other conditions.

The second position exception effect was consistent over sessions: session 1 (547 and 535 ms), $F_1(1, 99) = 25.16$, $F_2(1, 38) = 7.12$, $p = .011$; session 2 (561 and 548 ms), $F_1(1, 99) = 35.50$, $F_2(1, 38) = 8.33$, $p = .006$; session 3 (566 and 554 ms), $F_1(1, 99) = 17.91$, $F_2(1, 38) = 7.41$, $p = .010$, with no interaction, both $F < .4$.

A third position exception effect was absent in all sessions: session 1 (540 and 542 ms), session 2 (552 and 549 ms), and session 3 (557 and 559 ms), all $F < 1.05$, and this did not interact with session, both $F < 1.3$.

A position by irregularity interaction was present in each session: session 1, $F_1(2, 495) = 138.67$, $F_2(2, 85) = 15.45$; session 2, $F_1(2, 495) = 118.06$, $F_2(2, 85) = 14.83$; and session 3, $F_1(2, 495) = 98.60$, $F_2(2, 85) = 15.48$. The position by irregularity by session interaction was not significant, $F_1(4, 396) = 1.301$, $p = .269$, $F_2(4, 170)$, $p = .070$.

Summary of session effects. Overall, participants became slower over sessions. This might reflect motivation (these three half-hour sessions were embedded within three of nine one-hour sessions of the larger experiment), or a strategic change with experience. If the motivation explanation is correct, the lack of slowing for the slowest conditions (4-letter nonwords, 5-letter nonwords, first position exceptions) might be an artifact of more trials falling foul of the 1500 ms response recording cutoff. A strategic explanation would involve a shift of the response criterion (cf. Taylor & Lupker, 2001).

2.2.3. Comparison with previous mega-study data

To further establish that our findings were not unusual, we also extracted the mean RTs for these words from Spieler & Balota's (1997) undergraduate word naming data set. Of the 592 unique words in our data set, 534 were in the Spieler and Balota experiment. We examined our major effects of interest in terms of the condition means, presented in Table 4, and by-items analyses like those we used on our data.

More frequent words were read more quickly than less frequent words, $F_2(4, 80) = 5.71$, as in our data set. No significant effect of neighbourhood size was found, $F_2(1, 39) = 0.01$, as in our data set. Longer words were read more slowly than shorter words, $F_2(2, 60) = 10.93$, as in our data set. From Jared et al.'s (1990) stimuli, exception words were read more slowly than their regular controls if they had enemies more frequent than their friends, $F_2(1, 16) = 11.82$, $p = .003$, as in our data. This was also the case when the friends were more frequent, $F_2(1, 19) = 5.76$, $p = .027$; this effect was only significant in our data if a one-tailed correction was applied. Inconsistent words tended to be read more slowly than their consistent controls when their enemies were more freqeunt than their friends, but this was not significant, $F_2(1, 16) = 1.41$, $p = .252$. In our data, and in Jared's original study, the effect in this direction was however significant. No difference was found between consistent and inconsistent stimuli when friends were of higher frequency than enemies, $F_2(1, 18) = 0.00$, as in our data. From Rastle & Coltheart's (1999) stimuli, exception words were read more slowly than controls when the irregularity was in first position, $F_2(1, 7) = 14.87$, $p = .006$, and in second position, $F_2(1, 20) = 4.68$, $p = .043$, but no such effect was significant for third position, $F_2(1, 16) = 1.35$, $p = .263$. These were all the same as in our data, as was the presence of the interaction between position and irregularity (such that earlier irregularities produce more slowing), $F_2(2, 43) = 5.32$, $p = .009$.

In summary, of these 11 tests on Spieler & Balota's (1997) data, 7 showed the same signficant effects as in our data and 2 showed the same non-significant effects as our data (though one was marginal by-subjects in our data). One further test was significant where this was significant by-subjects but only marginal by-items in our data, and the last was significant in our data, where the effect was only numerically in the same direction in Spieler and Balota's data. Overall, our data were not unusual.

2.2.4. Raw effect estimation and reliability

Our purpose in collecting these data was to examine individual differences in these effects, so it was necessary to summarise each effect for each participant as a single number. A description of how we did so, and the reliability of the effects follows.

The reliability of our session-average estimates was calculated on the basis of the average test-retest reliability and the Spearman-Brown formula. This reliability is used to adjust estimated variance of the effect because the estimated effects will show variability that is greater than that of the underlying effects; indeed, the presence of measurement error or noise implies variability will appear even on measures for which participants are identical. Significance levels are omitted in this section: Observed correlations for 100 participants are significant at the 5% level if their absolute values are at least .197.

Frequency. The frequency effect was estimated by applying the linear contrast to participants' band means, reversing coding so that the expected effect was positive (positive effects if rare words were read slower than common words). Cross-session correlations were (1 and 2) .101, (1 and 3) .388, and (2 and 3) -.046. The mean correlation of .148 yields an estimate for the reliability of the average of .342.

In light of the negative correlation of second and third sessions, we examined methods that might improve the reliability. Using individual item frequencies was counterproductive (.120, .314, and -.119). Solutions involving different contrasts did improve reliability but did not improve correlations with other effects. We did not explore solutions that involved the full range of stimuli as these might artificially induce common variance with other effects.

Neighbourhood size. The simple difference between the high neighbourhood size and low neighbourhood size stimuli was used as the neighbourhood effect estimate. Cross-session correlations were (1 and 2) .064, (1 and 3) .102, and (2 and 3) .072. The mean correlation of .080 implied a reliability for the average of .206; given the weak effect, this is not surprising.

Length. The length effects were estimated as the linear contrast. Cross-session correlations for the word length effect were (1 and 2) .334, (1 and 3) .192, and (2 and 3) .216. The mean correlation of .247 implies a reliability for the average of .497. Cross-session correlations for the nonword length effect were (1 and 2) .424, (1 and 3) .270, and (2 and 3) .313. The mean correlation of .336 implies a reliability for the average of .603.

Lexicality. The lexicality effect was calculated as the difference between word and nonword responses for four-letter stimuli, to avoid structural correlations with the length effects (the four-letter condition has no weight in the calculation of the linear contrast for length). The cross-session correlations were (1 and 2) .613, (1 and 3) .632, and (2 and 3) .345; the mean of these is .530, implying a reliability for the average of .772.

Regularity (Jared). The exception cost was calculated from the Jared (2002) stimuli as the difference of the mean of both exception conditions from their matched controls, and from the Rastle and Coltheart (1999) stimuli as the difference of the mean of the first and second position exceptions and their matched controls. The cross-session correlations for the Jared measure were (1 and 2) -.014, (1 and 3) .138, and (2 and 3) .206, whose mean correlation of .110 implies a reliability for the average of .270. The cross-session correlations for the Rastle and Coltheart measure were (1 and 2) .309, (1 and 3) .332, and (2 and 3) .256, whose mean correlation of .299 implies a reliability for the average of .561.

Consistency. The consistency effect was calculated as the difference of the inconsistent stimuli with high-frequency enemies and their matched controls, because only the advantage for consistent words was present in the data. The cross-section correlations were (1 and 2) .113, (1 and 3) -.079, and (2 and 3) .012, whose mean of .016 implies a poor reliability for the average of .047.

Regularity (Rastle & Coltheart). The cross-session correlations for the measure of the regularity effect based on the Rastle and Coltheart stimuli were (1 and 2) .309, (1 and 3) .332, and (2 and 3) .256, whose mean correlation of .299 implies a reliability for the average of .561.

Position of irregularity. The position of irregularity effect was calculated as the double difference (interaction) subtracting the position 2 regularity effect (position 2 exceptions minus matched controls) from the position 1 regularity effect (position 1 exceptions minus matched controls). The cross-session correlations were (1 and 2) .264, (1 and 3) .338, and (2 and 3) .165, whose mean of .256 implies that the average has a reliability of .508.

2.2.5. Effect distributions

Of immediate concern for the study of individual differences is whether participants do in fact differ on the variables of interest. Figure 1 illustrates (crosses and solid lines) the distribution of the effect estimates calculated as described above. These distributions are, however, somewhat misleading as they include both true differences between participants and differences that arise from trial-to-trial variation (noise); indeed these plots would show variability even if participants were identical. We therefore estimated the (true-score) variance by multiplying the observed variance of an effect by our estimate of that effect's reliability (as indicated by classical test theory); these are illustrated in Figure 1 (dashed lines) by normal distributions with this estimate of the variance.

We also performed analyses based on the ANOVA Subject × Effect interaction to establish that effects varied across subjects; the statistics are superimposed on the corresponding panels of Figure 1. Significant variation was found for all but two of the effects; the interaction involving neighbourhood size approached significance, whereas consistency showed no such evidence of reliability.

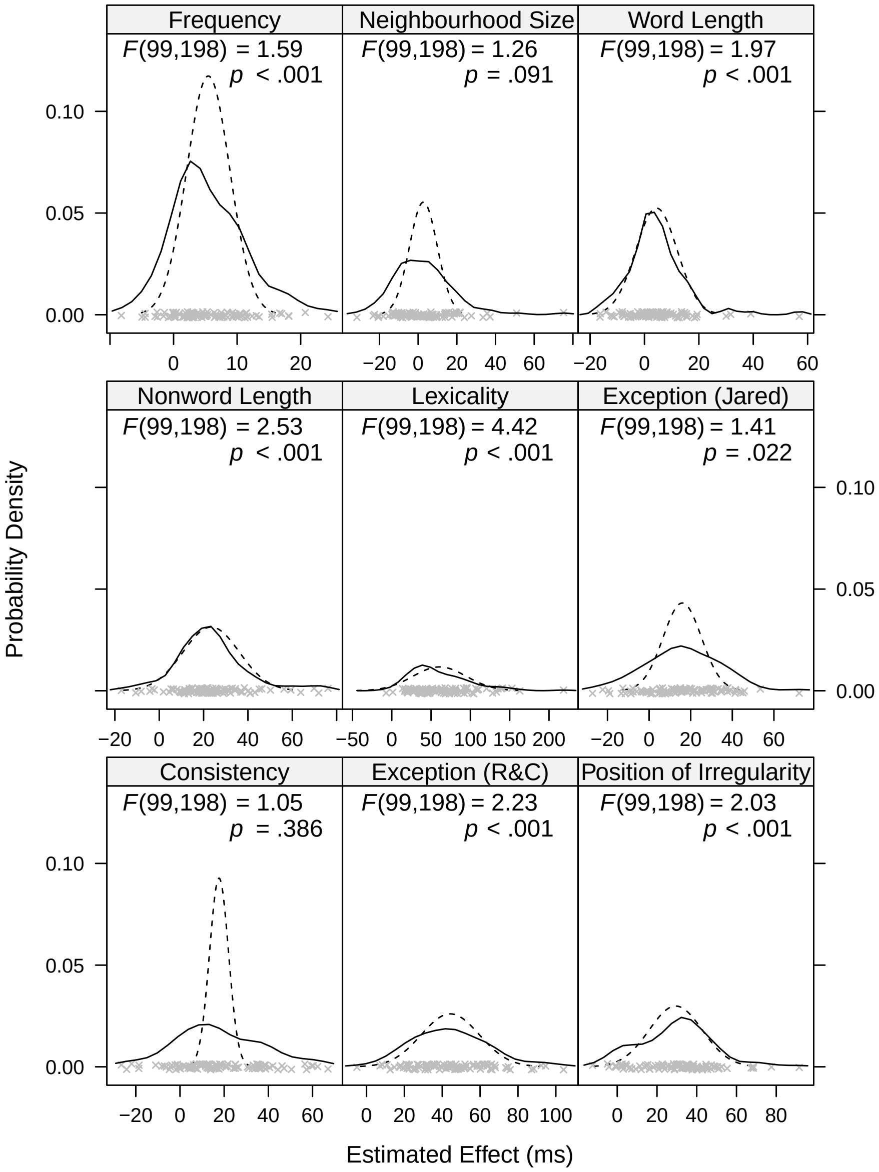
Variability in the neighbourhood size effect in naming is, however, evidenced from other more sensitive data (four participants reading 2820 words 50 times each: Adelman et al., 2013, indeed, these participants show inhibitory or null effects). Moreover, as the next section shows, both consistency and neighbourhood size participate in significant correlations, which should not be the case if there were no participant-based variability.

2.2.6. Effect correlations

We calculated the correlations between the effects as calculated above; for ease of interpretation, the frequency effect was reverse coded, so that a greater benefit from frequency was indicated by a larger number; these are presented in Table 6 (above the diagonal), as are correlations adjusted for attenuation using the test-retest reliabilities (below the diagonal). The high incidence of estimates of a perfect correlation for the consistency variable could indicate that the reliability estimate is an underestimate, or that the population correlation coefficient is undefined because there is no true variability. The perfect correlation estimated between the two measures estimated from Rastle & Coltheart's (1999) stimuli (exception and position of irregularity) is suggestive that there is more variability in some of the conditions than others: For instance, the position 1 exception effect may carry most of the variability in both effects. The overall high incidence of significant correlations is suggestive of a general speed influence that could be readily modelled by modifying the coefficient linking model cycles to RTs, rather than being indicative of the qualities of differing processes, as we go on to discuss.

Adjusting for general speed. A core difficulty in assessing individual differences in response time effects in terms of correlations is that a correlation will be induced if there is a general speed coefficient that is a multiplier on the central processes; in many models this is the slope for converting cycles into response

Figure 1: Distributions of effects: Crosses (×) indicate observed effect estimates, calculated as described in the text. Solid lines are Gaussian kernel density estimates of the distribution of effect estimates. Dashed lines are normal distributions estimating the distribution of the true effects, that is, with standard deviation adjusted for reliability. The reported *F*-statistic is for the Subject × Effect interaction (tested with the Subject × Effect × Session term as the MSE), indicating whether there is evidence for individual differences in the effects. Frequency is reverse-coded (relative to the regression coefficient) so that positive values indicate rare words were read slower than common words. R&C = Rastle and Coltheart.



