PLOS Complex Systems

RESEARCH ARTICLE

Modelling a complex cognitive system with limited data: Optimization and generalization in a computational model of reading aloud

Conrad Perry[1]*, **Marco Zorzi**[2,3], **Johannes C. Ziegler**[4]

1 Faculty of Health and Medical Sciences, Adelaide University, Adelaide, Australia, **2** Department of General Psychology and Padova Neuroscience Center, University of Padova, Padova, Italy, **3** IRCCS San Camillo Hospital, Venice-Lido, Italy, **4** Aix-Marseille University and Centre National de la Recherche Scientifique, Centre de Recherche en Psychologie et Neurosciences (UMR 7077), Marseille, France

* ConradPerry@gmail.com

OPEN ACCESS

Citation: Perry C, Zorzi M, Ziegler JC (2025) Modelling a complex cognitive system with limited data: Optimization and generalization in a computational model of reading aloud. PLOS Complex Syst 2(11): e0000074. https://doi.org/10.1371/journal.pcsy.0000074

Editor: Ning Cai, Beijing University of Posts and Telecommunications, CHINA

Received: April 1, 2025

Accepted: September 22, 2025

Published: November 6, 2025

Data availability statement: All data files and the R scripts used are available as Supplementary Materials. They can be downloaded here: https://osf.io/7xkjf

Funding: This research was supported by the Australian Research Council (Grant

Abstract

Optimizing the Connectionist Dual-Process Model of Reading Aloud (CDP; Perry et al., *Journal of Memory and Language, 134*, 104468) using large-scale empirical datasets has been shown to enable accurate predictions of independent datasets that were not used for optimization. Here, we investigated CDP's generalization performance when optimized on small datasets consisting of words, nonwords, or a combination of both. The results showed CDP's quantitative performance was similar on both small and large datasets except when optimized on small nonword-only datasets. Additionally, CDP's predictions generally surpassed those derived from regression-based models, suggesting it had good generalization performance. Using sloppy parameter analyses, we also found that a small number of parameters determined most of CDP's quantitative performance and that the parameters which did this were similar across both small and large datasets. These findings suggest that the CDP does not overfit the data, even when optimized on very small numbers of stimuli. They also give insight into the role the parameters play in generating psycholinguistic effects. More generally, the findings show that when an underlying cognitive architecture constrains behavior, complex systems like reading may be analyzed and understood using very limited data. This is important as it shows that computational modelling can be used in some situations where data is scarce but understanding the system remains crucial.

Author summary

Reading is a complex cognitive process that involves the coordination of multiple components—from recognizing letters to producing spoken words. Like many complex systems, it is often assumed that meaningful modelling of reading

DP210100936 to CP; https://www.arc.gov.au/ and the Institute of Convergence ILCB (France 2030, ANR-16-CONV-0002 to JZ; https://www.ilcb.fr/. MZ is supported by the Italian Ministry of Health (Ricerca Corrente to IRCCS San Camillo Hospital; https://www.sancamilloscientifico.it/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

requires large datasets. In this study, we show that this is not necessarily the case. Using a computational model of reading aloud—the Connectionist Dual Process (CDP) model—we demonstrate that accurate performance predictions on held-out datasets can be achieved even when the model parameters are optimized using a single small dataset. Indeed, CDP outperforms simpler statistical models, despite having the added challenge of generating the actual phoneme sequences to be used for speech production. We also identified which model parameters play a key role in shaping reading performance, offering insights into the dynamics of the reading system. Our findings suggest that even richly structured, complex systems like reading may be analyzed and understood using limited data, as long as the model incorporates the right assumptions about the underlying cognitive architecture. This opens the door to using computational modelling more broadly in situations where data may be scarce—but understanding the system remains crucial.

Introduction

Computational models in cognitive psychology can serve different purposes. Some are used to conceptualize problems without providing quantitative predictions [1]. Others use quantitative data in their development and evaluation [2]. In the area of reading aloud, models also differ considerably in their ability to fit human data, as measured by the amount of variance explained. Notably, across a large number of datasets, the Connectionist Dual-Process model [3–9] (CDP; see Fig 1) consistently outperforms other models, including the Dual-Route Cascaded model [10] and the Triangle model [11] (see Perry et al. [5,8]). These datasets have included reaction times (RTs) from large-scale datasets (also known as "mega-studies") [12,13] and small-scale experiments [14–16] including both word and nonword experiments. CDP represents a formal hypothesis about how the reading system works. The model includes many parameters because its representations and processes are grounded in experimental and neuropsychological evidence supporting their existence. Its development was not driven by the goal of minimizing the number of parameters or identifying the smallest set of processes capable of explaining the most variance across datasets. Indeed, we have previously shown that it is possible to remove many of the parameters in CDP without significantly impairing its quantitative performance in predicting reaction times in experiments with typical adult readers [5,17].

The latest version of an English CDP model is CDP++.parser [18], and this is the model that was used in all of the simulations below. Like all CDP models, CDP++.parser relies on two main reading routes: a lexical and a sublexical route. The lexical route contains orthographic and phonological representations of whole words. Separate semantic representations can also be accessed from this route, although this is not implemented in the model. The sublexical route computes the phonemes of words from graphemes (i.e., single letters or letter clusters that correspond to individual phonemes) without the help of whole word representations. When reading begins,

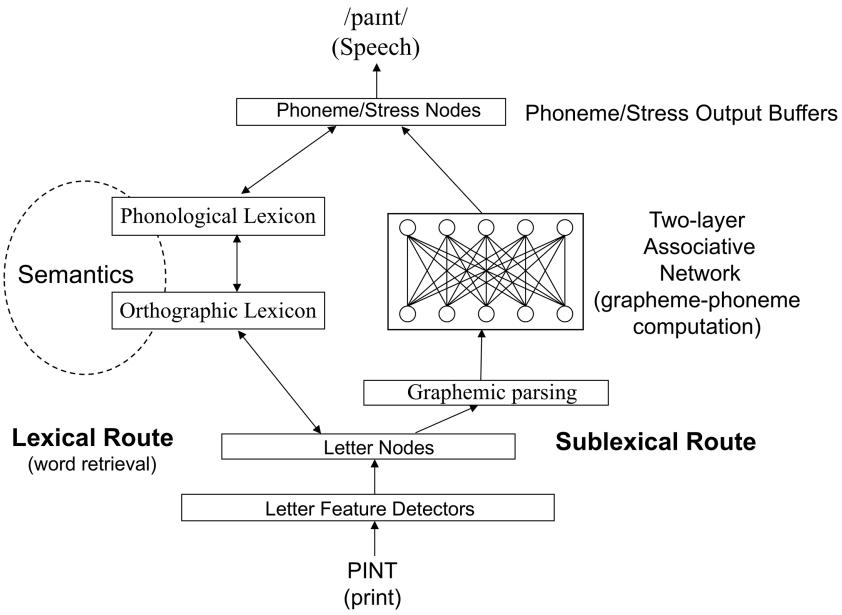

Fig 1. The CDP model of reading aloud. The figure displays the architecture of the computational model (reproduced from Perry et al. [2], with permission).

https://doi.org/10.1371/journal.pcsy.0000074.g001

these two routes are both activated and both contribute to the final phoneme sequence that is generated. CDP++.parser differs from earlier CDP models in two main ways. First, it can process disyllabic words—previously only CDP++ [6] had this capability, whereas earlier versions were limited to monosyllables. Second, the sublexical route operates differently. Notably, all versions of CDP use a graphosyllabic template that organizes graphemes in an onset–vowel–coda structure, mirroring phonological syllables. In all earlier versions of CDP (CDP++ [6], CDP+ [5], and CDP [4]), the mapping of letters to graphemes and their placement within the template was deterministic. In contrast, CDP++.parser uses a probabilistic mechanism when ambiguity exists—both in the selection of graphemes (e.g., the different pronunciations of "ph" in hi*ph*op vs. si*ph*on) and in their placement within the graphosyllabic template.

In terms of the specific computations, processing in CDP begins when a string of letters is presented, which activates letter features. Activation then flows to the letter level, from which both the lexical and sublexical routes can be activated. Within the lexical route, whole-word orthographic representations stored in an orthographic lexicon are activated from the letter level. These, in turn, activate whole-word phonological representations stored in a phonological lexicon. Finally, activation spreads to the phoneme output buffer, which uses an onset-vowel-coda structure and holds the phonemes to be used for speech production. Importantly, activation does not just flow one way and there are excitatory and inhibitory parameters between levels, as well as lateral inhibition at each representational level except for the feature level.

In the sublexical route, once letters are activated, the grapheme parser begins to extract letters one by one from left to right (following the reading direction in English). It then creates graphemes from them (i.e., single letters or letter clusters that correspond to individual phonemes) and assigns them to the graphosyllabic template. Phonemes are then computed from these graphemes via a simple two-layer associative network. This network has been trained on grapheme-phoneme correspondences from a set of rules and correspondences that exist in real words. The phonemes the network generates are then fed to the phonological output buffer. Since both routes are active simultaneously, phonemes from the lexical and sublexical routes may differ at a given position in the buffer. In such cases, the outputs from the two routes compete, and the final sequence of phonemes used for speech production is determined through this competition. Although not shown

in the simplified figure above, stress information is also represented in the lexicon, and stress output nodes are activated from the phonological lexicon and by the sublexical route. Further details of the model's architecture and dynamics can be found in Perry et al. [2,5,18] and a concise description of the CDP approach can be found in Zorzi [17].

Models and performance

The quantitative performance of models of reading aloud is widely considered as one of the key evaluation criteria [2,12,19]. It is typically assessed by correlating model-generated values with behavioral data from relevant datasets. The strength of these correlations is often compared to the performance of simple regression models that use psycholinguistic variables (e.g., word frequency and length, spelling-sound consistency) as predictors. However, defining what constitutes "good performance" is not straightforward. In this respect, Spieler and Balota [12] have suggested that a computational model should be at least as good as a multiple regression equation including all predictors known to be important. We have followed this recommendation and tested several versions of CDP as they were developed [5–8,9]. Earlier versions of CDP, which relied on hand-tuned parameters [5,6,8], achieved performance levels comparable to those of multiple regression equations. More recently, the CDP++.parser model [7], whose parameters were optimized using an automated algorithm, has shown improved performance [2]. At a minimum, this suggests that CDP++.parser does better than a regression equation. However, since regression equations may overfit the data, a simple comparison between model and regression performance may actually underestimate the true predictive power of CDP. This issue is explored in more detail below. The strength of correlations between models and data can also be used to compare different computational models. When such comparisons have been conducted, CDP (in its English version) has consistently outperformed alternative models [5]. The Italian version of CDP also outperformed a competing model [8], although the difference was smaller.

A model with high quantitative performance offers a valuable foundation for investigating how individual differences and atypical cognitive profiles modulate learning outcomes. For example, using CDP, we modelled the individual learning trajectories of 622 children, some with dyslexia and some without [20]. We found that single deficit models were able to explain the overall mean error pattern in both dyslexic and typical readers. However, only multi-deficit models were able to accurately predict the full distribution of errors across individuals. This distinction is crucial. If CDP had poor quantitative performance yet still produced similar shaped distributions, it would raise questions about whether its performance reflected the underlying cognitive mechanisms it was designed to model or whether it was instead driven by idiosyncratic properties of the model itself.

Another important criterion for evaluating models of reading aloud is generalization performance, that is, their ability to predict results from datasets they were not explicitly trained or optimized on. In earlier versions of CDP, we used a single set of model parameters to examine generalization performance across all benchmark effects. The latter included both small experiments targeting specific psycholinguistic factors and large-scale datasets capturing broader patterns of variability across items. In the latest version of our model [2], we took this further by optimizing CDP on several large datasets and then evaluating each optimized version on independent datasets. The results showed that CDP could accurately predict outcomes in held-out datasets, thus demonstrating strong generalization performance. It also outperformed regression models whose regression coefficients were fit to one dataset before being tested on others. These findings indicate that CDP's predictive success is not merely the result of parameter tuning or overfitting. Rather, it stems from the model's underlying architecture, which incorporates cognitively and neuropsychologically motivated representations and processes - features that are absent in standard regression models.

An aspect of generalization performance that has not yet been examined is how CDP's quantitative performance would change if it were optimized on small-scale experiments, specifically, those using 300 stimuli or fewer, as is the case for the experiments analyzed below. This question is particularly important because CDP has many parameters. It is therefore not clear whether it could still generalize well when optimized on limited datasets or whether such

optimization would result in overfitting. This will be tested below by optimizing CDP on well-known small-scale experiments and then examining the extent to which CDP can generalize to independent, held-out datasets. Poor generalization in this context would suggest that the model is overfitting, thereby compromising its predictions and the cognitive interpretability of parameter values. For comparison, we will also examine CDP's generalization performance on these small-scale experiments when the model is instead optimized on large datasets (ranging from 572 to 6714 stimuli in the present study).

Understanding the role of parameters in CDP

Apart from CDP's quantitative performance, another important strength is its interpretability, specifically, the ability to determine which parameters and representations contribute meaningfully to its quantitative performance. Despite CDP's relatively large number of parameters, we have shown that only a small subset significantly influences its quantitative performance when optimized on large datasets [2]. Indeed, most of the parameters had little to no effect on the model when varied within wide bounds. This suggests that, while CDP may appear complex, its functional behavior is driven by a tractable set of parameters.

The technique we used to determine which parameters were important in CDP is known as Sloppy Parameter Analyses (SPA) [21–24]. In SPA, the model is first optimized to fit a dataset as closely as possible. This initial optimization not only provides the best-fitting parameter set but also serves as a reference point for understanding the model's sensitivity to parameter changes. Interestingly, optimization itself offers insight into parameter importance. When the model is optimized multiple times on the same dataset, the resulting values for certain parameters may vary widely. This variability can have two distinct interpretations: (1) the parameter has little effect on model performance, so its exact value is not critical, or (2) the parameter is part of a degenerate set, where trade-offs with other parameters produce equivalent outputs.

Following optimization, SPA proceeds by introducing small deviations – either individually or in combination, from the optimized values. When a small change in a parameter causes a large drop in model performance (i.e., decreasing fit to the data), the parameter is classified as *stiff*. Alternatively, if the same change has little effect, the parameter is called *sloppy* because it can take many values without much effect on performance. Each parameter can thus be characterized by an index that quantifies its degree of stiffness or sloppiness and parameters with equal index values contribute equally to changes in model performance when perturbed by proportionately the same amount.

The SPA is performed around the optimum because the latter represents the best fit of the model to the real data. Evaluating parameter effects around a suboptimal model would reduce the validity of the results, as the derived parameter sensitivities would reflect a model less consistent with empirical data. A useful analogy is filling a container with putty to infer its internal shape. In this case, the putty would be the model. If the putty fills the container perfectly, the extracted shape provides an accurate representation -- an optimal fit. But if the putty is poorly applied, the resulting shape will be distorted and less informative. Similarly, analyzing a well-optimized model yields more valid and interpretable measurements of parameter influence than analyzing one that fits the data poorly.

The present study

Given that CDP performs well when optimized and tested on large datasets, it is important to examine its performance on small-scale experiments and compare it to other metrics, including multiple regression. This is particularly relevant because most of what we know about reading comes from small-scale, factorial experiments with a limited number of highly controlled items and a small number of participants. In our previous modelling work, we referred to the results of such experiments (when coming from influential and well-replicated studies) as "benchmark effects". While these experiments are carefully designed to isolate specific variables, some of these effects might be artificial or idiosyncratic due to the deliberate orthogonalization of psycholinguistic dimensions that covary in large datasets. A classic example is the consistency by frequency interaction [12,14]. Because small experiments often employ simple ANOVA designs and

compare a small number of variables that are deliberately chosen to be either high or low on a given dimension, they may not reflect the statistical properties of naturally occurring word distributions.

In this study, we test whether parameters that were optimized on small benchmark datasets can generalize to other benchmark datasets. Doing so allows us to understand which benchmark effects reflect core processing mechanisms of the reading system and which reflect more peripheral or task-specific factors with limited generalizability.

Another debate that modelling small datasets could inform is the impact of list difficulty and list composition on reading behavior [25–27]. It has long been argued that the difficulty of the stimulus set plays an important role in how people respond [25,26,28]. The explanations offered are typically verbal or qualitative, and results across conditions are often interpreted post hoc (but see [27]). When computational modelling has been used to address this issue, it is typically through the change of a single parameter chosen based on the modeller's intuition [6]. With the methods described here, rather than changes being made based on the modeler's intuition, differences between parameters that are found when CDP is optimized on different stimulus sets can be used. Crucially, these differences are not imposed by the modeler but emerge from the structure of the data and the model's learning mechanisms. As such, they provide explicit, testable predictions about how list composition may influence reading strategies—grounded in the statistical properties of the stimuli and the architecture of the model, rather than the beliefs of the experimenter.

If CDP can be meaningfully optimized on small datasets, it may also have practical value, particularly for modelling individual reading profiles based on datasets containing relatively few words [29]. This would open the door to testing whether personalized remediation programs, informed by CDP's predictions, are more effective than programs based solely on traditional reading assessments [30]. Because CDP makes explicit predictions about which parameters affect the reading performance of individuals, it could help identify the underlying difficulties faced by a given individual. For example, if a parameter that affects the contribution of sublexical phonology has little impact on the performance of a particular child, this might suggest that this child is underutilizing sublexical processing (i.e., decoding), potentially pointing to a suboptimal reading strategy. Such a result could be validated further and compared using targeted psycholinguistic tests. In this case, poor sublexical processing might be caused by phonological dyslexia, and the model could thus provide converging evidence for this diagnosis. Similarly, if a struggling reader shows unusually high sensitivity to parameters related to word frequency, it might suggest that the child has difficulty learning and updating word representations. Alternatively, it might reflect limited exposure to low-frequency words, perhaps due to reading environments dominated by informal sources like social media rather than richer linguistic input from fiction or educational texts. In this way, CDP modelling could help identify the cognitive source of reading difficulties and predict which types of remediation might be most effective for a given individual. Determining whether small dataset modelling provides reliable results is therefore a crucial first step toward such individualized applications, especially in contexts where large-scale generalization testing is not possible or meaningful.