		1.	2.	3.	4.	5.	6.	7.	8.	9.
		-Frq	Nei	WLen	NWLen	Lex	ExcJ	Cons	ExcRC	PoI
1.	(neg.) Frequency	—	**.223**	**.211**	**.471**	**.378**	**.227**	**.226**	**.252**	**.271**
2.	Neighbourhood Size	.840	—	-.046	.158	-.105	.104	.068	.047	.083
3.	Word Length	.512	-.143	—	**.398**	**.452**	.141	.078	-.002	.115
4.	Nonword Length	1.000	.448	.727	—	**.585**	**.287**	**.283**	**.331**	**.233**
5.	Lexicality	.736	-.263	.729	.858	—	**.303**	**.280**	**.309**	**.329**
6.	Exception (Jared)	.747	.441	.386	.712	.663	—	**.356**	**.268**	**.206**
7.	Consistency	1.000	.691	.512	1.000	1.000	1.000	—	**.332**	.125
8.	Exception (R&C)	.575	.138	-.005	.568	.470	.689	1.000	—	**.613**
9.	Pos. of Irregularity	.650	.099	.228	.421	.526	.557	.809	1.000	—

Table 6: Correlations between raw estimated effects, above diagonal; the frequency effect has been reverse-coded (sign-flipped) so that a greater benefit from higher frequency is indicated by a larger number. Correlations significant at $\alpha = .05$ ($|r| \geq .197$) are indicated in bold. The same values adjusted for attenuation using test-retest reliability are presented below the diagonal; these are not appropriate for significance testing against zero. R&C = Rastle and Coltheart.

times. Whilst in modelling applications, this parameter can simply be included for every participant, for a more qualitative understanding of correlations, removing such an effect is desirable. To do this when we assume (as in the models we use here) that the noise is independent from the central processes of interest requires a method that adjusts for central process variance with noise removed, rather than all the variance (as in z-scoring); our method for such adjustment is detailed in Appendix D.

Once these adjustments were made, we estimated reliabilities for the adjusted effect magnitudes using the correlations between sessions, in the same manner as for the unadjusted effect magnitudes (§2.2.4). The reliabilities were: frequency .114; neighbourhood size .197; word length .171; nonword length .408; lexicality .570; exception (Jared) .399; exception (Rastle & Coltheart) .238; consistency .000[7]; and position of irregularity .395.

Correlations among the effects adjusted for general speed using these variance components, and the disattenuated versions of the correlations are presented in Table 7. The apparently structural correlation between position of

[7] In fact, all three between-session correlations were negative, but a negative reliability is meaningless.

		1.	2.	3.	4.	5.	6.	7.	8.	9.
		-Frq	Nei	WLen	NWLen	Lex	ExcJ	Cons	ExcRC	PoI
1.	(neg.) Frequency	—	**.313**	.072	**.211**	.082	-.015	.123	.029	.093
2.	Neighbourhood Size	1.000	—	-.031	.024	**-.236**	-.195	.069	.115	.090
3.	Word Length	.515	-.169	—	.140	**.262**	-.072	.102	-.106	.055
4.	Nonword Length	.978	.085	.529	—	.174	.153	-.004	-.028	-.106
5.	Lexicality	.322	-.704	.839	.360	—	**.226**	.139	-.170	.054
6.	Exception (Jared)	-.070	-.696	-.276	.380	.474	—	.179	-.182	-.043
7.	Consistency	—	—	—	—	—	—	—	.030	-.074
8.	Exception (R&C)	.176	.531	-.524	-.089	-.461	-.590	—	—	**.505**
9.	Pos. of Irregularity	.438	.323	.213	-.265	.115	-.109	—	1.000	—

Table 7: Correlations between general-speed-adjusted estimated effects, above diagonal; the frequency effect has been reverse-coded (sign-flipped) so that a greater benefit from higher frequency is indicated by a larger number. Correlations significant at $\alpha = .05$ ($|r| \geq .197$) are indicated in bold. The same values adjusted for attenuation using test-retest reliability are presented below the diagonal; these are not appropriate for significance testing against zero; consistency had no test-retest reliability. R&C = Rastle and Coltheart.

irregularity and the related exception measure remained, along with five other significant correlations. Frequency was positively related with neighbourhood size and nonword length. As well as the positive correlation with frequency, neighbourhood size correlated negatively with lexicality. In addition to this negative correlation with neighbourhood size, lexicality correlated positively with word length and exception effect by the Jared measure. These relationships are illustrated in Figure 2.

Five other correlations approached significance: negative correlations of the Jared exception measure with neighbourhood size, consistency, and the Rastle & Coltheart measure of the exception effect; and lexicality's positive correlation with the nonword length effect and lexicality's negative correlation with the Rastle & Coltheart exception effect.

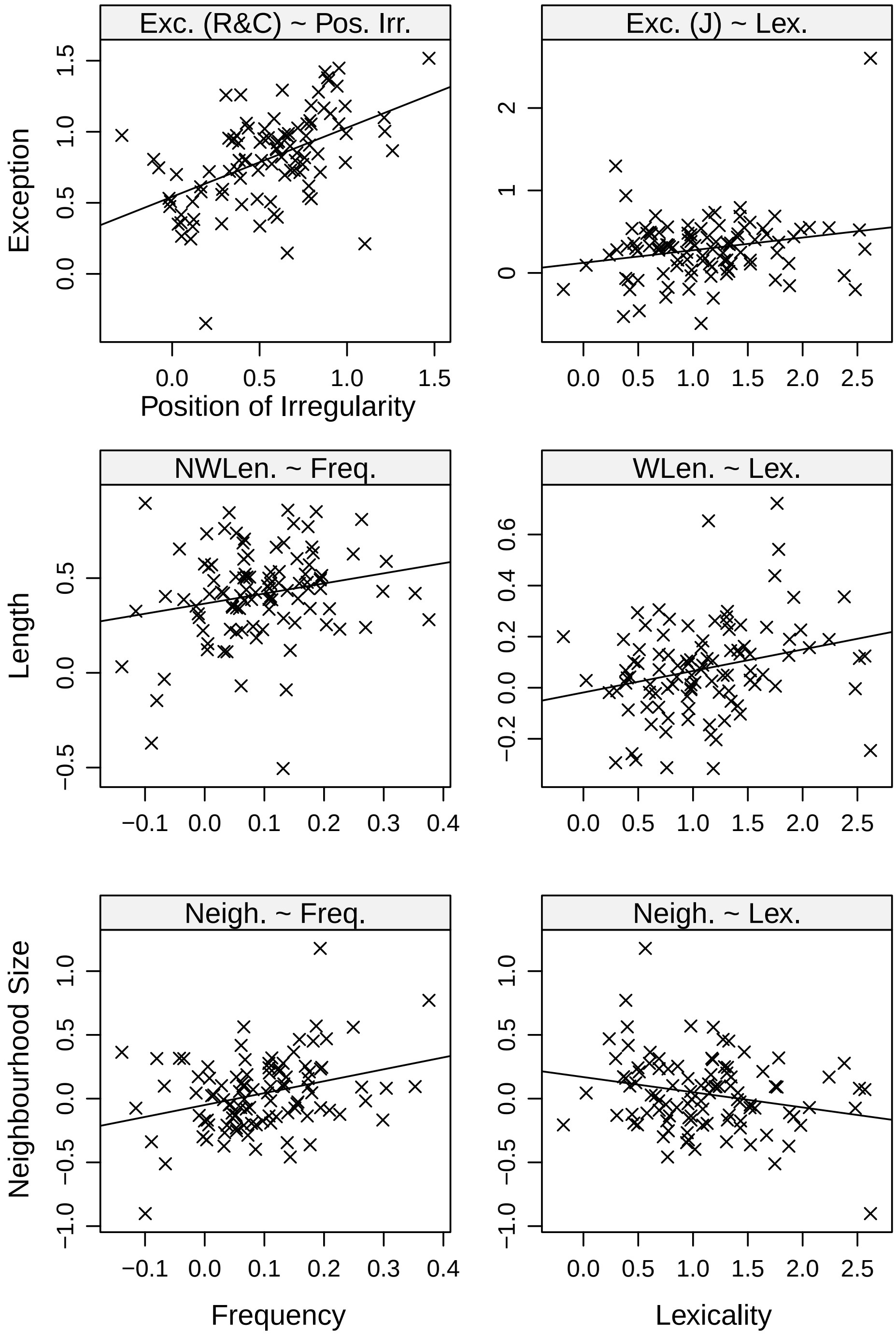
2.3. Discussion

2.3.1. Replication of standard effects

Our data largely replicated the major effects of interest, with the exception of neighbourhood size.

Frequency had its usual effect such that rare words required more process-

Figure 2: Scatterplots illustrating the significant individual differences correlations between general-speed adjusted effects.



ing time. The neighbourhood size effect was not significant. We suspect this is because there is a mixture of readers showing facilitatory and inhibitory effects, and our sample has a greater proportion of readers showing inhibitory neighbourhood size effects than the typical experiment showing the facilitatory effect. This seems likely because our sample was older than that in a typical experiment (with introductory psychology students), and the facilitatory effect is associated with younger participants (Spieler & Balota, 2000). Longer stimuli were read more slowly than shorter stimuli, with the effect being stronger for nonwords than real words. Consistent words were read more quickly than inconsistent words, only if their friends were more frequent than their enemies, as found by Jared et al. (1990). Exception words were read more slowly than regular words. The dependence of this on the position of the irregularity — greater effects for irregularities to the left (beginning) of the word — was replicated, but the evidence for a dependence on consistency was not.

2.3.2. Effects by session

For the most part, participants responded slower in later sessions, but the slowest response categories were immune to this effect. This could be artifactual, or relate to a change in response criterion (cf. Taylor & Lupker, 2001).

2.3.3. Comparison with mega-study data

There was great consistency between our data and the same stimuli in Spieler & Balota's (1997) data. The consistency effect did not reach significance in Spieler and Balota's data, but the effect was in the same direction in both data sets, and such effects are significant in their data in regression analyses over a larger set of items (Adelman & Brown, 2008a). The only other minor inconsistency was one test of the exception effect in our data required a one-tailed correction to reach significance, where such as correction was not needed in Spieler and Balota's data. Overall, this confirmed the typicality of our data.

2.3.4. Reliability and variability

There was clear evidence for reliability of seven of the nine measures as measures of individual traits, marginal evidence for such reliability of neighbourhood size effects, and no such evidence for consistency effects. This type of statistical evidence is ambiguous between very high actual session-to-session variability or very small (or absent) individual differences. Large relative session-to-session variability may partially result from relatively few items contributing to the effect estimate.

That the neighborhood size effect was correlated with other effects in this experiment, and can be predicted from age and external written language variables in this task (Adelman, Sabatos-DeVito & Marquis, in prep.) and lexical decision (e.g., Sears et al., 2008) suggest that individual differences are not absent, but the weak effects are largely outweighed by trial-to-trial noise.

Given the size and significance of the consistency effect, it is surprising that those who show the strongest such effects in one session, are not those showing the strongest effects in the next. We suggest that this is due to large priming effects that differ with different randomizations of the stimulus ordering that outweigh the individual differences. When word bodies are repeated within an experiment, the pronunciation at the first presentation can facilitate a consistent future presentation or inhibit an inconsistent one (Seidenberg et al., 1984). As such, the consistency effect may itself — in whole or in part — result from the probability that the most recent preceding occurrences of the word body in question (which need not be within the experiment itself) having been in a rime friend that primes the relevant pronunciation. Where such variation is intrinsic to the causation of the effect, lower estimated reliability should be expected than if an effect is caused by a fixed property of the underlying mechanism, because susceptibility to the effect interacts with the helpfulness of the stimulus ordering.

Moreover, to the extent that orthographic neighbours are more (or only) facilitatory when they are also phonological neighbours (Adelman & Brown, 2007; Peereman & Content, 1997), a similar consistency-based mechanism might be proposed for neighbourhood size effects. If the effect of frequency is linked to contexts rather than occurrences per se (Adelman et al., 2006), then a similar argument may be made for these contextual diversity effects: More contextually diverse words are more likely to have an appropriate context instantiated by a recently presented word than are contextually restricted words.

Overall, whilst it is clear that there are reliable individual differences, it also appears that within-individual variation may be very important for understanding effects in which priming might have a causal role.

2.3.5. Correlations among effects

Turning to the central aim of the experiment, raw correlations were positive among nearly all pairs of effect measures, indicative of overall general speed differences. The interpretation of these general speed differences may be difficult if there are genuinely two processes — lexical and nonlexical — of imperfectly correlated varying speeds. Correlations on measures adjusted for general speed (Table 7) revealed six significant relationships, and five that approached significance, which will be discussed in turn.

We consider the effects largely in terms of principles of recent versions of dual-route theory (Coltheart, 2012, or more concretely, Coltheart et al., 2001; Perry et al., 2007, for examples); we are not aware of previous examples of their use to explain these types of correlations. These principles are that: lexical access speed depends on frequency; nonlexical processing has a left-to-right component; for words, lexical and nonlexical processes can co-operate and compete to produce pronunciations, and co-operation produces faster responses than competition; words are read quicker than nonwords because the lexical route produces pronunciations rapidly.

Significant correlations.

Exception (Rastle & Coltheart) and Position of Irregularity.

As described above, the positive correlation between the exception effect from the Rastle & Coltheart stimuli and the position of irregularity effect (from the same stimuli) is explained by the position one exception effect for the same items contributing most of the variance to both these measures.

Frequency and Neighbourhood Size. Both frequency and neighbourhood size effects have been taken to be an indication of the quality of lexical representations (e.g., Sears et al., 2008). Moreover, neighbourhood effects can be produced by the lexical route of the DRC (Adelman & Brown, 2008a; Reynolds & Besner, 2002). In a lexical quality interpretation, participants who show these effects strongly have representations that are weak and therefore slow to be activated, and poorly differentiated from other representations (which in naming are coincidentally helpful, cf. Adelman & Brown, 2007; Peereman & Content, 1997). Their positive relationship can readily be explained by such a concept, possibly linked to letter-word connections in models, and hence lexical route strength or speed (linking into general speed). Any nonlexical contribution to neighbourhood size effects (Perry et al., 2007) would not directly give a positive link with the lexical frequency effect, though an indirect influence (i.e., cross-talk) could involve both routes.

Lexicality and Neighbourhood Size. The lexicality effect is an indication of the extent to which words are read more quickly than nonwords, which in a dual-route framework largely reflects the extent to which the lexical route operates more quickly than the nonlexical route (e.g., Coltheart et al., 1993). That is, strong lexicality effects reflect a fast lexical route, whereas strong neighbourhood effects reflect, according to the argument above, a slow (poorly specified) lexical route; this would explain the negative relationship.

Lexicality and Word Length. Length effects are often thought of as due to a left-to-right process in the nonlexical route (Coltheart & Rastle, 1994). The slower such a process operates, the greater difference will occur between shorter and longer words, at least insofar as the nonlexical route is a limiting factor. In the present account, a larger lexicality effect also should indicate a relatively inefficient nonlexical route, so the positive relationship might follow, so long as it is not so slow as to have no effect. Alternatively, word length effects may have both a lexical and nonlexical locus.

Lexicality and Exception (Jared). By the same token, a large lexicality effect suggests that the lexical route is operating quickly and the nonlexical route is operating slowly, so little variation can be contributed by the lexical route, as everything is fast. This also means effects in the nonlexical route would be large because its processing is protracted, and large when adjusted for general speed, because the nonlexical route is contributing most of the variance. The exception effect is one of these effects produced from nonlexical contributions, so should be large in this scenario where the lexicality effect is large, thus producing the positive correlation.