Given the importance of understanding CDP's performance on small-scale experiments, we examined its performance on a number of freely available classic datasets in which the item results had been published. We chose only experiments where the entire stimulus set was available. For each experiment, we optimized a version of CDP specifically on that dataset. We then tested each optimized model on all other experiments (which are therefore treated as held-out data). The selected experiments included ones that examined only words, words and nonwords together, and only nonwords. Using these three different groups is important because there is substantial literature suggesting that stimulus composition influences reading strategies, and CDP predicts that such differences can be captured through changes in parameter values.

To contextualize CDP's performance, we compared the results of CDP to that of a standard multiple regression model predicting reading performance from the psycholinguistic features of the stimuli. Like for CDP, we fitted a regression equation to each dataset and then tested its performance on the remaining datasets, using the coefficients derived from the fitted dataset. This approach allowed us to directly compare CDP and regression not only in terms of raw predictive power but also in their susceptibility to overfitting when generalized to new datasets [2].

Methods

The method was the same as documented in Perry et al. [2]. In particular, for each dataset, we optimized 27 parameters of CDP that were each constrained within a broad range of values (see Table 1). Each optimization run was carried out using the particle swarm algorithm, with code adapted from Kennedy and Eberhart [31]. This method was selected based on our previous work [2], which showed its efficiency and reliability for optimizing CDP. The default parameters included in the code were used, apart from the number of particles which was set to 22. Each time the algorithm was used, it was run for 26 cycles, resulting in the generation of 572 parameter sets, and only the parameter set which had the best fit to the dataset was used.

The cost function used for optimization was a Mean-Squared-Error (MSE) score between the model's predictions (in CDP cycles) and the human reaction times (RTs), with the human RTs scaled down by a factor of 6.5. The scaling constant was empirically chosen to bring the human RTs (in milliseconds) into the same numerical range as CDP's response latencies (in number of cycles) across datasets. Although this scaling factor might be considered a parameter, it is not a parameter of CDP itself and it serves a purely practical purpose, allowing CDP to produce results within the output range used in our prior studies. Moreover, keeping this value fixed avoided significant computational burdens and prevented inconsistent scaling across simulations. This would have allowed identical RTs in two datasets to be interpreted differently, undermining cross-experiment comparisons.

When CDP failed to produce the same response as human readers (i.e., a model error), a small penalty was added to the cost function. This error term was used when CDP generated an incorrect string of phonemes for a word or an unlikely string of phonemes for a nonword. Importantly, the cost function was designed to model mean item-level results—a common approach in computational modeling—rather than simultaneously modelling both reaction times and error rates. As a result, this evaluation does not capture potential speed-accuracy trade-offs at the individual level. For example, if a particular word in an experiment had a 5% error rate, and CDP produced the correct phonemes for the word, no error penalty would be given. Alternatively, if CDP produced a different set of phonemes than a word has even if it could have potentially been given by one of the participants (such as saying *pint* to rhyme with *mint*), an error penalty was still given. The actual equation used was:

$$(1)\quad C_{\theta_i} = \frac{1}{\text{Correct}_{\theta_i}} \sum\nolimits_{1..j} (o_{\theta_i j}/6.5 - e_{\theta_i j})^2 + 50 \times \text{NE}_{\theta_i} + 50 \times \text{NSE}_{\theta_i}$$

where C = is the overall cost (score), Correct is the number of words correctly produced by the model, o is the observed (i.e., actual) RT, e is the value the model produced (the expected value – i.e., the value the model produced for word j), NE = is the number of errors, NSE = number of stress errors, θ_i represents the model parameters $\theta_{i...N}$, where N in our case was 27 (i.e., all of the parameters investigated), and j represents the number of words the model was optimized on.

Optimization (as described above) was repeated 100 times per dataset. This yielded 100 independently optimized parameter sets for each dataset (i.e., the best parameter set from the 572 generated from each of the 100 repeats of each data set was chosen) and hence 100 estimates for each parameter within each dataset. This was done for two main reasons. First, repeated optimization increases the chance of finding model parameters as close to the optimum as possible (due to the stochastic nature of the optimization algorithm). Second, it allows us to examine the distribution of parameter values across runs. In this case, if a parameter has no meaningful effect on the performance of a model, it cannot be optimized, as there is no value it can take to bring the model closer to the optimum. Thus, its value after optimization will be essentially random. Similarly, even if a parameter has a very small effect on the model, it may still be difficult to optimize. This is because its effect on the error function may be very limited, and in complex models, the optimization procedure can only run for a finite amount of time. Thus, it may not get perfectly optimized. This will mean that if the same model is optimized many times, the parameter will take many different values. Alternatively, if a parameter has a strong effect on

Table 1. Parameters that the models were optimized on and the highest and lowest bounds which they could take (reproduced with modifications from Perry et al. [2], with permission).

	Acronym used	Lowest Level	Highest Level	Location	Notes
Feature to letter excitation	Feat_Let_Ex_ff	0.002	0.01	Lexical route	
Feature to letter inhibition	Feat_Let_In_ff	-2	-0.5	Lexical route	
Letters lateral inhibition	Let_LatIn	-1	0	Lexical route	
Letter to orthographic lexicon excitation	Let_OL_Ex_ff	0.01	0.1	Lexical route	
Letter to orthographic lexicon inhibition	Let_OL_In_ff	-2.5	-0.1	Lexical route	
Orthographic lexicon lateral inhibition	OL_LatIn	-0.2	0	Lexical route	
Orthographic lexicon to phonological lexicon excitation	OL_PL_Ex_ff	0.5	2.5	Lexical route	
Orthographic lexicon to letter excitation	OL_Let_Ex_fb	0	0.1	Lexical route	
Orthographic lexicon to letter inhibition	OL_Let_In_fb	-0.1	0	Lexical route	
Phonological lexicon lateral inhibition	PL_LatIn	-0.2	0	Lexical route	
Phonological lexicon to phoneme excitation	PL_Phn_Ex_ff	0.05	0.2	Lexical route	
Phonological lexicon to phoneme inhibition	PL_Phn_In_ff	-0.2	-0.02	Lexical route	
Phonological lexicon to orthographic lexicon excitation	PL_OL_Ex_fb	0.1	4	Lexical route	
Phoneme to phoneme lateral inhibition	Phn_Phn_LatIn	-0.2	0	Lexical route	
Phoneme to phonological lexicon excitation	Phn_PL_Ex_fb	0	0.2	Lexical route	
Phoneme to phonological lexicon inhibition	Phn_PL_In_fb	-0.25	-0.05	Lexical route	
Phonological lexicon to stress excitation	PL_St_Ex_ff	0.01	0.15	Lexical route	
Phonological lexicon to stress inhibition	PL_ST_In_ff	-0.2	0	Lexical route	
Stress to stress inhibition	St_LatIn	-0.2	0	Stress Output Buffer	
Minimum stress naming criterion	St_Over	0.01	0.6	Phoneme Output Buffer	This parameter represents the minimum level of activation stress nodes need to reach at the stress output buffer so that a word can be output
TLA excitation parameter	TLA_Ex_ff	0.03	0.12	TLA network	This parameter represents how strong phonology computed by the TLA network activates the phoneme output buffer
Letter to letter scantime	Let_Scan	1	20	Letter Level	This parameter represents how long it takes for the graphemic buffer to process each letter
Global activation rate	Global_Act	0.05	0.3	All representations	This parameter is used to change the slope of the sigmoid function (activation build-up) from input into a node.
Frequency modifier (both lexicons)	FreqMod	0.05	0.3	Lexical Route	These could potentially differ across lexicons, although we treated them as one parameter
Minimum naming criterion	Min_Naming	0.15	0.7	Phoneme Output Buffer	This parameter represents the minimum level of activation a phoneme node needs to reach so that it can be included in the phonemes that are output by the model

(Continued)

Table 1. (Continued)

	Acronym used	Lowest Level	Highest Level	Location	Notes
Grapheme parsing letter threshold	Let_Over	0.05	0.3	Letter Level	The level of activation which a letter must be over before graphemic parsing begins
Dead node level	DeadNode	0	20	TLA network	This parameter is used to signal further searching needs to be done with a grapheme that the model thinks should be used but only has a weak connection to any phoneme

Note: ff = Feedforward, fb = feedback, Ex = Excitatory, In = Inhibitory, Feat = Feature, Let = Letter, LatIn = lateral inhibition, OL = Orthographic Lexicon, PL = Phonological Lexicon, Phn = phonemes, St = Stress

https://doi.org/10.1371/journal.pcsy.0000074.t001

the performance of a model, then it is likely to be given a more similar value across multiple optimization runs because its value is important and thus likely to be more affected by the optimization procedure. However, there are cases where a parameter has a strong effect on model performance but still shows variability in its optimized values. This can occur when the effect of a parameter trades off with one or more other parameters (this is known as parameter *degeneracy*). The value of those parameters may thus differ across optimization runs even though individually they all affect the performance of the model.

For SPA to yield valid insights, it is essential that the model being analyzed is close to the true optimum, as discussed above. To make sure this was the case, we examined how the models performed using a number of different methods. There were a number of reasons to suggest that they were well optimized. First, across the 100 best runs of each of the 17 datasets, the average correlation between CDP and the actual data was $r = .50$ with an average standard deviation (SD) of.036. As can be seen from S1a Fig, in most of the experiments, the vast majority of the correlations were extremely similar. This suggests that there was not a lot of variability in terms of how well the models were optimized and thus that the large amount of variability in some of the parameters was unlikely to be due to differences caused by optimization.

Next, we examined the consistency of RT predictions across the top 20 (i.e., best-fitting) optimized models by computing pairwise correlations. This analysis was motivated by the observation that models with similar fit values may nonetheless rely on different parameter configurations, potentially reflecting optimization landing into different regions of the parameters' state-space. To do this, we calculated Spearman's rank correlation coefficients between the RT predictions of all possible combinations of optimization runs. We chose Spearman correlation because it is robust to outliers that the model occasionally produces (though results are very similar using Pearson correlation). The logic of this analysis is that if the top models tend to produce very similar results, both in terms of overall fit (as examined above) and the predictions they make on each item, it would suggest that they converge to a more similar region of the parameters' state-space than models that produced similar fit values but use quite different values for each item. As can be seen from S1b Fig, apart from the pseudohomophone experiment, all of the top 20 runs in each experiment correlated strongly with each other. This suggests that the RT values across the runs was very similar, making it unlikely that multiple high-performing models reflect optimization to substantially different parts of the parameters' space.

We also examined the root-mean-squared-error (RMSE) values calculated across each possible pairing of results in each experiment with the top 20 optimized models in each experiment (see S1c Fig). They also showed relatively small variability. As a comparison, the RMSE was computed from the results of each run of each experiment. The minimum value found across all experiments was 70.33, which was multiple times larger than any RMSE value we calculated between experiments. Thus, the difference between different runs of the same experiment is much smaller than the variability from items within each experiment. Taken together, these results suggest that the top-performing models tend to converge on a relatively similar region of the parameters' state-space, rather than achieving similar fits from divergent areas of the parameters' state space.

After the optimization phase, SPA was performed on the best optimized model from each dataset. The reader is referred to Perry et al. [2] for a full mathematical description of SPA. In that study, the scores were generated using two different error functions (either the log of their parameter values was taken or parameter values were normalized by the size of pre-specified low and high boundaries). Both approaches produced very similar results. Here, for the sake of simplicity, we only present results using the log function. SPA yields a score for each parameter that quantifies its relative influence on model performance. Specifically, it reflects how much a proportionate change in a given parameter affects model error, compared to changes in other parameters. The scores for different parameters are directly comparable with each other, providing a clear indication of which parameters the model's behavior is most sensitive to.

Stimuli

The following datasets were used to examine CDP. They can be divided into three categories across two dimensions. One of the dimensions is the type of stimuli they used: word only, nonword only, or a mix of both (i.e., words and nonwords). The second dimension distinguishes between small experiments designed to test specific psycholinguistic manipulations and large-scale datasets without explicit manipulations. With the large-scale datasets, there was no dataset that included a mix of both words and nonwords.

Word only experiments

- Jared et al. [14]: All 4 experiments (160 words per experiment). These experiments examined the effect of spelling-sound consistency and word frequency.
- Rastle and Coltheart [32]: Both experiments examined word stress. Experiment 1 had 160 words with different types of stress and Experiment 2 had 120 words which examined stress irregularity using high and low frequency words.

Nonword only experiments

- Andrews and Scarrat [15]: There were two experiments in Andrews and Scarrat. However, we only used the data from Experiment 2 because the data from Experiment 1 was not available. Experiment 2 used 128 nonwords with orthographic bodies (i.e., the orthographic equivalent of the rime) that varied on the number of likely pronunciations they could generate.
- McCann and Besner [33]: This experiment examined the pseudohomophone effect, where 72 nonwords that would typically be pronounced like words (e.g., *kat*) were compared with 72 control nonwords that would not typically be pronounced like words.

Mixed word and nonword experiments

- Weekes [34]: This experiment used 100 nonwords and 200 words. Word and nonword length were examined as was the frequency of the words (high vs. low).
- Ziegler et al. [16]: This experiment used 80 words and 80 nonwords. Body neighborhood (i.e., how commonly the orthographic equivalent of a rime occurs) was also manipulated with a high vs. low manipulation.

Word datasets

The data used here was first reported in Perry et al. [35]

- Spieler and Balota (Young) [12]: This is a dataset of 2998 monosyllabic words read aloud by younger participants.

- Spieler and Balota (Old) [12]: This is a dataset of 2998 monosyllabic words read aloud by older participants.
- Chateau and Jared [36]: This is a dataset of 901 disyllabic words, all of which were 6 letters long.
- Seidenberg and Waters [37]. This is a dataset of 1329 monosyllabic words (which will be called the Waters dataset to avoid confusion with the Seidenberg et al. nonwords).
- Treiman et al. [38]. This is a dataset of 1327 monosyllabic words.
- Yap and Balota [39]. This is a dataset of 6714 monosyllabic and disyllabic words.

Nonword datasets

- Seidenberg et al. [40]. This is a database of 572 nonwords.

Results

Optimization results

Standard Deviations (SDs) of each parameter were calculated from 100 independent optimization runs per experiment and are presented in Fig 2. Note that each parameter value was first normalized by the difference between the minimum and maximum values that it could take during optimization (see Table 1). This normalization enables direct comparison of SDs across parameters with different scales. A SD of around .25 or higher indicates random distributions. Although this threshold is somewhat arbitrary, it is close to the SD obtained when random values are generated from a uniform distribution between 0 and 1, which is .29 (note that the possible range for the raw parameters after division by the difference between the maximum and minimum values of each parameter is also one). However, smaller SDs may be found when the parameter values are randomly distributed over only part of the entire possible range. Based on visual inspection of the histograms, we therefore adopted the slightly more conservative threshold of .25 to flag potentially sloppy parameters. As the SD values decrease from around .25, the distributions begin to show more structured, non-random patterns (i.e., similar values are given across different simulation runs). Low SDs thus indicate stiff parameters. For instance, the *Global Activation* parameter tends to be relatively stiff across most datasets, unlike the *Feature-to-Letter Excitation* parameter. The *Minimum Naming Criterion* differs depending on the datasets. Histograms with the actual distributions are provided in S2 Fig.

The distribution of parameter SDs from the small-scale experiments was also very similar to that observed for the large datasets. A Wilcoxon rank sum test examining the SD values across the two groups was not significant, $W = 27328$, $p = .20$ (see S3 Fig for histograms of the SD values in these two groups). This suggest that when CDP is optimized on a small number of items, it yields parameter distributions comparable to those obtained from optimization on large datasets. As illustrated in Fig 2, parameters that have a narrow range of values when optimized multiple times on a large dataset tend to have a narrow range of values when optimized multiple times on a small-scale experiment. Conversely, parameters that have a broad range of values when optimized multiple times on a large dataset tend to have a broad range of values when optimized multiple times on a small-scale experiment.

Results for word experiments

Given that the small-scale experiments produced results that were similar to those from the large datasets, it follows that the parameters that had narrow distributions were also essentially the same. This included the *Global Activation* and *Letter Scanning Time* parameters. *Global Activation* affects the speed at which the activation function rises across all representations. As argued in our previous work [35], this parameter is the most influential in CDP's ability to successfully