Frequency and Nonword Length. Finally, the positive relationship between frequency and nonword length is harder to explain, because they appear to reflect relatively pure lexical and nonlexical processes respectively, and so should be separately linked to these separate speeds. If a positive link between these two speeds were the source, though, it should have been extracted in the general speed correction. That it remains after this correction required explanation. One possible explanation is that the strong frequency effect occurs when there is more generalised activation (and more competition) in the lexical route (due to the poor differentiation of representations); as such, the lexical route interferes with, and slows, output from the nonlexical route, and disproportionately more so for (more slowly read) longer nonwords. An alternative explanation, that is less consistent with the traditional dual route conception, is that when the lexical route is operating rapidly, it is sometimes able to contribute a lexical analogy pronunciation before the slower, length-sensitive nonlexical route can generate a response, reducing the influence of the nonlexical route.

Marginal correlations.

Lexicality and Nonword Length. Likewise, a large lexicality effect indicates an efficient lexical route, so that once general speed is accounted for, most variation is due to the nonlexical route, where the nonword length effect is generated.

Neighbourhood Size and Exception (Jared). When the neighbourhood size effect is large, this implies that the lexical route is not only slow, but also important in determining RTs for words; as such, factors that reflect nonlexical influence, including costs from exceptions, should be relatively less important (once general speed is adjusted for), generating the observed negative correlation.

Exception (Jared) and Exception (Rastle & Coltheart). These

nominally similar effects correlate negatively, perhaps because the Rastle and Coltheart measure is dominated by the cost for irregularities in the initial position — mostly involving consonants — and the Jared measure is based on later irregularities — mostly involving vowels. Possibly these reflect different influences of nonlexical pronunciation. Effects of early irregularities might reflect inhibitory processes, possibly at the lexical level, whilst the effect of later irregularities might reflect a lack of facilitation for the correct pronunciation, rather than inhibition per se.

Exception (Jared) and Consistency. These inconsistent and irregular stimuli both mostly involve differences in the pronunciation of the vowel, so might be subsumed under some similar form of nonlexical process.

Exception (Rastle & Coltheart) and Lexicality. According to the description above, inhibition to lexical access is involved in the exception effect with the Rastle & Coltheart stimuli. For such inhibition to be effective, the lexical process cannot outpace the nonlexical process too severely, which would be associated with a small lexicality effect; this would produce a negative correlation.

2.3.6. Verbal analyses vs. simulated models

However, it is unclear whether this pattern of correlations is necessarily to be expected in a complete dual-route model. It is one thing to describe these causes verbally, but it is another to show that they can all occur in the same single system. Indeed, arguments could have been made for some of these correlations in the opposite direction, such as a negative correlation between frequency and nonword length. Part of this is due to nonlinearity in the effect of speed in a system with routes that can co-operate in parallel: a fast system can have little effect on RTs because there is little range in its contribution, and a slow system can have little effect on RTs because it is outpaced by a faster system. Determining what can in fact occur in a dual-route system therefore requires simulations of a dual-route model.

3. MODELLING: DUAL-ROUTE CASCADED MODEL

The most well-known implemented dual-route model is the dual-route cascaded model (DRC: Coltheart et al., 1993, 2001). Although published versions of this model lack semantics and the ability to read polysyllabic words, it is complete in the sense that both the lexical and nonlexical route are implemented in some detail, and it is able to produce predictions for any monosyllabic stimulus, and can produce all the effects considered here (see regressions performed by Adelman & Brown, 2008a). This model stands as a conjunctive hypothesis about the processes by which people read aloud; that is, its authors consider it a theory in the sense of a set of falsifiable statements about the details of cognition, rather than only one possible implementation of some broader principles. It combines a lexical route that is based on an extension of the interactive-activation and competition model (McClelland & Rumelhart, 1981) with a set of spelling-sound rules that are applied in a temporally left-to-right fashion.

In our simulations, we sought to examine whether the model's description of the structures and processes of reading aloud is sufficiently accurate to allow simulation of these data. Does adjustment of numerical parameters allow the model to capture the variation and correlations in our data? Or does the model make assumptions about the mechanisms and parameters underlying the item effects that are shown to be incorrect by our data?

3.1. Method

3.1.1. Models

Version of the Model. Coltheart et al. (2001) described the DRC with explanations for the motivation for its structure; we (Adelman & Brown, 2008a) redescribed it with some details that we used in our re-implementation to produce the behaviour of the program that Coltheart et al. made available but which were not mentioned in their paper. For present purposes, one minor change was made to the model in the way that the grapheme-phoneme conversion (GPC) route is initiated, and the phonological lexicon frequencies were changed to be the same as used by Perry et al. (2007) (which were provided to them by Coltheart). In terms of the GPC route, in the earlier version of the model, the cycle on which the GPC first converted the first letter was directly specified by a GPC delay parameter (α: on the tth model cycle, the left-most $\lfloor 1 + (t - \alpha)/\beta \rfloor$ letters are considered). The modification was to set a threshold on letter activation to initiate GPC processing: That is the GPC delay was not directly a parameter, instead the start of GPC processing (α) was set to be the first cycle on which a letter in the left-most channel exceeded a threshold parameter. This change has been adopted in unpublished modifications to the DRC by Coltheart and colleagues (DRC 1.2.1, used to model findings by, e.g., Mousikou, Coltheart, Finkbeiner & Saunders, 2010), in response to criticisms such as those of Blais & Besner (2007) relating to the response of the model to degraded stimuli. The rate at which letters are identified in the model does not vary substantially across words (only across stimulus qualities) so the effect of the old parameter is mimicked by the new parameter.

Parameter Settings. Two-hundred-and-fifty thousand potential parameter sets were each produced by independently randomly selecting each parameter from a set of values based upon those found to be optimal for average data by Adelman & Brown (2008a) and, for parameters also in the CDP+, the value used by Perry et al. (2007). The values used are presented in Appendix E.

3.1.2. Stimuli and Procedure

All but two stimuli from the experiment were used for simulation, the exceptions being DIRE, which is bisyllabic in most accents including the model's, and MOULD, as only MOLD is in the model's vocabulary. All 250,000 parameter sets were tested by sequentially presenting the stimuli to the model to obtain responses and RTs in model cycles; a trial was ceased, counted as an error and excluded from analysis if more than 300 cycles were required. Responses that did not match the model's stored pronunciation for words were also counted as errors. If and when 60 stimuli had timed out or been erroneously pronounced by the model with a given parameter set, that simulation was ceased, and the parameter set rejected; 2,674 sets were retained.

3.1.3. Selection to Correspond to Participants

For each participant, a multiple regression was performed for each parameter set that had been simulated (and retained), with the response being the participant's mean RT for each word[8], and the predictors being the model's RT in cycles and the initial phoneme of the model's response. For each participant, the parameter set whose regression produced the lowest mean-squared-error (highest R^2) was retained, and the regression's predictions were treated as the predictions of the DRC (i.e., the DRC was permitted to treat the effect of first phoneme as explicable but outside the model's scope). The mean R^2 obtained was 36.10%.

3.2. Results

3.2.1. Standard Effects

We first examined whether the DRC had produced the effects observed in the data; the predicted means for each condition are in Table 4. The corresponding ANOVAs follow, and are also summarized in Table 5.

[8] One might imagine performing a similar calculation on condition means or effects, but this discards information, and results in fewer points to use for fitting than parameters to fit them.

Frequency. Predicted RT decreased with increasing frequency, $F_1(4, 396) = 84.98$, $F_2(4, 80) = 113.4$.

Neighbourhood size. The model predicted an inhibitory effect of neighbourhood size, $F_1(1, 99) = 6.76$, $p = .011$, $F_2(1, 41) = 6.19$, $p = .017$. However, it would be unrealistic to expect a predicted effect of 1 ms to reach significance in the data.

Length and lexicality. The model predicted longer RTs for longer words, $F_1(2, 198) = 52.35$, $F_2(2, 60) = 132.6$, and the same pattern for nonwords, $F_1(2, 198) = 106.2$, $F_2(2, 60) = 152.2$. The difference between words and nonwords, such that nonwords were slower, was significant $F_1(1, 495) = 1056$, $F_2(1, 150) = 4995$, as was the interaction such that the length effect for nonwords was greater, $F_1(2, 495) = 13.61$, $F_2(2, 150) = 53.88$. The estimated means after partialling out neighbourhood size were: 544, 549, and 552 ms for words and 594, 605, and 621 ms for nonwords.

Regularity and Consistency.

Regularity. Jared's (2002) exception stimuli with high frequency enemies were predicted to have slower responses than their matched controls, $F_1(1, 99) = 98.20$, $F_2(1, 17) = 13.00$, $p = .002$, as were those with high frequency friends, $F_1(1, 99) = 67.75$, $F_2(1, 19) = 19.20$.

Consistency. The regular-inconsistent stimuli with high frequency enemies were predicted to have slower responses than their matched controls, though
this was significant only by-subjects, $F_1(1, 99) = 46.15$, $F_2(1, 19) = 0.58$, $p =$
1.1 .456. The predicted 1 ms effect was, however, tiny compared to that in the data.
The regular-inconsistent stimuli with high-frequency friends showed the opposite pattern, which was again only significant by-subjects, $F_1(1, 99) = 17.80$, $F_2(1, 19) = 0.04$, as in the data.

Position of irregularity. Exception words whose irregularity was in the first position were predicted to be read more slowly than their regular controls, $F_1(1, 99) = 311.0$, $F_2(1, 19) = 12.65$, $p = .002$, as were the second position exceptions, $F_1(1,99) = 51.89$, $F_2(1,38) = 22.44$, and the third position exceptions, $F_1(1,99) = 23.85$, $F_2(1,28) = 7.01$, $p = .013$. The regularity by position interaction was significant, $F_1(2,495) = 132.2$, $F_2(2,85) = 12.23$.

3.2.2. Recovery of individual differences in effects

Correlations between the participants' effects and the corresponding effects in the simulations were: frequency .447; neighbourhood size -.039; word length .476; nonword length .784; lexicality .885; exception (Jared) .498; exception (R&C) .686; consistency .239; position of irregularity .671. Correlations using the corrections for general speed (which for the model is equivalent to z-scoring as there is no trial-to-trial variance in the model predictions) were: frequency .338; neighbourhood size .034; word length .319; nonword length .527; lexicality .647; exception (Jared) .394; exception (R&C) .556; consistency .084; **1.11** position of irregularity .572. Again, the criterion for significance is $|r| \geq .197$ (or $|r| \geq .164$ one-tailed).

3.2.3. Effect correlations

We examined the correlations that the predicted RTs would produce between the effects, presented in Table 8. Most of the significant correlations were positive, again consistent with the influence of general speed. The exception was neighbourhood size's correlation with frequency, which is negative, contrary to data. Overall, of 24 correlations that reached significance in the data (Table 6), 19 were predicted to be significant in the correct direction, 2 not significant but in the correct direction, 2 not significant and in the incorrect direction, and 1 significant in the incorrect direction.

The correction for general speed produced the correlations in Table 9, in which the notable negative correlation between neighbourhood size and frequency persists. Of the 6 correlations significant in the data after the adjustment (Table 7), 3 were predicted significant in the correct direction (including the artifactual correlation involving the two measures from the Rastle & Coltheart stimuli), 2 not significant but in the correct direction, and 1 significant in

		1.	2.	3.	4.	5.	6.	7.	8.	9.
		-Frq	Nei	WLen	NWLen	Lex	ExcJ	Cons	ExcRC	PoI
1.	(neg.) Frequency	—	**-.287**	**.374**	**.568**	**.592**	.191	**.757**	**.341**	**.234**
2.	Neighbourhood Size		—	-.170	-.072	-.186	-.077	-.133	-.112	-.144
3.	Word Length			—	**.644**	**.481**	.167	.015	**.478**	**.241**
4.	Nonword Length				—	**.666**	-.098	**.385**	**.250**	**.270**
5.	Lexicality					—	**.261**	**.452**	**.435**	**.206**
6.	Exception (Jared)						—	.110	**.631**	-.052
7.	Consistency							—	**.211**	**.198**
8.	Exception (R&C)								—	**.605**
9.	Pos. of Irregularity									—

Table 8: Correlations between raw DRC predicted effects; the frequency effect has been reverse-coded (sign-flipped) so that a greater benefit from higher frequency is indicated by a larger number. Correlations significant at $\alpha = .05$ ($|r| > .197$) are indicated in bold. R&C = Rastle and Coltheart.

		1.	2.	3.	4.	5.	6.	7.	8.	9.
		-Frq	Nei	WLen	NWLen	Lex	ExcJ	Cons	ExcRC	PoI
1.	(neg.) Frequency	—	**-.244**	**.232**	**.339**	**.459**	.148	**.774**	**.216**	.064
2.	Neighbourhood Size		—	-.086	.056	-.092	-.058	-.134	-.038	-.067
3.	Word Length			—	**.577**	**.250**	.087	-.083	**.323**	.057
4.	Nonword Length				—	**.437**	**-.262**	**.198**	.013	.007
5.	Lexicality					—	.110	**.412**	.107	-.161
6.	Exception (Jared)						—	.080	**.565**	**-.211**
7.	Consistency							—	.172	**.553**
8.	Exception (R&C)								—	.140
9.	Pos. of Irregularity									—

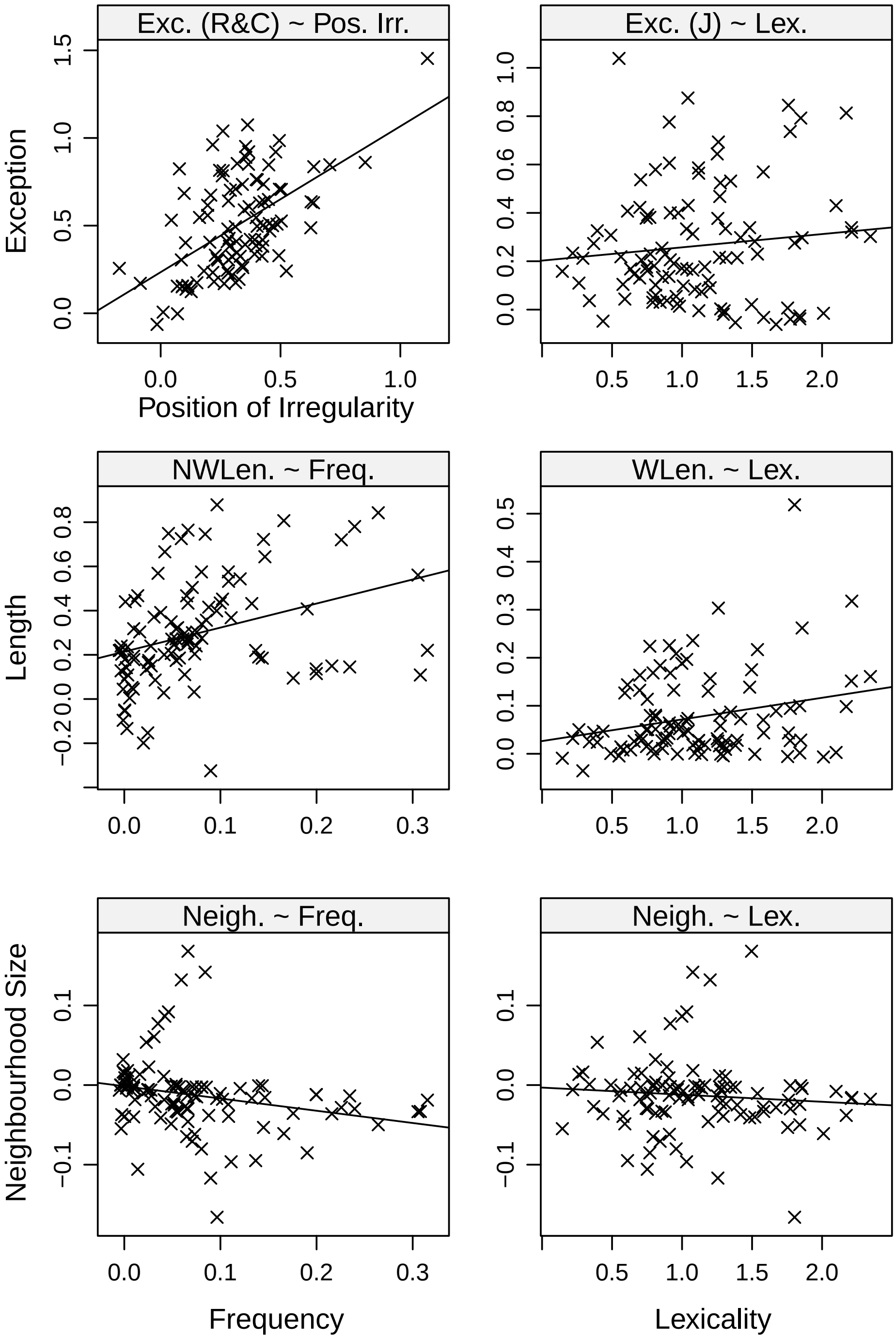
Table 9: Correlations between general-speed-adjusted DRC predicted effects; the frequency effect has been reverse-coded (sign-flipped) so that a greater benefit from higher frequency is indicated by a larger number. Correlations significant at $\alpha = .05$ ($|r| > .197$) are indicated in bold. R&C = Rastle and Coltheart.

the incorrect direction; again, this was the correlation between frequency and neighbourhood size effects. These relationships are illustrated in Figure 3.

3.3. Discussion

Parameter modification allowed the DRC to adjust to much of the individual variation in effects shown in the data: As well as simulating the average effects of frequency, word and nonword length, lexicality, exception, and position of irregularity, in the data, the simulated individual differences correlated

Figure 3: Scatterplots illustrating the pertinent individual differences correlations between general-speed adjusted effects simulated with the DRC.



with observed individual differences in these effects, both before and after the adjustment for general speed.

However, closer examination revealed some problems; in particular, the model did not succeed in capturing phenomena relating to the neighbourhood size effect. First, although the model can produce a small facilitatory effect of neighbourhood size under some parameter settings (Adelman & Brown, 2008a; Coltheart et al., 2001; Reynolds & Besner, 2002, and the sets chosen for 20 of the participants here), the mean effect for the fitted parameters was inhibitory, suggesting that parameter sets that produce facilitatory neighbourhood size effects fit other effects less well than those producing inhibitory neighbourhood size effects. Second, no correlation emerged between the observed and predicted neighbourhood size effects. Third, the model predicted that participants showing strong frequency effects would show more inhibitory neighbourhood size effects, when in the data, participants showing strong frequency effects showed more facilitatory neighbourhood size effects.

This occurs because of a trade-off between frequency and neighbourhood effects that was noted by Forster (1976). Consider two neighbours of differing frequency, such as BRIGHT and BLIGHT. The neighbourhood effect requires that a word's neighbours are activated when that word is presented. When the lower frequency item (BLIGHT) is presented, not only is its higher frequency neighbour (BRIGHT) activated, it has greater activation due to its frequency. If the combination of these two effects is too great, the lower frequency item will be mistaken as its higher frequency neighbour. As such, parameter sets that produce both effects strongly are not acceptable due to their unreasonably high error rate, and thus were rejected as candidates to represent participants in the modeling. Thus, the parameter sets that were retained in the modeling to represent individual participants show *either* a strong frequency effect *or* a facilitatory neighbourhood size effect (and a weaker frequency effect), thus producing a negative correlation.

One possible explanation of the pattern in the human data is given by the idea of lexical precision and the lexical tuning hypothesis (e.g. Andrews, 2012;

Andrews & Hersch, 2010; Castles, Davis, Cavalot & Forster, 2007): Experience with a word increases the quality of the representation of that word so that it is easier to distinguish from its neighbours. That is, a higher frequency word is not only more easily recognized, it is more easily rejected when it is not the stimulus. As such, the higher frequency neighbor (BRIGHT) of a low frequency stimulus (BLIGHT) usually ceases to be a candidate for identification sufficiently early that misidentification does not occur.

This is consistent with Andrews and Hersch's (2010) criticism of interactive-activation based models: These models simulate an individual with perfect orthographic knowledge — perfect spelling — rather than allowing for imprecise representation, which appears to be a key source of variability between even individuals who can competently read. Modulation of the parameters of the model could not mimic this, probably because this kind of variation in lexical quality is selective to low-frequency words, whereas the parameters are not. Whilst such individual differences in lexical quality of known words are correlated with vocabulary, the importance of spelling over vocabulary as a predictor of the priming effects in these studies implicates lexical quality as the actual cause[9].

Furthermore, the model underestimated the overall magnitude of regularity and consistency effects. This may be because nonlexical influences operate directly only on pronunciation, and the effect on lexical activation is weak because it is indirect. That is, exceptional spelling-sound correspondences primarily affect pronunciation, rather than lexical access itself. Alternatively, the problem could be to do with the assembly of nonlexical phonology itself, which the CDP+ model was designed to improve relative to the DRC.

[9] Moreover, allowing vocabulary to vary with an additional parameter (minimum frequency of word known) did not improve the model's ability to capture the frequency–neighbourhood-size correlation.

4. MODELLING: CONNECTIONIST DUAL PROCESS MODEL

The first complete version of the connectionist dual process model, the CDP+ (Perry et al., 2007), can be seen as a modification of the DRC to incorporate graded spelling-sound correspondences that are learned by the delta rule, and a more structured phonological representation; its authors argue it offers a more complete explanation of the item effects attested in the literature.

4.1. *Method*

Modelling with CDP+ was similar to that performed for the DRC, except the model was, of course, that described by Perry et al. (2007)[10], and the parameters based upon those used in that paper, detailed in Appendix E; where a parameter was common to the DRC and CDP+, the same parameter range was used as for the DRC. 3,505 parameter sets were retained as passing the fewer-than-60-errors criterion. The mean R^2 obtained was 36.97%.

4.2. *Results*

4.2.1. *Standard effects*

We checked the CDP+'s predictions for the standard effects with these parameters; the predicted means for each condition are in Table 4. The corresponding ANOVAs follow, and are also summarized in Table 5.

Frequency. A significant effect of shorter response times for more frequent words was found, $F_1(4, 396) = 47.77$, $F_2(4, 80) = 18.36$.

Neighbourhood size. The words with lower neighbourhood size were predicted to have more rapid responses than the matched words with high neighbourhood size. This miniscule (less than 0.1 ms) difference was only significant by subjects, $F_1(1, 99) = 4.12$, $p = .045$, $F_2(1, 39) = 0.00$. This is suggestive of the model being very sensitive to a slight mismatch of the neighbourhood size

[10] There has been further development of this model, primarily in terms of extension to multisyllabic vocabulary, but this is the version that was available at the time we began the time-consuming simulations, and the one most comparable to the published version of the DRC.

stimuli (that is not significant across pairs), and an effect of this magnitude obviously cannot be treated as a substantive prediction of the model.

Length and lexicality. Predicted RTs increased with length for words, $F_1(2,198) = 21.26$, $F_2(2,60) = 30.71$, and nonwords, $F_1(2,198) = 95.15$, $F_2(2,60) = 21.75$. Nonwords were read slower, $F_1(1,495) = 1014$, $F_2(1,150) = 1091$, and the length effect was stronger for nonwords, $F_1(2,495) = 13.61$, $F_2(2,150) = 13.41$. The estimated means adjusting for neighbourhood size were 552, 553, and 554 ms for words, and 596, 603, and 614 ms for nonwords.

Regularity and consistency.

Regularity. Jared's (2002) exception stimuli with high-frequency enemies had greater predicted RTs than their controls, which was significant by-subjects, $F_1(1,99) = 84.12$, but not by-items, $F_2(1,17) = 0.01$, and a similar pattern held for the exceptions with low-frequency enemies, $F_1(1,99) = 67.83$, $F_2(1,19) = 0.10$.

Consistency. The regular-inconsistent stimuli with high-frequency enemies were given longer predicted RTs than their controls by a miniscule amount (different in the third decimal place), again only significant by subjects, $F_1(1,99) = 6.063$, $F_2(1,19) = 0.04$. This predicted effect is not comparable in size to that in the data. For those regular-inconsistent stimuli with low-frequency enemies, the slower predictions than their controls again reached significance only by subjects, $F_1(1,99) = 7.436$, $p = .008$, $F_2(1,19) = 0.73$; this is not the direction in the data.

Position of irregularity. Exception words whose irregularity was in first position were predicted to have RTs longer than their matched controls, $F_1(1,99) = 193.0$, $F_2(1,19) = 5.34$, $p = .032$, but those in second position were not, $F_1(1,99) = 1.47$, $p = .228$, $F_2(1,38) = 0.12$. The prediction that third position irregulars would be read more quickly than their controls was significant only by subjects, $F_1(1,99) = 30.34$, $F_2(1,28) = 3.66$, $p = .066$; this is not the direction in the data. The position by regularity interaction was significant, $F_1(2,495) = 141.1$, $F_2(2,85) = 9.48$.

4.2.2. Recovery of individual differences in effects

Correlations between the participants' effects and the effects in the corresponding CDP+ simulations were: frequency .382; neighbourhood size .020; word length .464; nonword length .612; lexicality .931; exception (Jared) .151; exception (Rastle and Coltheart) .514; consistency .224; position of irregularity .491. With the adjustments for general speed applied, the correlations were: frequency .221; neighbourhood size .113; word length .317; nonword length .558; lexicality .766; exception (Jared) .001; exception (Rastle & Coltheart) .557; consistency .239; position of irregularity .438.

4.2.3. Effect correlations

We examined the correlations that would emerge from the predicted RTs, presented in Table 10. Again, many positive correlations were obtained, consistent with the influence of general speed. Nevertheless, three of the significant correlations were negative, two involving consistency. Of these, one — the correlation between consistency and exception (Jared stimuli) effect — was surprisingly contrary to a significant positive correlation in the data. Overall, of the 24 correlations significant in the data (Table 6), 7 were predicted significant in the correct direction and 6 not significant in the correct direction. Of the 11 predicted in the wrong direction, 3 were significant.

The correlations using the predictions adjusted for general speed in Table 11 showed negative correlations involving consistency and effects assumed to be nonlexical, and a positive correlation with frequency; the corresponding effects in the data did not reach significance. Of the six general-speed-adjusted correlations significant in the adjusted data (Table 7), two were predicted significant in the correct direction, two significant in the incorrect direction, and two not significant in the correct direction. These relationships in the simulations are illustrated in Figure 4. The correlation between lexicality and word length was

		1.	2.	3.	4.	5.	6.	7.	8.	9.
		-Frq	Nei	WLen	NWLen	Lex	ExcJ	Cons	ExcRC	PoI
1.	(neg.) Frequency	—	.186	**.490**	-.040	**.565**	-.093	**.422**	.103	.089
2.	Neighbourhood Size		—	.128	-.122	.091	**.286**	.004	**.343**	.225
3.	Word Length			—	.095	**.557**	.022	-.014	.002	.099
4.	Nonword Length				—	**.345**	.023	-.071	**-.225**	-.189
5.	Lexicality					—	-.123	**.260**	-.155	-.011
6.	Exception (Jared)						—	**-.298**	**.365**	.159
7.	Consistency							—	**-.214**	-.111
8.	Exception (R&C)								—	**.929**
9.	Pos. of Irregularity									—

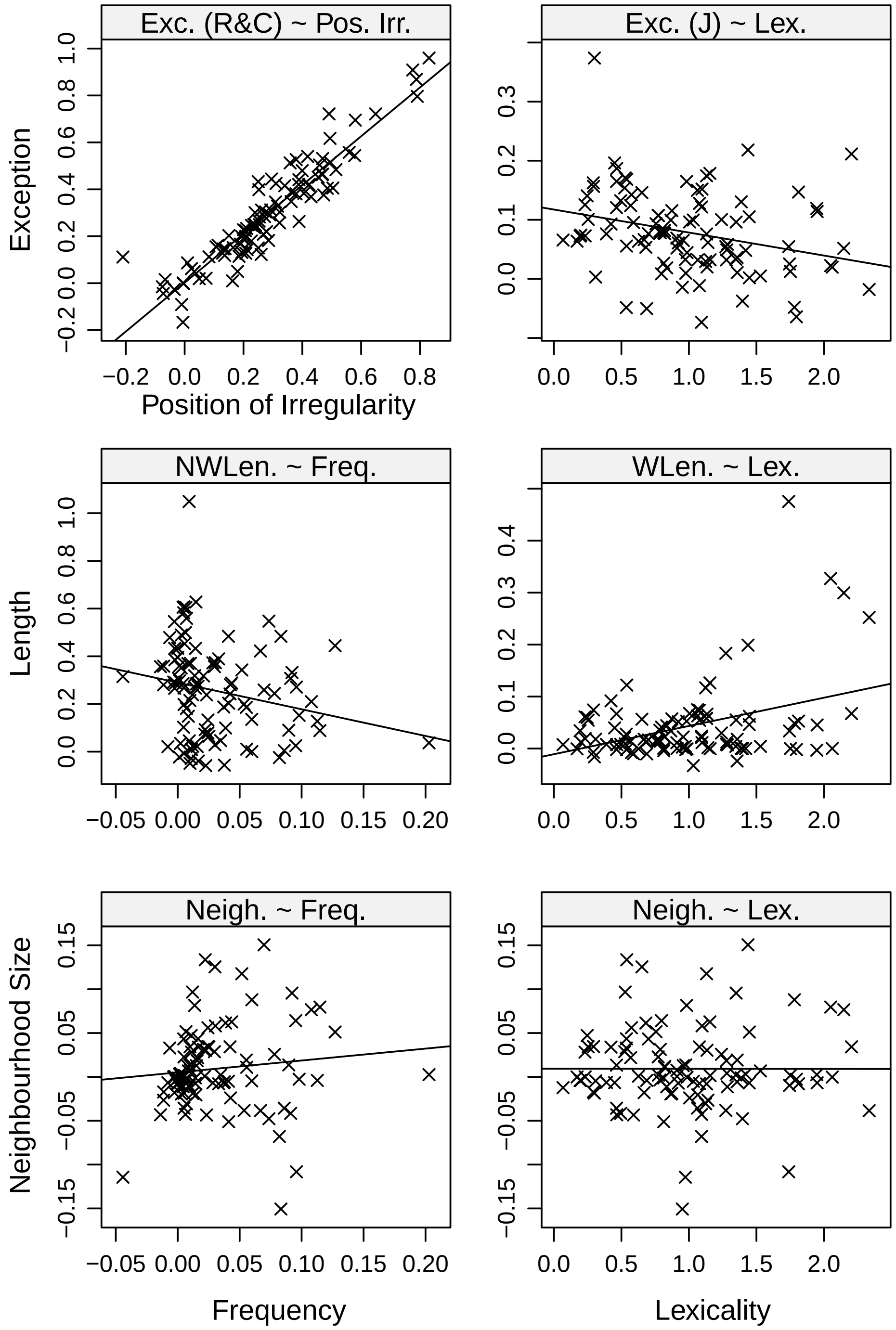
Table 10: Correlations between raw CDP+ predicted effects; the frequency effect has been reverse-coded (sign-flipped) so that a greater benefit from higher frequency is indicated by a larger number. Correlations significant at $\alpha = .05$ ($|r| > .197$) are indicated in bold. R&C = Rastle and Coltheart.