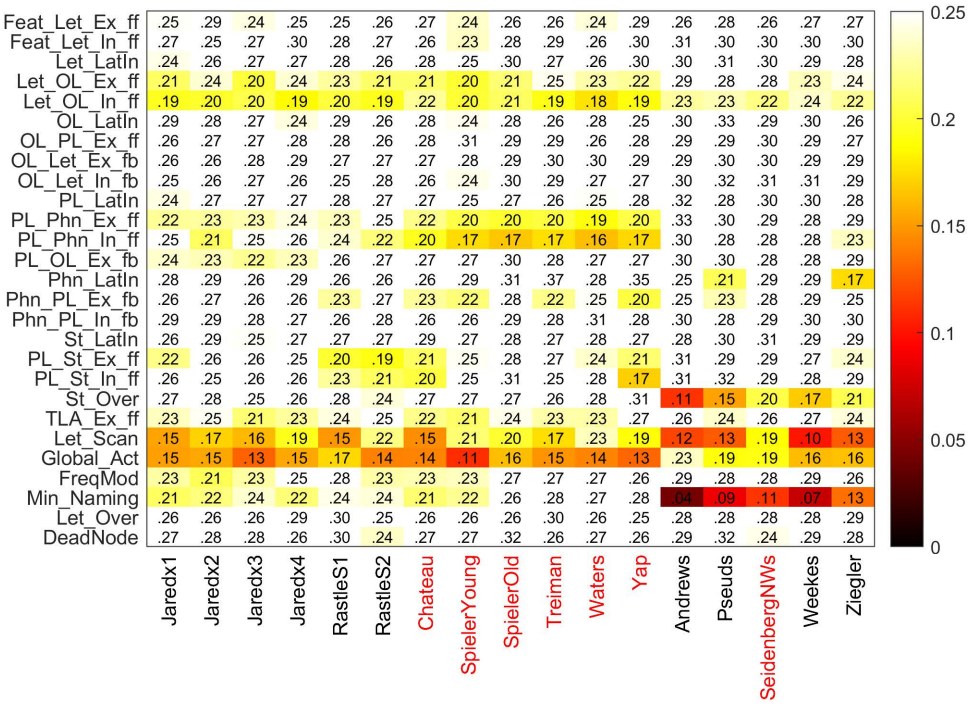

Fig 2. Parameter distributions across experiments. The figure presents normalized standard deviations of the parameter distributions across the different experiments examined. Cells highlighted with darker colors denote lower parameter variability. Anything above .25 is white. *Notes: Experiment names in black are small experiments and those in red are larger datasets. Jaredx1-x4 = Experiments 1- 4 of Jared et al. [14], RastleS1-S2 = Experiments 1 and 2 of Rastle and Coltheart [32], Chateau = Chateau and Jared [36], SpielerYoung and SpielerOld = Spieler and Balota [12], Treiman = Treiman et al. [38], Waters = Seidenberg and Waters [37], Yap = Yap and Balota [39], Andrews = Andrews and Scarratt [15], Pseuds = McCann and Besner [33], SeidenbergNWs = Seidenberg et al. [40], Weekes = Weekes [34], Ziegler = Ziegler et al. [16]. ff = Feedforward, fb = feedback, Ex = Excitatory, In = Inhibitory, Feat = Feature, Let = Letter, LatIn = lateral inhibition, OL = Orthographic Lexicon, PL = Phonological Lexicon, Phn = Phoneme, St = Stress, Stress_Over = stress over criterion, Let_Scan = letter scanning time, Act = activation, FreqMod = Frequency modifier, Min_Naming = Minimum Naming criterion, Let_Over = Letter Over criterion.

https://doi.org/10.1371/journal.pcsy.0000074.g002

simulate human RTs. *Letter Scanning Time* affects how quickly each letter is processed before being inserted in the graphemic buffer, and it also consistently showed stiff behavior across datasets. There were also a number of lexical parameters that had broader distributions but whose values were not random. For example, the Inhibition from the Letter Level to the Orthographic Lexicon needs to be strong enough to stop incorrect word forms from being activated. As long as the inhibition is strong enough, the specific value has little effect on the model. As a result, this parameter's SDs were moderate across all experiments (between .19 and .24), suggesting that while it is not tightly constrained, its values fall within a functional range. This can be seen from S2 Fig. Across all experiments and datasets, values near zero (i.e., no inhibition) are absent, and the distributions skew toward more negative values, indicating consistent inhibitory effects. Importantly, broader parameter distributions do not necessarily reflect lack of importance: in some cases, they might simply reflect some degree of degeneracy (i.e., parameters which trade-off with each other to produce the same result). The SPA results below show how these can be identified.

Results from nonword experiments

Since there was only one large dataset of nonwords used (the Seidenberg Nonwords), we need to be careful comparing it to the two small-scale experiments. Nonetheless, the results were broadly consistent across the three datasets. One notable difference was that the Letter Scanning Time was somewhat less constrained in the Seidenberg dataset (SD = .19)

compared to the Andrews (SD = 0.12) and pseudohomophone (SD = 0.13) experiments. The Minimum Naming Criterion was also particularly constrained in the Andrews experiment, with an SD of just 0.04. Consistent with prior findings from Perry et al. [35], parameters in the lexical route were generally weakly constrained with only two parameter distributions having an SD below .25. These were both found with the pseudohomophone data, with the Lateral Inhibition parameter of the phonological lexicon and Phoneme to Phonological Lexicon Excitation parameters having SDs of .23 and .21, respectively. This is not so surprising because for CDP to produce a pseudohomophone effect, activation needs to feedback from the phonemes. Therefore, when optimized on data with a strong pseudohomophone manipulation, these parameters may affect CDP more than when CDP is optimized on nonwords without such a property. Finally, the Minimum Naming Criterion was quite tightly constrained around lower values as can be seen from the histograms (with SDs of .04, .05, and .11, for the three nonword experiments), unlike with the word only data where no experiment produced a distribution with a SD of less than .21. This means CDP predicts that when only nonword stimuli are used, people will tend to set a low minimum naming criterion (threshold) before responding compared to when only words are used.

Results from the mixed (word and nonword) experiments

The results from the mixed experiments closely resembled those from the nonword-only experiments, particularly in terms of the Minimum Naming Criterion, which consistently was set to low values (See S2 Fig). This pattern contrasts with the word-only datasets, where the criterion was generally higher. These findings offer insights into how heterogeneous lists are processed. Specifically, it suggests that inclusion of nonwords leads to a lowering of the response threshold for initiating pronunciation. Thus, the parameter values do not appear to reflect an averaging of stimulus properties across the list. Instead, the presence of more difficult, slowly activated stimuli, such as nonwords, drives the system to adopt a more permissive criterion. This is because if the Minimum Naming Criterion is set too high, it can cause nonword errors because phonemes might not reach the necessary threshold to enter the final sequence of phonemes that is generated by the model. If parameters were chosen by hand rather than by optimization, increasing the Global Activation parameter could potentially allow phonemes to reach the criterion more quickly, allowing more nonwords to reach the criterion. However, the Global Activation parameter was remarkably similar across experiments, suggesting that the model does not rely on global activation adjustments to accommodate the slower activation of nonword stimuli.

Sloppy parameter analyses

To examine which parameters had the greatest impact on CDP's performance, we conducted SPA following the procedure outlined in Perry et al. [35]. A key component of this analysis is the computation of a Hessian Matrix (i.e., a square matrix of second-order partial derivatives), which captures the curvature of the cost function around its optimum. It is computed using finite difference scores, where differences between cost function scores at the optimum are compared to cost function scores after small perturbations to individual parameters away from the optimum. This estimates the extent to which parameters and combinations of them cause differences in the cost function – that is, which parameters and combinations are responsible for the quantitative behavior of the model.

Based on the Hessian Matrix, a set of eigenvalues and eigenvectors can be derived for each experiment through matrix decomposition. The strength of an eigenvalue reflects how sharply the cost function changes when moving in the direction defined by its associated eigenvector. This is conceptually similar to principal components analysis, where eigenvalues indicate the strength of principal components and eigenvectors their direction. As in principal component analysis, the direction of the eigenvectors typically do not align with the axes defined by individual parameters. Thus, a movement in space that is perfectly aligned with an eigenvector usually requires adjustments to several parameters simultaneously.

To assess how "sloppy" or "stiff" a specific CDP parameter is, we examine how closely the axis corresponding to that parameter aligns with the eigenvectors of the Hessian When dealing with single parameters, if the parameter axis aligns closely with an eigenvector associated with a large eigenvalue, the parameter will be stiff, meaning that small changes in

the parameter lead to large changes in the cost function. In contrast, if a parameter axis aligns with an eigenvector with a small eigenvalue, the parameter will be sloppy, having relatively little impact on the cost function. However, it is important to note that the stiff/sloppy distinction is continuous rather than binary. Some parameters may not strongly align with any single eigenvector but instead align more moderately with one or more strong eigenvectors. In such cases, the overall influence of the parameter may not be either highly stiff nor fully sloppy.

To calculate the extent to which a parameter is aligned with different eigenvectors whilst taking into account their strength, we used the parameter ranking function proposed by Kardynska et al. [24]. This function provides an index where higher values indicate greater stiffness. Further details and a more detailed description of this procedure can be found in Perry et al. [35].

Fig 3 presents the eigenvalues obtained from the Hessian decomposition across all of the experiments. Each of the blue bars represents the strength of an eigenvalue from a given experiment. For example, in the Seidenberg Nonwords dataset, seven eigenvalues were extracted with values ranging from over 10000 (10^5) to less than one. Due to the use of a logarithmic scale, this wide range appears approximately linear in the graph. This log-linear distribution can be seen across all experiments. This means that despite having many parameters, the effective dimensionality of the model's parameter space is relatively low. That is, most of the variance in model behavior is governed by a small number of stiff dimensions, with the remaining dimensions contributing very little. Such a pattern is consistent with findings from other complex systems [22,23,41–44]. This low-dimensional structure makes the analysis of CDP tractable because there are only a small number of dimensions that need to be investigated to understand most of its behavior. By contrast, if a large number of eigenvalues contributed more evenly to performance, the blue bars would be close to each other even with