		1.	2.	3.	4.	5.	6.	7.	8.	9.
		-Frq	Nei	WLen	NWLen	Lex	ExcJ	Cons	ExcRC	PoI
1.	(neg.) Frequency	—	.109	**.326**	**-.217**	**.421**	**-.209**	**.450**	-.187	-.018
2.	Neighbourhood Size		—	.114	-.194	-.001	**.261**	-.048	**.315**	**.209**
3.	Word Length			—	-.032	**.381**	-.001	-.096	.030	.081
4.	Nonword Length				—	.134	-.033	-.156	**-.302**	**-.289**
5.	Lexicality					—	**-.289**	**.212**	**-.311**	**-.199**
6.	Exception (Jared)						—	**-.315**	**.312**	.073
7.	Consistency							—	**-.204**	-.110
8.	Exception (R&C)								—	**.933**
9.	Pos. of Irregularity									—

Table 11: Correlations between general-speed-adjusted CDP+ predicted effects; the frequency effect has been reverse-coded (sign-flipped) so that a greater benefit from higher frequency is indicated by a larger number. Correlations significant at $\alpha = .05$ ($|r| > .197$) are indicated in bold. R&C = Rastle and Coltheart.

predicted correctly, as was the artifactual correlation between exception (Rastle & Coltheart) and position of irregularity effects. The incorrectly simulated correlations were that between frequency and nonword length and that between exception (Jared) and lexicality, which both were simulated as negative when they were positive in the human data. Neighbourhood size's correlations with frequency and lexicality effects were in the correct direction (positive and negative, respectively) but not significantly so, and practically zero for lexicality.

Figure 4: Scatterplots illustrating the pertinent individual differences correlations between general-speed adjusted effects simulated with the CDP+.



4.3. Discussion

Parameter modification allowed the CDP+ to capture much of the individual variation in effects. Moreover, the CDP+'s R^2s for the individual participants were better than those for the DRC. However, it did so in a way that was inconsistent with some aspects of the correlational structure of the data, and some effects were not predicted reliably with the chosen parameters.

In terms of effects, the most notable feature was the very weak prediction of consistency effects with the parameters chosen to fit these data. Given that these are a major motivation for the CDP+, and these effects were consistent in Perry et al.'s (2007) simulations, it is surprising that the present simulations did not obtain them.

Given that there do exist parameter values that would produce the effect, the problem must be that parameter values that produce the effect were not selected because they fit the data less well in terms of R^2. In particular, given that there was also no second-position exception effect in the predictions (also present in Perry et al.'s (2007) simulations), it appears parameters were selected that slowed the nonlexical route so that it only processed the first letter during word processing.

The selection of parameters that effectively remove the second-position exception effect and consistency effect could in principle have occurred either because of a present effect problem — that is, the parameters that would produce the missing effects compromise an effect (or the magnitude of an effect) that accounts for more variance. However, it is difficult to see how candidate variables for producing this type of problem — such as frequency and lexicality — could affect the second-position exception effect in this way for the CDP+ but not the DRC, given their similar structure. Even if this is the case, this is not a flaw in the parameters, it is a problem with the model: The model cannot predict the magnitudes of effects shown simultaneously in the data with a single parameter set (per participant).

An absent effect problem — that is, the parameters with a second-position exception effect and a consistency effect introduce a predicted effect that does not occur in the data — might instead explain the problems in these simulations, as this could readily be attributed to a single difference between the CDP+ and the DRC. For instance, the graded (proportional) spelling-sound consistencies to which CDP+ is sensitive may include some to which people are not sensitive; this would also explain why people more often read nonwords like the DRC than like the CDP+ (Pritchard et al., 2012).

Indeed, many of the stimuli have vowels in second position, and most vowel graphemes have several possible pronunciations, so precisely how people process the vowels (e.g., are correspondences used for the whole body/rime of the word, rather than for the vowel alone) may strongly influence the fits.

Two correlations were predicted with the wrong direction in these simulations. One was a predicted negative correlation between frequency and nonword length; this appears to reflect a straightforward trade-off between the lexical process producing the frequency effect and the nonlexical process producing the nonword length effect. Whilst the DRC has the same structure, this trade-off is less marked in the DRC because in that model the nonlexical contribution is the same for all regular words of matched length.

The other problematic correlation was a predicted negative correlation between exception (Jared) and lexicality effects; this probably relates to the role of the nonlexical route: When the nonlexical route is more efficient, this introduces an exception effect (albeit a slight one for these stimuli), and nonwords are read more quickly, reducing the lexicality effect[11].

Given that the introduction of graded spelling-sound correspondences to the nonlexical route did not allow the CDP+ to improve on the DRC's patterns of predictions for our individual differences data, we instead explored a variety of alternative models, including a modification to the DRC's lexical route designed to mimic the alternative suggestion that differences in lexical repre-

[11] Because the parameter sets chosen for the DRC do show second-position exception effects, the size of these effects is more modulated by the size of the contribution of the nonlexical route, which is greater when the lexical route is slow (and lexicality effects are large).

sentation quality (rather than acccessibility) are linked to frequency (see §3.3).

5. MODELING: DUAL-ROUTE CASCADED MODEL WITH FREQUENCY-WEIGHTED CONNECTIONS

We explored a variety of modifications to the DRC that might make it more consistent with the data; we chose DRC as the base model because its base performance had fewer inconsistencies with the data, and because models that did not require new training for each parameter set could be explored more rapidly. These modifications included: using CDP+'s phonological representation; using a vowel-centered orthographic representation; modifying the functional form of frequency effects (cf. Adelman & Brown, 2008b); allowing the parameter controlling the frequency effect to be different for the phonological lexicon than the orthographic lexicon; introducing a direct set of connections from the nonlexical route to the phonological lexicon; allowing participants to vary in vocabulary size by removing low-frequency items; and moving the locus of the frequency effect from bias on lexical units to be in the weights connecting orthography and phonology.

Of the various modifications we examined, the last-named — making frequency have its effect through the connection strength between the orthographic and phonological lexical units (see Besner, Moroz & O'Malley, 2011, for arguments in favour of this mechanism for frequency effects), rather than the bias on those units — made the most progress towards resolving the inconsistencies with data that the DRC showed (without introducing new incorrect predictions), when combined with the changes to orthographic and phonological representations (to be vowel-centered and onset-vowel-coda, respectively; the former was the more important change). We will call this model DRC-FC. We now describe in detail how DRC-FC differs from DRC, and our simulations with the DRC-FC model.

5.1. Method

5.1.1. Model: Differences between the DRC and DRC-FC

Implementation of frequency. The direct influence on the input (the biased input) to the word units in the orthographic and phonological lexicons was removed; this is equivalent to setting the frequency scaling parameter to zero. Instead, the excitation weight on each connection from an orthographic unit to a phonological unit was adjusted by multiplying by a value representing the orthographic frequency from CELEX (the same value as used to produce the biasing input to the orthographic lexical units in DRC). Similarly, the excitation weight on each connection from a phonological unit to an orthographic unit was adjusted by multiplying by a value representing the phonological frequency from CELEX (the same value as used to produce the biasing input to the phonological lexical units in DRC). The multiplers were calculated by dividing the log. orthographic/phonological frequency of the word by the log. orthographic/phonological frequency of the most frequent word (as is used for the bias in the base DRC), and multiplying by a frequency weighting parameter (which was common to both directions of connection).

Orthographic coding scheme. Orthographic representations were left-padded with spaces so that the first vowel (counting Y as a vowel if it was not the initial letter) was in fourth position, and the right-padding with spaces reduced accordingly. The grapheme-phoneme translation system ignored these preceding spaces (but did not add additional following spaces), so that when nonlexical processing was initiated, the first letter, not a space, was immediately processed. This, combined with the following modification, was examined because of the possiblity it might allow rime-based neighborhood or consistency processing along the lexical route.

Phonological coding scheme. Phonological representations were modified to use separate slots for onset, vowel, and coda, as in the CDP+ (rather than the left-alignment of the DRC) but the representation of the blank (missing) phoneme as a distinct unit (rather than a stable near-zero activity) was retained from the DRC. Thus, the phoneme representation for items with fewer than three phonemes in the onset had the blank phoneme unit active in some of the onset slots. This permitted the rule for timing of initiation of pronunciation of the DRC to be used with the minor alteration that the terminating blank needed to be after the vowel. It was therefore necessary for the nonlexical route to activate these blank units to produce a pronunciation. It did so at the same time it activated the vowel, and the strength of the input to these blank units was the same as that to the vowel unit.

5.1.2. Parameter settings, stimuli, procedure and per-participant selection

All other aspects of the methods were the same as those for the DRC, with the exception that a new range of values was needed for the frequency weighting parameter of the new frequency mechanism; the range of all parameter values is given in Appendix E. After the removal of parameters sets producing too many errors 3,548 parameter sets remained. The mean R^2 of the fitted sets was 35.72%.

5.2. Results

5.2.1. Standard effects

We examined DRC-FC's predictions for the standard effects with these parameters; the predicted means for each condition are in Table 4. The corresponding ANOVAs follow, and are also summarized in Table 5.

Frequency. Shorter response times were predicted for more frequent words than for less frequent words, $F_1(4, 396) = 32.39$, $F_2(4, 80) = 6.12$.

Neighbourhood size. The words with higher neighbourhood size were predicted to have more rapid responses than the matched words with low neighbourhood size. This small (0.6 ms) difference was only significant by subjects, $F_1(1, 99) = 4.31$, $p = .041$, $F_2(1, 39) = 1.45$, $p = .235$.

Length and lexicality. The longer a word, the slower its predicted latency, $F_1(2, 198) = 34.04$, $F_2(2, 60) = 36.01$, and likewise for nonwords, $F_1(2, 198) = 50.32$, $F_2(2, 60) = 31.82$. Nonwords were read slower, $F_1(1, 495) = 1106$, $F_2(1, 150) = 3558$, and the length effect was stronger for nonwords, $F_1(2, 495) = 6.51$, $p = .001$, $F_2(2, 150) = 16.83$. The estimated means adjusting for neighbourhood size were 550, 551 and 553 ms for words, and 602, 604 and 612 ms for nonwords.

Regularity and consistency.

Regularity. Jared's (2002) exception stimuli had greater predicted RTs than their controls, whether in the group of items with higher-frequency enemies, $F_1(1, 99) = 75.79$, $F_2(1, 17) = 9.48$, $p = .007$, or low-frequency enemies, $F_1(1, 99) = 52.30$, $F_2(1, 19) = 13.16$, $p = .002$.

Consistency. Differences between the regular-inconsistent stimuli and their regular controls were small and not significant, for both the items with high frequency enemies $F_1(1, 99) = 2.54$, $p = .114$, $F_2(1, 19) = 0.01$, and those with low-frequency enemies, $F_1(1, 99) = 2.10$, $p = .151$, $F_2(1, 19) = 0.21$.

Position of irregularity. First position exceptions were predicted to have RTs longer than their matched controls, $F_1(1, 99) = 222.26$, $F_2(1, 19) = 8.51$, $p = .009$, as were second position exceptions, $F_1(1, 99) = 15.80$, $F_2(1, 38) = 6.64$, $p = .014$. The cost for third position irregularities was small and significant by-subjects, $F_1(1, 99) = 9.24$, $p = .003$, and significant with one-tailed correction by-items, $F_2(1, 28) = 3.00$, $p = .094$. The position by regularity interaction was significant, $F_1(2, 495) = 113.79$, $F_2(2, 85) = 10.30$.

5.2.2. Recovery of individual differences in effects

Correlations between the participants' effects and the effects in the corresponding DRC-FC simulations were: frequency .269; neighbourhood size .121; word length .475; nonword length .597; lexicality .878; exception (Jared) .405; exception (Rastle and Coltheart) .591; consistency .324; position of irregularity .600. With the adjustments for general speed applied, the correlations were:

		1.	2.	3.	4.	5.	6.	7.	8.	9.
		-Frq	Nei	WLen	NWLen	Lex	ExcJ	Cons	ExcRC	PoI
1.	(neg.) Frequency	—	**.235**	**.297**	**.463**	**.412**	-.133	**.569**	.002	.170
2.	Neighbourhood Size		—	**.615**	.174	.071	**-.198**	**-.279**	**.716**	**.375**
3.	Word Length			—	**.421**	**.411**	**.294**	-.145	**.442**	**.248**
4.	Nonword Length				—	**.408**	.065	**.294**	.101	.062
5.	Lexicality					—	**.425**	**.361**	**.275**	.036
6.	Exception (Jared)						—	.162	**.718**	-.003
7.	Consistency							—	.088	-.001
8.	Exception (R&C)								—	**.638**
9.	Pos. of Irregularity									—

Table 12: Correlations between raw DRC-FC predicted effects; the frequency effect has been reverse-coded (sign-flipped) so that a greater benefit from higher frequency is indicated by a larger number. Correlations significant at $\alpha = .05$ ($|r| > .197$) are indicated in bold. R&C = Rastle and Coltheart.

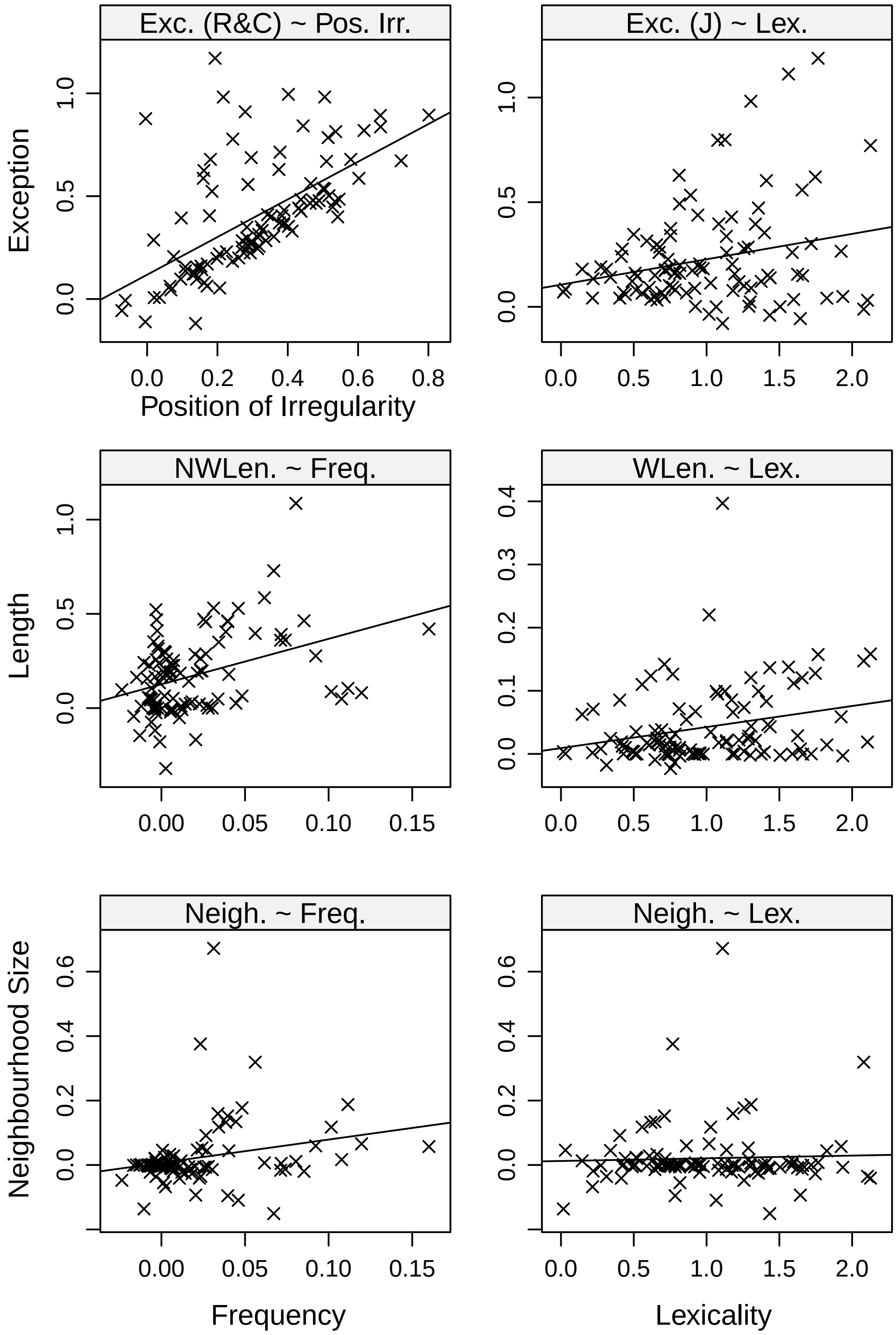
frequency .167; neighbourhood size .156; word length .274; nonword length .485; lexicality .670; exception (Jared) .294; exception (Rastle & Coltheart) .467; consistency .234; position of irregularity .537.

5.2.3. Effect correlations

We examined the correlations of the effects given by the predicted RTs, presented in Table 12. The influence of general speed was again apparent in the abundance of positive correlations. Two correlations were nevertheless significantly negative in the predicted data, where those in the observed data were not significant (postive). Overall, of the 24 correlations significant in the data (Table 6), 14 were predicted significant in the correct direction and 8 not significant in the correct direction. Of the 2 predicted in the wrong direction, neither was significant.

The correlations using the predictions adjusted for general speed in Table 13 showed four negative correlations, none of which was significant in the data. Of the six general-speed-adjusted correlations significant in the adjusted data (Table 7), five were predicted correctly (and significantly), and one not significant in the incorrect direction (see Figure 5). The negative correlation between neighborhood size and lexicality was the effect not produced by DRC-FC.

Figure 5: Scatterplots illustrating the pertinent individual differences correlations between general-speed adjusted effects simulated with the DRC-FC.



		1. -Frq	2. Nei	3. WLen	4. NWLen	5. Lex	6. ExcJ	7. Cons	8. ExcRC	9. PoI
1.	(neg.) Frequency	—	**.249**	**.316**	**.386**	**.269**	**-.297**	**.464**	-.161	.024
2.	Neighbourhood Size		—	**.587**	.137	.041	**-.200**	**-.263**	.153	**.282**
3.	Word Length			—	**.414**	**.264**	.117	-.168	**.300**	.072
4.	Nonword Length				—	**.235**	-.054	**.207**	-.010	-.029
5.	Lexicality					—	**.245**	**.285**	-.054	**-.295**
6.	Exception (Jared)						—	.099	**.624**	-.171
7.	Consistency							—	-.004	-.104
8.	Exception (R&C)								—	**.598**
9.	Pos. of Irregularity									—

Table 13: Correlations between general-speed-adjusted DRC-FC predicted effects; the frequency effect has been reverse-coded (sign-flipped) so that a greater benefit from higher frequency is indicated by a larger number. Correlations significant at $\alpha = .05$ ($|r| > .197$) are indicated in bold. R&C = Rastle and Coltheart.

5.3. Discussion

Like its relatives, DRC-FC accommodated individual differences in effects by changes in parameters. Unlike the other models, in its general-speed-adjusted data, it correctly produced a significant positive correlation between frequency and neighbourhood size effects. In common with the other models, it did not produce a consistency effect comparable to the data, nor a negative correlation between neighbourhood size and lexicality effects.

Introducing vowel-alignment in orthographic and phonological representations did not improve the model's reproduction of the consistency effect. If (as previously discussed) the consistency effect largely results from residual priming of a word's friends and enemies — rather than purely the stimulus activating its own friends and enemies — then this would explain why all the models considered here — where activities are reset for each trial — do not fully capture the effect.

The key improvement with the DRC-FC is correctly catpuring the positive correlation between frequency and neighbourhood size. DRC-FC's ability to capture this pattern — perhaps counterintuitively — appears to rely upon the two effects occuring at different stages within the same route. Where the DRC has both effectively occurring at the lexical units, high frequency tends to increase lexical inhibition, making neighbourhood size effects inhibitory for those showing strong frequency (and overall). In the DRC-FC, a stronger frequency effect can increase the influence of neighbours by passing more activation from orthographic to phonological layers. Alternatively, if these effects are now controlled by different parameters, these parameters are free to correlate in whichever direction is necessary to capture the relationship between the effects.

From a learning perspective, it seems reasonable that the strength of association between orthographic and phonological forms should depend on the number of learning episodes. Whilst this may appear to run counter to other findings that suggest that orthographic units are sensitive to frequency, there could be additional loci of the frequency effect beyond that in DRC-FC, and some tasks that appear to be wholly orthographic may actually drive participants to make decisions that are based on phonological activity (Rastle & Brysbaert, 2006). It still, however, seems implausible that there is no purely visual-orthographic frequency effect, but this may occur on the links between letters and words. One way in which such an account could avoid the BLIGHT-BRIGHT problem is if the frequency effect operates more effectively on inhibitory than facilitatory inputs, so that the frequency benefit is only seen for exact matches.

6. GENERAL DISCUSSION

We now recapitulate the main empirical and modelling findings, consider notable aspects of the empirical findings, consider the implications for models, and conclude with future directions.

6.1. Summary of findings

6.1.1. Effects in the data

The data broadly showed the typical patterns that we sought: Frequent words were read more quickly than rare ones (Forster & Chambers, 1973); the slowing by length Frederiksen & Kroll (1976)was greater for nonwords than words (e.g., Weekes, 1997); words violating spelling-sound rules were read slower than those that did not (e.g., Seidenberg et al., 1984), especially if those violations were early in the word (Rastle & Coltheart, 1999); (some) words whose pronunciations were inconsistent with those that shared an orthographic body (vowel onwards) were read slower than those with no such inconsistency (e.g., Jared, 2002). The evidence was not strong for a facilitatory neighbourhood size effect (see Andrews, 1997), because that effect was moderated by age, with our older participants showing less facilitation, and possibly inhibition, consistent with past research (Spieler & Balota, 2000).

6.1.2. Correlations in the data

We sought correlations among key item effects in word naming, and found many. Most of these could be attributed to general speed, but others could not: positive relationships with frequency effects for neighbourhood size and nonword length effects; positive relationships with lexicality for word length and exception effects; and negative relationships with neighbourhood size for lexicality and exception effects.

6.2. Modeling the data

Whilst we were able to offer an interpretation in terms of dual route theory, we were concerned that most correlations could have been predicted either way. This is a general problem with attributing individual differences to routes without a detailed implementation of their properties and how their contributions are combined. This is because a route's effect may be weak because it is fast or because it is too slow. If it is too fast, its fastest and slowest responses differ little in speed. If it is too slow, it may be outpaced by the other route and have little influence on the result. We therefore ran implemented dual route models — the DRC and CDP+, and a modified DRC, the DRC-FC — to see if they could capture the data by changing parameters on a participant by participant basis.

In these simulations, we selected for each participant the parameter values giving the best correlation with his or her individual item mean response times (once first phoneme was partialed out). This technique gives a more complete view of model performance than techniques that fit only condition means or effects. These non-item-level model selection procedures would indeed have found parameters that made better predictions about the effects we built into the experiment. However, the parameters could make very bad predictions about other unexamined (and possibly unknown) effects that contribute to the response times of the items in the experiment and still be selected. Regardless of the effects that we intend to use to *interpret* the modeling data, all effects place constraints (of the compatibility type discussed in §1.1.3) on how the cognitive mechanisms must operate.

Whilst, unsurprisingly, the fitting procedure gave the models some capability to predict a large effect of some variable for those who showed a greater effect of that variable, more detailed examination suggested that the models did not capture this variation in a way that was compatible with the data.

6.2.1. DRC

The parameters best fitting the subject-and-item-level data for the DRC led the model to produce no consistency effect, a negative correlation between the frequency and neighborhood size effects, and no correlations between the lexicality effect and either the neighbourhood size or exception effect, inconsistent with the data. The model with these parameters did correctly capture the three other correlations of interest, as well as the other overall patterns, albeit not always with their full magnitude.

6.2.2. CDP+

The parameters best fitting the subject-and-item-level data for the CDP+ led the model to produce no consistency effect, an exception effect that was restricted to first-position irregularities, a negative correlation between lexicality and exception effects, a negative correlation between frequency and nonword length effects, and no correlations of neighborhood size with either frequency or regularity, inconsistent with the data. The model with these parameters did correctly capture the other two correlations of interest, as well as the other overall effects, albeit not always with their full magnitude.

6.2.3. DRC-FC

The parameters best fitting the subject-and-item-level data for the DRC-FC led the model to produce no consistency effect, and no correlations between the lexicality effect and the neighbourhood size effect, inconsistent with the data. The model with these parameters did correctly capture the five other correlations of interest, as well as the other overall effects, albeit not always with their full magnitude.

6.3. Interpretation of the data

6.3.1. General speed does not accommodate all individual differences

There was a strong overall pattern of positive correlations in the raw effects. Adjustments for general speed indicated that many of these positive correlations were attributable to a general speed multiplier affecting all the (signal) processes. However, five significant non-artefactual correlations persisted, which were indicative that systematic individual differences exist that are not due to general speed. Verbal dual route interpretations could be placed on these, but we sought to replace these verbal interepretations with the modeling.

6.3.2. Lexical quality and reading experience

The concept of lexical quality seemed important for interpreting the data, despite its having no clear analog in the models of interest. Lexical quality (e.g., Andrews & Hersch, 2010; Perfetti, 1992) refers to the idea that the specific orthographic, phonological or semantic-syntactic information associated with a particular word can vary in how well-specified it is; being poorly specified means that particular information — such as the ordering of adjacent I and E — about a word is missing. The degree of underspecification of a word's information can vary so that receptive language outperforms production (e.g., recognizing a correctly spelled word despite spelling it incorrectly yourself). Although lexical quality is a concept that applies at the level of the individual word within the individual person, it also can be applied as an average over the words a person knows: individuals who have learned a language more — due to greater experience and/or greater efficiency of learning — will have on average better lexical quality across lexical items. This offers a natural explanation of otherwise difficult phenomena, such as the observed correlation of neighbourhood size and frequency effects, because it naturally links both neighbourhood and frequency effects to (lack of effective learning) experience. In particular, the constraint-based approach implied by lexical quality means that as words become better learned, they might be more activated by perfect matches but they should also be more readily distinguished from close mismatches.

Consistent with this, older participants (having more reading experience) also showed less facilitatory neighbourhood size effects, consistent with finer-tuned input representations — that is, higher lexical quality — reducing the generalization that would support such facilitation.

This interpretation also offers a possible view on analyses that suggest frequency effects can be effectively predicted from vocabulary size. Developmentally, lexical quality is determined by exposure to lexical items and by ability to form and refine representations from such exposure. These are the same factors that would be responsible for vocabulary size; indeed vocabulary can be seen as a person-level measure of lexical quality computed on an all-or-none basis over possible lexical items (i.e., all words in the dictionary). It is therefore possible to interpret results that suggest an influence of vocabulary size as instead being due to lexical quality (on average for an individual over the relevant lexical entries).

6.3.3. Changes within individuals within the experiment

Session-to-session changes. For the most part, the average participant responded slower in the later sessions of naming, although such patterns may not bear out at the level of the individual participant (see Adelman et al., 2013, for data for many sessions). This was not associated with an increase in the sizes of the effects of interest, which might be expected from a stretching of the main process of naming. That is, the slowing occurred in the "intercept" part of the process unaffected by the identity of the stimulus. Indeed, the only significant interactions with session were such that effects became smaller over sessions, because the slowest conditions (exception words, long nonwords) were relatively immune to the slowing effect. This may reflect participants strategically compensating for their slowing only for those responses they perceived at risk of falling after stimulus offset.