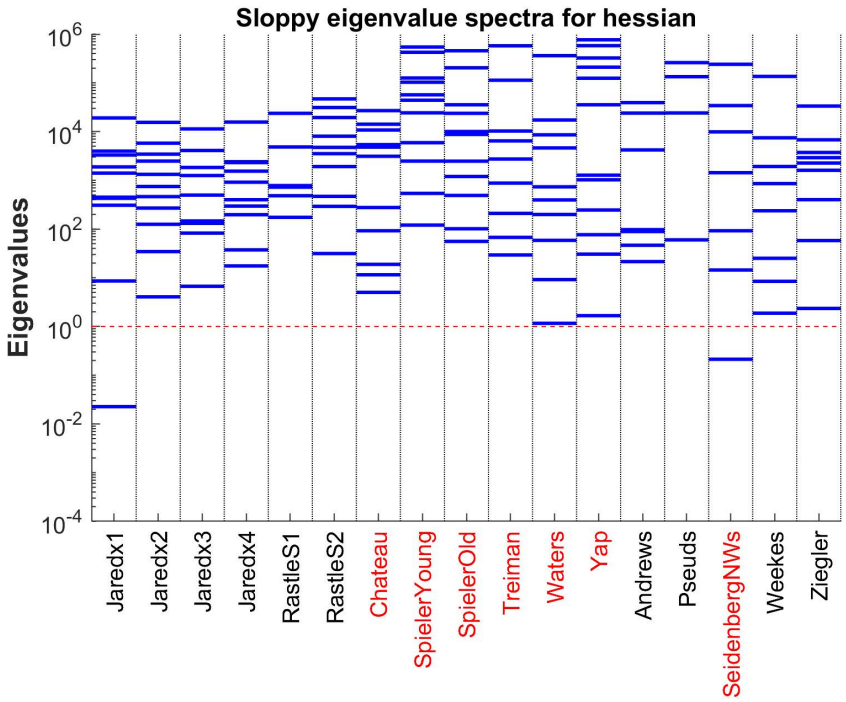

Fig 3. Eigenvalues found in the different experiments. *Notes: The Y axis has log scaling. Jaredx1-x4 = Experiments 1- 4 of Jared et al. [14], RastleS1-S2 = Experiments 1 and 2 of Rastle and Coltheart [32], Chateau = Chateau and Jared [36], SpielerYoung and SpielerOld = Spieler and Balota [12], Treiman = Treiman et al. [38], Waters = Seidenberg and Waters [37], Yap = Yap and Balota [39], Andrews = Andrews and Scarratt [15], Pseuds = McCann and Besner [33], SeidenbergNWs = Seidenberg et al. [40], Weekes = Weekes [34], Ziegler = Ziegler et al. [16].

https://doi.org/10.1371/journal.pcsy.0000074.g003

simple linear scaling. This would mean that many dimensions of the parameters' space would need to be investigated to understand how the model works. This pattern was not observed.

One difference between the small experiments and large datasets was in the magnitude of the highest eigenvalues. Large datasets generally produced higher maximum eigenvalues. The only exception was in the Chateau dataset which had the fewest items among the large-scale word datasets. This trend led to a significant difference between the size of the log-transformed eigenvalues between small and large datasets, as confirmed by a Wilcoxon rank sum test ($W = 7240$, $p = .017$). Histograms illustrating this difference can be seen from S4 Fig.

Fig 4 presents the alignment between individual model parameters and the eigenvectors of the Hessian. The results show that the number of parameters that contribute to model performance (i.e., stiff parameters) was similar across small-scale experiments and the large datasets. This can be seen from S5 Fig, which shows histograms of parameter ranking values grouped by dataset size. A Wilcoxon rank sum test examining the differences between these groups was not significant ($W = 25023$, $p = .72$).

These results are important because they suggest that CDP does not overfit the data when optimized on small numbers of items. If overfitting were occurring, we would expect to see aberrant or unstable patterns to emerge in the SPA, leading to systematic differences between the small-scale experiments and the large datasets. However, no such differences were observed, indicating that even with limited data, the optimization process yields stable and interpretable parameter structures.

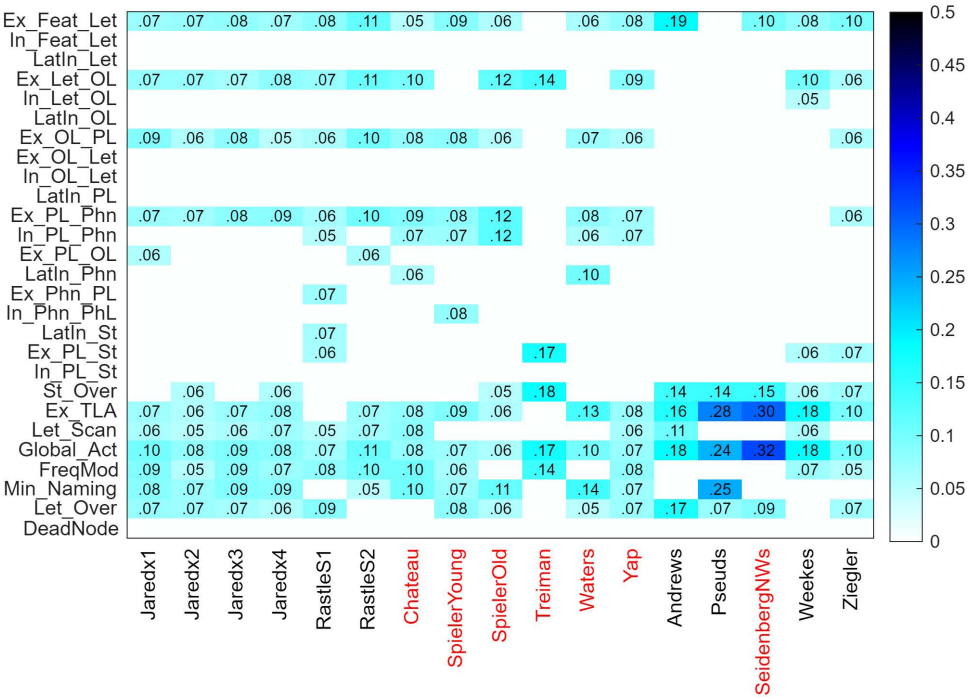

Fig 4. Parameter rankings across experiments. *Notes: The higher the number, the stiffer the parameter is. Experiment names in black are small experiments and those in red are larger datasets. Values less than 0.05 are not represented and are left as blank cells. Jaredx1-x4 = Experiments 1- 4 of Jared et al. [14], RastleS1-S2 = Experiments 1 and 2 of Rastle and Coltheart [32], Chateau = Chateau and Jared [36], SpielerYoung and Spieler-Old = Spieler and Balota [12], Treiman = Treiman et al. [38], Waters = Seidenberg and Waters [37], Yap = Yap and Balota [39], Andrews = Andrews and Scarratt [15], Pseuds = McCann and Besner [33], SeidenbergNWs = Seidenberg et al. [40], Weekes = Weekes [34], Ziegler = Ziegler et al. [16]. ff = Feed-forward, fb = feedback, Ex = Excitatory, In = Inhibitory, Feat = Feature, Let = Letter, LatIn = lateral inhibition, OL = Orthographic Lexicon, PL = Phonological Lexicon, Phn = Phoneme, St = Stress, Stress_Over = stress over criterion, Let_Scan = letter scanning time, Act = activation, FreqMod = Frequency modifier, Min_Naming = Minimum Naming criterion, Let_Over = Letter Over criterion.

https://doi.org/10.1371/journal.pcsy.0000074.g004

Parameter stiffness in word experiments

The parameter rankings were highly similar across the large-scale word datasets and the small-scale word experiments. This reinforces the conclusion that the model was not overfitting when optimized on smaller item sets. Several parameters affected the processing of CDP according to SPA, yet still exhibited large variability in their estimated values, with a quasi-random distribution within their possible bounds (Feature-letter excitation, Orthographic Lexicon to Phonological lexicon excitation, Letter Over). This suggests that there is some degree of degeneracy with those parameters because they affected model performance but were not strongly constrained.

Parameter stiffness in nonword experiments

In the nonword experiments, the TLA Excitation Feedforward parameter (*TLA_Ex_ff*) was consistently stiff. This parameter determines how strongly the sublexical route activates phonemes in the phoneme output buffer, which combines input from the lexical route and sublexical route before producing a spoken output. For nonwords, this route typically provides the majority of the phonological information, so stiffness in this parameter reflects its critical role in nonword processing. The global activation parameter was also stiff, likely serving to align the overall model RTs with observed RTs across the experiments. None of the lexical parameters after the feature level meaningfully influenced the performance of the model.

One interesting result was that the Minimum Naming Criterion (i.e., the level of activation nonwords must reach before they can be read aloud) was stiff with the pseudohomophones but not with the other two nonword experiments. This asymmetry may be explained by how errors affect the cost function. Because errors are incorporated into the cost function using a constant term (see Equation 1 above), even a small number of errors created from movements in parameters can have a disproportionately large effect on the cost function. This effect is more likely with nonwords than with words, simply because the model is more error-prone on nonword inputs, providing more opportunities for such threshold effects to occur. Thus, this pattern may reflect this occurring in the pseudohomophone results but not the other two nonword experiments.

Parameter stiffness in mixed (word and nonword) experiments

The mixed experiments showed intermediate patterns of parameter stiffness. The results from the Weekes dataset more closely resembled the nonword-only experiments, while the results from the Ziegler et al. dataset aligned slightly more with the word-only experiments. However, in both datasets, parameter stiffness generally fell between the word and nonword datasets.