Trial-to-trial changes. To be consistent with the models in question and typical word naming experimental analyses, we have excluded possible effects of long-term priming from our analyses. The implied assumption is that such effects will average out over the several different orderings of the stimulus list and result in all words being read on average the same amount faster than the baseline notionally unprimed condition. This assumption is problematic for the analyses here, however, because each participant only saw three orderings, which does not allow for a great deal of averaging out, and the natural estimate of reliability is the test-retest one. Worse, though, there are good reasons to suppose the assumption is wrong. First, words may differ in their susceptibility to priming (e.g., Kinoshita, 2006); if a truly neutral baseline is not achieved, more primeable words will have shorter RTs on average, but not reliably. Second, words may differ in the number of possible preceding words that would prime them[12]; words with more related primes will on average be more primed, thus having shorter RTs. The latter type of process may be particularly relevant to the consistency effect, as there is long-term rime priming (Seidenberg et al., 1984), and, indeed, this was an effect that showed low consistency in our data. Clearly, models that do not adapt to the trial sequence could not give a correct

[12] Indeed, this may be the source of primeability — words that have many related primes may *typically* be in the primed state, and thus be difficult to prime.

account of the effect if this were true.

6.4. Modeling interpretations

6.4.1. DRC

For the DRC, the key problems in the simulations related to effects that have previously been identified as problematic for this model and its relatives. The patterns with the neighbourhood size effect — its overall inhibitory effect in the model (albeit to a degree that would be undetectable in the data) and its negative correlation with the frequency effect — were suggestive that a modification to the model is needed that weakens its knowledge of the spelling correspondences of lower-frequency words; it is unclear how this could occur without broader changes in its letter level to remove the heavily criticised assumption of perfect knowledge of letter position (e.g., Davis, 2010; Grainger & van Heuven, 2003). A representation that is somewhat redundant could have the desired effect. The somewhat weak but correct consistency effect with flexibility in parameters is suggestive of an incomplete ability to account for the effect. Indeed, although significant, the effect was so small once converted from cycles to milliseconds that we would expect essentially no power to detect the effect in the data, where a strong effect was in fact observed. As models become more and more refined, increased attention to the precision of predicted effects will be warranted.

Given that (for instance) CRATE and LATE are wholly dissimilar in the DRC model both orthographically and phonologically, the absence of a strong consistency effect is perhaps not surprising. A complete sensitivity to rime structure requires some alteration to the orthographic coding, as was already indicated, and to the phonological coding, perhaps to reflect the linguistic structure of the syllable, as in CDP+[13].

[13] However, in modeling not presented in detail here, such changes were not sufficient to produce a consistency effect of appropriate magnitude when the parameters are constrained by other effects in the data.

6.4.2. CDP+

For the CDP+, the problems appeared more systemic. The model parameters that best fitted the data had an excessively severe position of irregularity effect, such that late irregularities and inconsistencies did not exert an influence from the nonlexical route. This is suggestive that properties of the predictions from the nonlexical route were discrepant from the data in a way that meant that fits to the data were better when the nonlexical route's influence was minimised. Whilst it is known that the CDP+ model can produce consistency effects on the average of conditions selected to differ in consistency (Perry et al., 2007), examining only differences in condition means can obscure other problems in the predictions (see, e.g., Besner, Twilley, McCann & Seergobin, 1990), and the item R^2 criterion was chosen to prevent models taking advantage of this obfuscation opportunity. Although the ability to fit consistency condition means has previously been taken as an advantage for the more graded nonlexical route of the CDP+, there are pseudoword stimuli for which human naming responses are more like those of the DRC than those of the CDP+ (Pritchard et al., 2012). One possible consideration here is that the CDP+ is trained on monosyllabic words, whereas no such constraint applies in humans' learning of spelling-sound relationships. If this is the problem, then the newer CDP++ (Perry, Ziegler & Zorzi, 2010) would resolve the problem.

However, the problems the CDP+ has could also be more to do with the structure of the network that learns the spelling-sound correspondences. The network's structure is sensitive to a form of consistency that is correlated with, but not the same as, the type of consistency manipulated in experiments showing the effect in humans. CDP+ is sensitive only to grapheme-level consistency, not rime-level consistency. CDP+'s nonlexical route has no way to represent specifically that OOD is often pronounced /ʊd/ (as in hood) or OOM is often pronounced /uːm/; it instead would have to represent that (1) OO makes /ʊ/ and /uː/ both somewhat likely; and (2) that a post-vocalic M makes /uː/ a more likely vowel and a postvocalic D makes /ʊ/ a more likely vowel, and these effects would be independent. (Over the whole vocabulary, this might not be the overall learned pattern — the point here is to illustrate the types of representations possible in the two-layer network.) That is, the effect of the orthographic consonant on the phonological vowel can not be specific to the particular orthographic vowel context: /uː/ would (ignoring the influence of the other learning patterns) be more activated by the nonlexical route in response to RIM than RID, and /ʊ/ is more activated by the nonlexical route in response to RID than RIM. This differs substantially from other connectionist accounts (e.g., Plaut et al., 1996) because it uses a network with no hidden units to map from graphemes to phonemes, and so can not learn larger units than a grapheme. In common with these models, however, the position-specific nature of the spelling-sound relationships learned by CDP+ is a likely cause of problems with nonwords that have graphemes in atypical positions, where people do clearly attempt to generalize across position.

6.4.3. DRC-FC

The changes made to the DRC to make the DRC-FC circumvented some problems with the fit to the effect correlations. By placing the frequency effect on the links between orthography and phonology, the deleterious effects of the combination of neighborhood activity and a benefit in bias for frequent words is achieved. By limiting the influence of frequency on orthographic nodes to feedback from the phonological nodes, higher-frequency neighbors were unable to overwhelm items, even when neighborhood size effects were large. However, it seems unlikely that there are no purely orthographic frequency effects, as this model would suggest. Moreover, in this case, there is no obvious single root cause that produces neighbourhood and frequency effects — the positive correlation between them is likely produced by covarying parameters, rather than a constraint in the model. Moreover, the model also still did not produce frequency and length effects of a magnitude consistent with the data.

6.4.4. PDP

Other connectionist, parallel-distributed-processing (PDP: e.g., Harm & Seidenberg, 1999; Plaut et al., 1996; Seidenberg & McClelland, 1989) models were, however, excluded from our consideration for three main reasons. The first is that it is at odds with the intent (and epistemology) of the modellers who produced them: Whilst the modellers responsible for the DRC and CDP+ treat the details of each effect as a constraint that should be included into a complete true model, modellers using PDP emphasise that the explanation or principle of a certain phenomenon is true, and explicitly deny that the model is itself a representation of the truth at any other level of abstraction. As such, an attempt to falsify such a model is meaningless, as it would falsify no claim of the author. The second reason is a corollary of this approach. Given that no model is intended to be correct, we would have no rule to select the correct model for our data: no model is correct, no model supersedes an earlier one, and each phenomenon deserves and requires its own model. The final reason follows from both of the preceding ones: no PDP model qualitatively captures all of the effects, because no model is intended be true and each phenomenon may and indeed should be simulated separately from any other according to the arguments of these PDP modellers (Seidenberg & Plaut, 2006).

Nevertheless, in light of these data, and some previous indications, we can offer some idea how some of the patterns seen here might be accommodated in models related to those seen in the PDP literature. In many respects, the considerations mirror those of trade-offs between routes in dual-route models insofar as there are factors in PDP models that produce trade-offs between item-specific processing and processing based on generalization across items. For instance, the number of hidden units between orthographic and phonological layers affects the extent of generalization: Lower numbers of hidden units tend to require greater compression of the input information, producing greater generalization, because having hidden units that read only one or a few words cannot work when such units are scarce. Similarly, modulation of an input gain parameter can control item-based versus generalization-based processing, with higher levels of input gain leading to more item-based processing. By either mechanism, more item-based processing would be expected to be associated with less benefits from generalization, slowing low-frequency words and pseudowords, but giving relative benefits to inconsistent or irregular words. That is, one could naïvely predict the frequency and lexicality effects would be positively correlated with each other and negatively correlated with exception effects.

An alternative consideration is the amount of learning that has taken place. First, as reading becomes more efficient, general speed should increase, reducing all effects. Further, as we are considering relatively expert readers, we would expect that the asymptote of learning is a consideration for the more frequent words, so that at greater levels of learning frequency effects should diminish, and more rapidly so than just the general speed effect. At the same time, as learning at the orthographic level proceeds, spurious activity for neighbors should be eliminated, reducing the potential for neighborhood facilitation.

Some research on modeling nonword pronunciation with PDP models has been offered as evidence that PDP models account for individual differences (Zevin & Seidenberg, 2006). However, the individual differences in which pronunciation was chosen for nonwords was caused in these models by minor random variations in the training schedule. Whilst it is easy to see how such variation could produce a bias in how novel stimuli are processed, it is in no way clear that such variation should offer an appropriate account of systematic shifts in response times to stimuli with particular characterstics, let alone how such effects should correlate.

Indeed, as with the other models, this kind of verbal analysis can only offer a set of possible considerations, which may be incomplete, and only some of which may be relevant for the particular model structure and parameters selected to fit each participant's data. To assess such predictions would require an appropriate PDP model in terms of *both* producing response times and covering an adequate range of effects. To our knowledge, this would be a new model (the model of Chang et al., 2012, might, for example, be extended to produce response times rather than only error scores that may be presumed to be monotonic in response time). We do not attempt any such PDP model development here, and indeed proponents of the PDP approach, and indeed well-known proponents of the PDP approach have argued that the development of such a model would be undesirable (Seidenberg & Plaut, 2006); proponents of competing approaches view this as an argument for unfalsifiability (Rastle & Coltheart, 2006).

6.5. Conclusion and future directions

We replicated the key theoretical phenomena of reading aloud, and demonstrated that reliable individual differences are observed even after accounting for differences in general speed. We then identified critical correlations among these effects (within individual participants), and used these correlations as criteria for testing computational models of reading aloud. After showing that DRC and CDP+ capture some but not all of these correlations, we developed and tested a model based on DRC, the DRC-FC, which did indeed yield more accurate predictions of these correlations.

These data offer important new constraints on theories of visual word recognition and highlight areas of potential improvement within DRC and CDP+ accounts of word naming. Extending models to accommodate individual differences is an important future step to allow more complete theories of reading to be more accurate and more applicable to real-world concerns about reading and reading acquisition. Moreover, investigations of individual differences in more naturalistic reading tasks similarly need to be extended beyond the comparison of good and poor readers (e.g., Ashby et al., 2005), informative as these may be. However, more than just accommodating these differences, models must also offer an understanding of these individual differences. Such an understanding would specify the processes — and the parameters of these processes — that differ to underlie differences in the effects. Such an understanding would be augmented by linking these parameters to other, more general, cognitive abilities. Such model-specific steps would be premature without a model that can accommodate the effects. We view the present data as one empirical step towards a more global model of reading.

Acknowledgements

This work was funded by ESRC Grant RES-062-23-0545. We thank Gordon Brown, Marc Brysbaert, Victor Kuperman and two anonymous reviewers for their constructive comments.

References

Adelman, J. S. (2012). Methodological issues with words. In J. S. Adelman (Ed.), *Visual word recognition Vol. 1: Models and methods, orthography and phonology* (pp. 116–138). Hove, England: Psychology Press.

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

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.

Adelman, J. S., & Brown, G. D. A. (2008b). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *114*, 214–227.

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.

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1037–53.

Adelman, J. S., Sabatos-DeVito, M. G., & Marquis, S. J. (in prep.). Individual differences in word naming: External predictors of item effects. Manuscript in preparation.

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.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*, 439–461.

Andrews, S. (2012). Individual differences in skilled visual word recognition and reading: The role of lexical quality. In J. S. Adelman (Ed.), *Visual word recognition Vol. 2: Meaning and context, individuals and development*. Hove, England: Psychology Press.

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

Andrews, S., & Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds? *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1052–1086.

Ashby, J., Rayner, K., & Clifton, J., C. (2005). The reading patterns of highly-skilled and average readers: Evidence from eye movements. *Quarterly Journal of Experimental Psychology*, *58A*, 1065–1086.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (Release 2).

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.

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

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual Word Recognition, Vol. 1: Models and Methods, Orthography and Phonology* (pp. 90–115). Hove, England: Psychology Press.

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.

Besner, D., Moroz, S., & O'Malley, S. (2011). On the strength of connections between localist mental modules as a source of frequency-of-occurrence effects. *Psychological Science*, *22*, 393–398.

Besner, D., Twilley, L., McCann, R. S., & Seergobin, K. (1990). On the association between connectionism and data: Are a few words necessary? *Psychological Review*, *97*, 432–446.

Blais, C., & Besner, D. (2007). Reading aloud: When the effects of stimulus quality distinguish between cascaded and thresholded components. *Experimental Psychology*, *54*, 215–224.

Brown, S. D., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments and Computers*, *35*, 11–21.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990.

Butler, B., & Hains, S. (1979). Individual differences in word recognition latency. *Memory & Cognition*, *7*, 68–76.

Castles, A., Davis, C., Cavalot, P., & Forster, K. (2007). Tracking the acquisition of orthographic skills in developing readers: Masked priming effects. *Journal of Experimental Child Psychology*, *97*, 165–182.

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

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition*, *28*, 143–153.

Coltheart, M. (1980). Reading, phonological recoding, and deep dyslexia. In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (pp. 197–226). London: Routledge.

Coltheart, M. (2012). Dual-route theories of reading aloud. In J. S. Adelman (Ed.), *Visual word recognition Vol. 1: Models and methods, orthography and phonology* (pp. 3–27). Hove, England: Psychology Press.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589–608.

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

Coltheart, M., & Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1197–1211.

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.

Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods*, *40*, 791–794.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review*, *117*, 713–758.

Diependale, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, *66*, 843–863.

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

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales, & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 257–287). Amsterdam: North-Holland.

Forster, K. I. (2012). A parallel activation model with a sequential twist. In J. S. Adelman (Ed.), *Visual word recognition Vol. 1: Models and methods, orthography and phonology* (pp. 52–69). Hove, England: Psychology Press.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627–635.

Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 361–379.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 674–691.

Grainger, J., Muneaux, M., Farioli, F., & Ziegler, J. C. (2005). Effects of phonological and orthographic neighbourhood density interact in visual word recognition. *Quarterly Journal of Experimental Psychology*, *58A*, 981–998.

Grainger, J., & van Heuven, W. (2003). Modeling letter position coding in printed word perception. In P. Bonin (Ed.), *The Mental Lexicon* (pp. 1–24). New York: Nova Science.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491–528.

Jared, D. (1997). Spelling-sound consistency affects the naming of high frequency words. *Journal of Memory and Language*, *36*, 505–529.

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language*, *46*, 723–750.

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, *29*, 687–715.

Kinoshita, S. (2006). Additive and interactive effects of word frequency and masked repetition in the lexical decision task. *Psychonomic Bulletin & Review*, *13*, 668–673.

Kuperman, V., & Van Dyke, J. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*, 42–73.