Generalization performance across datasets

Given that the models optimized on small experiments did not produce aberrant parameter patterns compared to the models optimized on the large datasets, it seems reasonable to test their generalization performance on unseen data. This was done by using the parameters from the best-optimized model in each experiment to predict all other datasets. This approach allows us to assess the predictive performance of the models and the extent to which meaningful parameters values can be estimated from relatively small stimulus sets. As a comparison, we used regression equations in a similar manner. Although more complex methods exist (e.g., [45]), we chose regression because that is the standard method for comparing cognitive models to data in this research area (e.g., [5,10,12] and it is commonly used to examine the influence of psycholinguistic features of the stimuli in larger data sets (e.g., [46,47]).

For the regression models, we used predictors that were relevant for the experiment based on the list composition. With word-only experiments, predictors included log frequency, spelling-sound consistency, letter length, and orthographic neighborhood (henceforth orthographic N). We also tested models with a frequency by consistency interaction term, as this effect is well-established in reading aloud studies [14], but this did not change the results in any meaningful way (for

the results, see S1 Table). For nonword-only experiments, we used letter length and orthographic N. For the mixed experiments, we used log frequency, letter length, orthographic N, a dummy variable coding word vs. nonword status and the interaction term between letter length and lexicality. In these cases, we also used coefficients derived from mixed datasets to predict word-only and nonword-only datasets, applying the relevant subset of predictors in each case (i.e., for word only experiments, frequency, spelling-sound consistency, letter length, and orthographic N; and for nonword only experiments, letter length and orthographic N).

Overall, the results showed that CDP, even when optimized on small stimulus sets, exhibits good generalization performance both in terms of RTs (see Table 2) and error rates (see Table 3), typically better than regression (see Table 4). However, there were a few notable exceptions. When the model was only optimized on small word-only experiments (i.e., not the large datasets or mixed experiments), it often performed poorly when predicting the nonword data. In contrast, CDP models trained on large-scale word-only datasets produced low error rates on the nonwords. This suggests that there may not be enough information in the small experiments so that parameters that are associated with accurate nonword processing can be reliably found. A speculative reason for this is that the words in the large datasets span a broader range of psycholinguistic dimensions, whereas the small experiments typically manipulate variables orthogonally, using items with extreme values. As a result, the large datasets may simply have more information in them. The optimization would also be less biased on the large datasets than in the small experiments where even a factor of minor importance would inflate the cost scores. Furthermore, the lexical processes engaged during word reading may obscure sublexical processes that are more critical for nonword pronunciation, making it more difficult for the model to generalize to nonwords without dedicated nonword training data. In summary, the limited generalization from small-scale word experiments to nonword data may be explained by gaps and/or biases in the training data used for model optimization.

Apart from these discrepancies in the error rates on nonwords, the only other case where generalization seemed to be poor was in the first experiment of Rastle and Coltheart [32], which involved disyllabic words and stress manipulation. This contrasts with most other datasets, which used only monosyllables. The poor generalization to this dataset might be attributed to the word stress parameters being difficult to optimize without disyllabic stimuli. However, since the models were able to perform well on Rastle and Coltheart's second experiment that also used disyllabic words, this can only be a partial explanation.

To further explore generalization, we examined prediction accuracy based on whether a model was optimized on a small-scale experiment or a large-scale dataset. Specifically, we calculated average *r* values for cross-predictions among word-only datasets across four combinations: Small predicts small (CDP: .55, Regression: .52) small predicts large (CDP: .44, Regression: .34), large predicts small (CDP: .57, Regression: .43) and large predicts large (CDP: .44, Regression: .36). In all cases, CDP outperformed regression, reinforcing the conclusion that CDP generalizes well and does not overfit—even when optimized on small item sets designed for specific factorial manipulations.

The results from the two mixed experiments showed that CDP was able to predict the data with a similar efficacy as the regression equations in terms of both the words it was optimized on and generalizing to other mixed sets. The results from CDP when optimized on mixed sets also predicted the other experiments well. This was not the case for the regression models which generalized extremely poorly on all of the non-mixed experiments apart from the Chateau dataset, often producing negative *r* values. This suggests the coefficients found with the regression equation, whilst fitting the dataset they were optimized on well, were not generalizing to other datasets. This is important because it shows that in some circumstances, the fit to the data of the regression equations based on psycholinguistic parameters may not necessarily be readily interpretable in terms of the way people read. Thus, CDP clearly makes a difference in predictions because its parameters are linked to processes/components.

Performance on nonwords was more mixed. CDP was poor at predicting the pseudohomophone data when not optimized on pseudohomophones. When CDP was optimized on pseudohomophones, it also tended to be poor at predicting other data. This suggests that datasets with large numbers of pseudohomophones may induce strategies which are at

Table 2. R values from correlations between CDP and the reaction times of experiments when CDP is optimized on different datasets.

		Experiment Type																
		Words (Small Experiments)						Nonwords			Mixed		Words (Big datasets)					
Dataset model optimized on	Overall Mean	Jared X1	Jared X2	Jared X3	Jared X4	Rastle Stress 1	Rastle Stress 2	Andrews	Pseuds	Seid NWs	Weekes	Ziegler	Chateau	Spieler Young	Spieler Old	Treiman	Waters	Yap
Jared X1	.50	**.65**	.71	.71	.67	.34	.61	.23	.02	.28	.69	.78	.49	.46	.50	.41	.31	.60
Jared X2	.48	.64	**.71**	.68	.63	.31	.59	.08	.17	.28	.60	.72	.49	.44	.49	.41	.28	.61
Jared X3	.48	.54	.66	**.74**	.70	.19	.51	.24	.09	.29	.69	.76	.43	.46	.49	.41	.33	.58
Jared X4	.44	.59	.65	.74	**.71**	.03	.36	.15	.09	.32	.67	.75	.26	.44	.49	.41	.30	.52
Rastle Stress 1	.48	.52	.59	.61	.60	**.49**	.58	.29	.14	.37	.70	.73	.44	.44	.43	.35	.33	.55
Rastle Stress 2	.49	.56	.59	.68	.64	.42	**.70**	.25	.03	.32	.68	.72	.52	.42	.48	.40	.27	.61
Andrews	.46	.55	.58	.66	.59	.26	.44	**.42**	.10	.37	.69	.74	.37	.43	.48	0.35	.21	.56
Pseuds	.29	.35	.42	.42	.48	.04	.25	.12	**.29**	.07	.49	.40	.29	.28	.29	.21	.17	.35
Seid NWs	.43	.40	.50	.64	.57	.13	.54	.18	.06	**.42**	.66	.72	.38	.44	.47	.38	.25	.56
Weekes	.47	.44	.53	.63	.56	.37	.61	.37	.07	.38	**.71**	.73	.44	.45	.48	.38	.28	.60
Ziegler	.39	.43	.52	.57	.57	.19	.47	.11	.06	.18	.64	**.78**	.41	.37	.39	.30	.21	.48
Chateau	.49	.57	.63	.68	.66	.39	.65	.22	.09	.27	.67	.70	**.56**	.45	.48	.41	.33	.61
Spieler Young	.41	.55	.63	.68	.67	.14	.39	.20	.06	.27	.54	.43	.36	**.46**	.44	.36	.32	.53
Spieler Old	.50	.55	.61	.70	.66	.41	.68	.18	.16	.25	.67	.72	.55	.45	**.49**	.41	.32	.62
Treiman	.49	.58	.63	.71	.66	.38	.66	.15	.13	.19	.65	.74	.51	.45	.50	**.43**	.30	.62
Waters	.43	.58	.63	.64	.63	.08	0.38	.27	.12	.30	.67	.70	.25	.44	.43	.36	**.35**	.51
Yap	.49	.56	.62	.71	.66	.44	.62	.22	.18	.29	.67	.68	.52	.43	.49	.40	.29	**.62**
N. Words		120	120	120	120	120	120	128	84	588	300	160	901	2998	2998	1327	3129	6714
Max		.65	.71	.74	.71	.49	.70	.42	.29	.42	.71	.78	.56	.46	.50	.43	.35	.62
Min		.35	.42	.42	.48	.03	.25	.08	.02	.07	.49	.40	.25	.28	.29	.21	.17	.35
Median		.55	.62	.68	.64	.31	.58	.22	.09	.29	.67	.72	.44	.44	.48	.40	.30	.58
Mean		.53	.60	.66	.63	.27	.53	.22	.11	.28	.65	.69	.43	.43	.46	.38	.29	.56

Note: Note: Jared X1-X4 = Experiments 1–4 of Jared [14], Rastle Stress 1 and 2 = Experiments 1 and 2 of Rastle and Coltheart [32], Andrews = Andrews and Scarratt [15], Pseuds = McCann and Besner [33], Seid NWs = Seidenberg et al. [40], Weekes = Weekes [34], Ziegler = Ziegler et al. [16], Chateau = Chateau and Jared [36], Spieler Young and Old = Experiments 1 and 2 from Spieler and Balota [12], Treiman = Treiman et al. [38], Waters = Seidenberg and Waters [37], Yap = Yap and Balota [39].

https://doi.org/10.1371/journal.pcsy.0000074.t002

Table 3. Error rates (%) of CDP when CDP is optimized on different datasets.