Mathey, S. (2001). L'influence du voisinage orthographique lors de la reconnaisance des mots écrits [The influence of orthographic neighbourhood on visual word recognition]. *Revue Candienne de Psychologie Expérimentale/Canadian Journal of Experimental Psychology*, *55*, 1–23.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375–407.

McClung, N. A., O'Donnell, C. R., & Cunningham, A. E. (2012). Orthographic learning and the development of visual word recognition. In J. S. Adelman (Ed.), *Visual word recognition Vol. 2: Meaning and context, individuals and development* (pp. 173–195). Hove, England: Psychology Press.

Morton, J. (1964). A preliminary functional model for language behaviour. *International Audiology*, *3*, 216–225.

Mousikou, P., Coltheart, M., Finkbeiner, M., & Saunders, S. (2010). Can the drc computational model of reading offer a valid account of the masked onset priming effect? *Quarterly Journal of Experimental Psychology*, *63*, 984–1003.

Mulatti, C., Reynolds, M. G., & Besner, D. (2006). Neighborhood effects in reading aloud: New findings and new challenges for computational models. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 799–810.

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.

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.

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the connectionist dual process (CDP++) model. *Cognitive Psychology*, *61*, 106–151.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired reading: Computational principles in quasiregular domains. *Psychological Review*, *103*, 56–115.

Pritchard, S. C., Coltheart, M., Palethorpe, S., & Castles, A. (2012). Nonword reading: Comparing dual-route cascaded and connectionist dual-process models with human data. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 1268–1288.

Rastle, K., & Brysbaert, M. (2006). Masked phonological priming effects in English: Are they real? Do they matter? *Cognitive Psychology*, *53*, 97–145.

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

Rastle, K., & Coltheart, M. (2006). Is there serial processing in the reading system; and are there local representations. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing* (pp. 3–24). Hove, UK.: Psychology Press.

Reynolds, M., & Besner, D. (2002). Neighbourhood density effects in reading aloud: New insights from simulations with the DRC model. *Canadian Journal of Experimental Psychology*, *56*, 310–318.

Reynolds, M., & Besner, D. (2004). Neighbourhood density, word frequency, and spelling-sound regularity effects in naming: Similarities and differences between skilled readers and the dual route cascaded computational model. *Canadian Journal of Experimental Psychology*, *58*, 13–31.

Roberts, M. A., Rastle, K., Coltheart, M., & Besner, D. (2003). When parallel processing in visual word processing is not enough: New evidence from naming. *Psychonomic Bulletin & Review*, *10*, 405–414.

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.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Seidenberg, M. S., & Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing* (pp. 25–49). Hove, UK.: Psychology Press.

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.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, *8*, 411–416.

Spieler, D. H., & Balota, D. A. (2000). Factors influencing word naming in younger and older adults. *Psychology and Aging*, *15*, 225–231.

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.

Thompson, G. B., Connelly, V., Fletcher-Flinn, C. M., & Hodson, S. J. (2009). The nature of skilled adult reading varies with type of instruction in childhood. *Memory & Cognition*, *37*, 223–234.

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

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 53–79.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979.

Yates, M. (2005). Phonological neighbors speed visual word processing: Evidence from multiple tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1385–1397.

Zevin, J. D., & Seidenberg, M. S. (2006). Simulating consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language*, *54*, 145–160.

Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, F. X., & Perry, C. (2008). Developmental dyslexia and the dual route model of reading: Simulating individual differences and subtypes. *Cognition*, *107*, 151–178.

Zorzi, M. (2000). Serial processing in reading aloud: No challenge for a parallel model. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 847–856.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? A connectionist dual-route model. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1131–1161.

Appendix A. Frequency stimuli

Stimuli used for estimating the frequency effect are presented in Table A.1.

Appendix B. Neighbourhood stimuli

Stimuli used for estimating the neighbourhood size effect are presented in Table B.1.

Appendix C. Length stimuli

Stimuli used to estimate length effects and the lexicality effect are presented in Table C.1.

1 ppm	3 ppm	9 ppm	27 ppm	81 ppm
boo	bug	ban	bay	box
bash	bout	bike	bomb	base
brawl	brink	bride	brick	brown
dud	din	den	dad	die
dent	dope	dock	damp	desk
foal	foil	fork	file	film
gull	gosh	goat	goal	game
hoot	hiss	hint	hide	hell
chomp	chunk	cheer	chain	chair
cud	kin	cab	kid	cup
clove	clown	clash	cloud	clean
lurk	lick	lust	link	lack
mush	moth	mill	mess	milk
nigh	knob	nail	knee	note
pelt	putt	pond	pack	pain
rink	rake	ripe	rope	rock
seep	sage	suck	song	send
shank	shoal	shine	sheet	share
steed	stunt	stack	storm	stone
tab	tar	tub	tin	tax
teak	toil	tilt	tent	team

Table A.1: Stimuli used for estimating frequency effect

Appendix D. Adjusting for general speed

A core difficulty in assessing individual differences in response time effects in terms of correlations is that a correlation will be induced if there is a general speed coefficient that is a multiplier on the central processes; in many models this is the slope for converting cycles into response times. Whilst in modelling applications, this parameter can simply be included for every participant, for a more qualitative understanding of correlations, removing such an effect is desirable. Consider a numerical example. Suppose that the to-be-modelled cognitive item-based task's response times are affected by three (standardised, orthogonal) variables of the items — A, B, and C — in a manner that is consistent with a linear regression, and the residual noise process has an independent, irrelevant, weight (in contemporary models of word naming, this is true: only item mean RTs come from central parts of the model; RT variability

low *N*	**high** *N*
bird	bell
buzz	buck
beard	*bitch
disc	deed
dumb	dole
dirt	doll
dawn	dare
foul	fame
fact	face
gulf	gang
garb	gush
growl	graze
helm	hilt
hurl	hash
howl	hack
high	home
yak	yam
cusp	coop
keel	cone
cube	cork
cult	cart
keen	cake
kiss	cave
coal	cope
leave	light
mesh	mole
mere	mile
romp	rave
rich	rate
surf	sane
self	seed
stump	stale
void	vest
veil	vice
wisp	wane
zinc	zeal
loin	loon
noun	*nope
crypt	creed
nil	nap
loaf	lest
loud	lake

Table B.1: Stimuli used for estimating neighbourhood size effects. Pairs with an asterisked item were not included in the analyses.

Words			Nonwords		
3	4	5	3	4	5
bib	bile	budge	bul	beab	beich
bin	bunk	bulge	bym	bibe	bouse
beg	barn	boost	bov	bobe	beash
bid	bond	bunch	bes	booc	boarm
dig	dull	depth	dyt	deef	dunch
dry	drop	drive	dro	dran	drine
fun	fill	fault	fud	filk	fodge
gut	gale	goose	gol	gise	gunch
gun	gate	guess	gaz	goaf	gudge
hag	hoop	hunch	hin	hean	hulch
hug	hawk	haunt	hab	heek	hetch
hut	hang	hence	hol	hont	haise
cod	cape	curve	kem	coth	cange
cap	code	coach	kuc	cose	kaunt
cut	cost	court	kav	kive	kutch
lob	lobe	lurch	lec	lenk	lerge
lid	lump	lodge	loy	lilk	lirge
leg	loss	lunch	lig	loog	louch
mar	meek	munch	moy	marf	medge
mob	mode	marsh	mib	meap	mouch
mix	myth	mount	mup	moop	metch
nip	nape	nudge	nev	noof	nalve
net	neat	nurse	nar	nowl	ninch
pun	pert	purge	pag	peef	petch
peg	pike	punch	pem	perb	pedge
pen	pump	pitch	poy	paim	purpe
rib	rout	roost	rus	rull	raint
rub	rude	ridge	ruv	roog	rorse
sin	sing	solve	suz	soob	sudge
tip	tune	toast	tol	telp	tarch
web	weep	wedge	wec	wiln	wouse

Table C.1: Stimuli used for estimating length and lexicality effects

within items is treated as a purely statistical issue). Participant 1 has regression coefficients 5, 15, and 5 respectively for these variables, and residual standard deviation (s.d.) 25; Participant 2 has coefficients 10, 30, and 10, and s.d. 50; Participant 3 has coefficients 10, 30, and 10, and s.d. 100; Participant 4 has coefficients 10, 10, 30, and s.d. 200; and Participant 5 has coefficients 5, 15, and 15, and s.d. 50. Adjusting for general speed means identifying Participants 1, 2 and 3 as examples of the same 1:3:1 (A:B:C) ratio pattern, Participant 4 as an example of a different 1:1:3 pattern, and Participant 5 as different from all of them, having a 1:3:3 pattern.

A common solution — and that used by Yap et al. (2012) — is to use standardised regression coefficients for regressions separately on each participant (or at least per-participant z-scored RTs in separate analyses). If there were no residual noise, that is, all variability in observed item means were due to item properties, this would have the desired effect. However, if there is contamination with noise that is not wholly proportional to overall speed (i.e. participants are not equally reliable) as in the example, then the strength of this noise will affect the adjusted strength of the item variables. In the example, whilst Participant 1 and 2 are correctly identified with the same set of strengths, Participant 3 receives weaker corrected strengths for all the item variables, due to the greater noise influence. It is true that more variance in the individual trials is explained by those variables for the first two participants, but this is unlikely to be theoretically relevant: Response time variability is not usually attributed to the central cognitive process of interest, instead in whole or in part being attributed to other processes, such as variability in response execution; certainly, contemporary naming models give no account of within-participant, within-item response time distributions. Perhaps more concerning is the comparison of Participants 2 and 4. Standardising Participant 4's highly variable response times will give a much greater compression of the effects than occurs for Participant 2; as a consequence, Participant 4's standardised effect for C will be less than Participant 2's standardised effect for C, despite this variable accounting for the majority of the systematic effect for Participant 4, and another variable

(B) doing so for Participant 2.

Another (incomplete) solution might be to use something related to the actual ratio of the observed regression coefficients, such as weighting the coefficients so that they (or their squares) add to one. If all effects are known, then this achieves the goal (up to the quality of the data). However, if the empirical regression lacks one or more variables then it does not produce the correct pattern: In our example, suppose C is omitted, then (assuming the remaining coefficients are perfectly estimated, for simplicity) whilst Participant 4 is correctly identified as different (1:1), the other four participants are all identified as having a 1:3 (A:B) ratio pattern. That is, the relatively smaller influence of A and B for Participant 5 is not detected.

These two possibilities combined isolate the problem: What is needed is a reasonably good estimate of the residual standard deviation for a given participant that remains after accounting for all the variance caused by item variables. The z-scoring approach uses the whole observed variance as the estimate of the systematic variance, underestimating the residual variance as zero. The naïve ratio approach uses observed variance attributable to known variables as the estimate of the systematic variance, overestimating the systematic variance by including variance due to unknown variables as well as some noise variance. Clearly, the true variance due to item variables — both known and unknown together — is the same as the true variance due to items. If a participant has, however, completed each item only once, directly estimating (by analysis of variance of the data for an individual participant) the variance due to items exhausts all the variance (because item and trial are aliased), including variance due to noise; no correction is available because there is no independent estimate of the noise. On the other hand, if, as here, participants read the words more than once, the estimated item variance does not soak up all the noise (only part of it); it is the item by session interaction that can not be separated from the noise in this design. That is, the variance of the item means is less contaminated (as an estimate of variance due to items) with noise variance than the variance of all the trials, but a further step can be taken: An unbiased variance

Factor	df	Sums of squares	Mean square (MS)
Session	2	99994	49997
Item	683	2402613	3518
Residual	1309	1478906	1130

Variance component for Item = (MS(Item) - MS(Residual))/(#Sessions)
= (3518-1130)/3 = 795.98

$$\text{Divisor} = \sqrt{795.98 \times 683/684} = 28.19 \text{ ms}$$

Table D.1: Calculation of the estimate of the variance component associated with Item.

component estimate for the items can be calculated. This variance component can be used instead of the total variance as the square of the denominator in a *z*-score-like calculation; in the following, we do so using the (unbiased) ANOVA estimate of the variance component for simplicity. If the assumption that the noise process is scaled as part of the general speed is true, this method produces results equivalent to those produced by plain *z*-scoring (all adjusted RTs are multiplied by the same constant); if it is not, then it does not allow differences in the noise process to affect the interpretation of the speed of the other processes.

To make this concrete, Table D.1 shows the ANOVA table from which the variance component was calculated for one participant.

The standard deviation estimated from the variance component for use in standardisation for each participant had mean 54.3 ms and standard deviation 14.7 ms; it correlated with the participants' overall standard deviations .661, with their mean RTs .522, and with the individual effects: frequency .473; neighbourhood size .148; word length .264; nonword length .665; exception (Jared) .380; Exception (Rastle & Coltheart) .581; consistency .388; position of irregularity .384; and lexicality .616.

Appendix E. Parameter ranges used in DRC and CDP+ simulations

Table E.1 presents the lowest and highest value used for each parameter of the models. For parameters than can take only integer values, 21 equally spaced values were used with equal probability; for other parameters, 51 equally spaced values were used with equal probability.

Parameter	DRC, CDP+ or DRC-FC?	Min.	Max.
Activation	All	0.18	0.38
Frequency scaling	DRC and CDP+	0.02	0.6
Frequency weighting	DRC-FC	12	360
Stopping criterion	All	0.25	1
Resting criterion	CDP+	0.00001	0.00101
Feature-letter excitation	All	0.001	0.02
Feature-letter inhibition	All	−0.25	−0.05
Letter-orthography excit.	All	0.04	0.24
Letter-orthography inhib.	All	−0.8	−0.4
Letter-letter inhib.	All	0	0.1
Letter decay	All	0	0.2
Orthography-phonology excit.	All	0.1	2.0
Orthography-letter excit.	All	0.2	0.8
Orthography-letter inhib.	All	0	.1
Orthography-orthography inhib.	All	−0.2	0
Orthography decay	All	0	0.1
Phonology-phoneme excit.	All	0.05	0.25
Phonology-phoneme inhib.	All	−0.4	0
Phonology-orthography excit.	All	0.02	2
Phonology-phonology inhib.	All	−0.2	0
Phonology decay	All	0	0.1
Phoneme-phonology excit.	All	0	0.4
Phoneme-phonology inihb.	All	−0.2	−0.05
Phoneme-phoneme inhib.	All	−0.4	−0.05
Phoneme decay	All	−0.02	−0.01
Nonlexical route strength	All	0.01	0.2
Nonlexical route threshold	CDP+	0.005	0.25
Temperature	CDP+	1.0	11.0
Learning rate	CDP+	0.005	0.2
Nonlexical route start	DRC and DRC-FC	0	20
Nonlexical route step	All	0	20
Phonics cycles	CDP+	0	200
Vocabulary cycles	CDP+	0	200

Table E.1: Parameters ranges used for DRC and CDP+ simulations.