		Experiment Type																
		Words						Nonwords			Mixed		Words (Big datasets)					
Dataset CDP optimized on	Overall Mean	Jared X1	Jared X2	Jared X3	Jared X4	Rastle Stress 1	Rastle Stress 2	Andrews	Pseuds	Seid NWs	Weekes	Ziegler	Chateau	Spieler Young	Spieler Old	Treiman	Waters	Yap
Jared X1	3.13	**0.00**	1.67	0.00	0.00	0.83	1.67	7.03	15.48	11.56	3.67	4.38	1.44	0.50	0.50	0.83	0.53	0.49
Jared X2	9.69	0.00	**0.83**	0.00	0.00	3.33	2.50	32.81	39.29	38.95	12.67	19.38	2.55	0.63	0.63	0.98	0.53	1.22
Jared X3	6.26	0.00	0.83	**0.00**	0.00	5.00	9.17	15.63	25.00	18.88	8.00	10.63	3.11	0.93	0.93	1.21	0.90	1.97
Jared X4	13.65	0.00	0.83	0.00	**0.00**	3.33	10.00	56.25	61.90	43.03	15.33	21.25	2.89	0.77	0.77	1.36	0.68	1.52
Rastle Stress 1	1.95	0.00	0.83	0.00	0.00	**0.00**	0.00	5.47	10.71	7.31	2.33	0.00	0.89	0.77	0.77	1.21	0.90	0.51
Rastle Stress 2	4.07	0.00	1.67	0.00	0.00	1.67	**0.00**	10.94	23.81	16.33	3.33	3.75	1.00	0.53	0.53	0.98	0.60	0.37
Andrews	5.27	6.67	9.17	7.50	7.50	4.17	4.17	**3.91**	11.90	4.42	3.67	0.00	6.99	3.14	3.14	4.75	3.24	6.81
Pseuds	7.40	6.67	8.33	6.67	6.67	12.50	25.83	4.69	**9.52**	5.61	4.33	0.00	13.43	3.30	3.30	4.60	2.93	10.19
Seid NWs	2.53	0.83	2.50	0.83	0.83	2.50	1.67	5.47	10.71	**3.91**	2.67	0.00	2.77	1.27	1.27	2.19	1.13	2.52
Weekes	1.59	0.00	0.83	0.00	0.00	0.00	0.00	3.91	11.90	3.91	**1.33**	0.00	0.78	0.53	0.53	1.06	0.60	0.39
Ziegler	2.82	0.00	1.67	0.00	0.00	4.17	4.17	5.47	11.90	6.80	3.00	**0.00**	3.55	0.93	0.93	1.81	0.75	2.92
Chateau	1.93	0.00	1.67	0.00	0.00	0.00	0.00	4.69	11.90	7.82	3.00	0.00	**0.67**	0.30	0.30	0.45	0.15	0.27
Spieler Young	2.16	0.00	1.67	0.00	0.00	0.00	0.00	8.59	13.10	5.61	3.33	0.63	0.67	**0.20**	0.20	0.53	0.08	0.37
Spieler Old	1.95	0.00	1.67	0.00	0.00	0.83	2.50	5.47	9.52	6.80	2.00	0.00	1.66	0.17	**0.17**	0.38	0.08	0.42
Treiman	2.23	0.00	1.67	0.00	0.00	0.83	0.00	5.47	11.90	10.20	3.33	0.00	1.00	0.30	0.30	**0.45**	0.15	0.25
Waters	2.06	0.00	1.67	0.00	0.00	0.83	2.50	4.69	10.71	6.80	2.67	0.00	2.22	0.23	0.23	0.38	**0.08**	0.63
Yap	1.67	0.00	1.67	0.00	0.00	0.00	0.00	4.69	10.71	5.10	3.00	0.00	0.67	0.20	0.20	0.45	0.08	**0.19**
N. Words		120	120	120	120	120	120	128	84	588	300	160	901	2998	2998	1327	1329	6714
Max		6.67	9.17	7.50	7.50	12.50	25.83	56.25	61.90	43.03	15.33	21.25	13.43	3.30	3.30	4.75	3.24	10.19
Min		0.00	0.83	0.00	0.00	0.00	0.00	3.91	9.52	3.91	1.33	0.00	0.67	0.17	0.17	0.38	0.08	0.19
Median		0.00	1.67	0.00	0.00	0.83	1.67	5.47	11.90	6.80	3.33	0.00	1.66	0.53	0.53	0.98	0.60	0.51
Mean		0.83	2.30	0.88	0.88	2.35	3.77	10.89	17.65	11.94	4.57	3.53	2.72	0.87	0.87	1.39	0.79	1.82

Note: Jared X1-X4 = Experiments 1–4 Jared et al. [14], Rastle Stress 1 and 2 = Experiment 1 and 2 of Rastle and Coltheart [32], Andrews = Andrews and Scarratt [15], Pseuds = McCann and Besner [33], Seid NWs = Seidenberg et al. [40], Weekes = Weekes [34], Ziegler = Ziegler et al. [16], Chateau = Chateau and Jared [36], Spieler Young and Old = Experiments 1 and 2 from Spieler and Balota [12], Treiman = Treiman et al. [38], Waters = Seidenberg and Waters [37], Yap = Yap and Balota [39].

https://doi.org/10.1371/journal.pcsy.0000074.t003

Table 4. R values from regression equations fit to one dataset and used to predict the others.

		Experiment Type																
		Words						Nonwords			Mixed		Words (Big datasets)					
Dataset regression coefficients taken from	Overall Mean	Jared X1	Jared X2	Jared X3	Jared X4	Rastle Stress 1	Rastle Stress 2	Andrews	Pseuds	Seid NWs	Weekes	Ziegler	Chateau	Spieler Young	Spieler Old	Treiman	Waters	Yap
Jared X1	.40	**.54**	.60	.63	.61	.32	.44	.32	-.11	.37			.44	.39	.33	.40	.25	.42
Jared X2	.40	.52	**.61**	.65	.62	.36	.45	.31	-.11	.37			.42	.41	.33	.41	.24	.40
Jared X3	.41	.51	.61	**.65**	.62	.42	.56	.34	-.10	.37			.43	.41	.33	.41	.23	.35
Jared X4	.41	.52	.61	.65	**.62**	.41	.56	.36	-.09	.38			.44	.41	.33	.41	.24	.36
Rastle Stress 1	.38	.24	.44	.51	.48	**.50**	.63	.49	.24	.38			.42	.44	.31	.35	.23	.06
Rastle Stress 2	.41	.42	.55	.63	.60	.49	**.66**	.52	.04	.44			.44	.36	.32	.38	.19	.13
Andrews	.38							**.54**	.16	.45								
Pseuds	.37							.49	**.24**	.39								
Seid NWs	.38							.54	.17	**.45**								
Weekes	-.01	-.30	-.20	-.15	-.15	.17	.16	-.38	.09	-.39	**.70**	.72	.49	-.23	.06	-.07	-.16	-.45
Ziegler	.14	-.16	.03	.17	.15	.30	.30	-.15	.14	-.31	.67	**.75**	.49	.04	.23	.13	.04	-.40
Chateau	.42	.40	.49	.55	.53	.44	.63	.50	.24	.40			**.50**	.44	.31	.34	.30	.24
Spieler Young	.40	.30	.44	.49	.45	.43	.49	.54	.16	.45			.43	**.49**	.33	.35	.30	.36
Spieler Old	.36	.45	.47	.50	.49	.08	.16	.49	.00	.43			.43	.41	**.36**	.30	.31	.48
Treiman	.42	.51	.61	.65	.62	.40	.53	.40	-.08	.39			.44	.43	.33	**.41**	.25	.38
Waters	.41	.43	.47	.52	.50	.27	.41	.54	.14	.45			.49	.45	.34	.32	**.33**	.46
Yap	.32	.41	.35	.39	.39	-.01	.14	.54	.12	.45			.32	.40	.33	.24	.31	**.48**
N. Words		120	120	120	120	120	120	128	84	588	300	160	901	2998	2998	1327	1329	6714
Max		.54	.61	.65	.62	.50	.66	.54	.24	.45	.70	.75	.50	.49	.36	.41	.33	.48
Min		-.30	-.20	-.15	-.15	-.01	.14	-.38	-.11	-.39	.67	.72	.32	-.23	.06	-.07	-.16	-.45
Median		.43	.48	.53	.52	.38	.47	.49	.12	.39	.69	.73	.44	.41	.33	.35	.24	.36
Mean		.34	.43	.49	.47	.33	.44	.38	.07	.32	.69	.73	.44	.35	.30	.31	.22	.23

Note: Jared X1-X4 = Experiments 1–4 of Jared [14], Rastle Stress 1 and 2 = Experiment 1 and 2 of Rastle and Coltheart [32], Andrews = Andrews and Scarratt [15], Pseuds = McCann and Besner [33], Seid NWs = Seidenberg et al. [40], Weekes = Weekes [34], Ziegler = Ziegler et al. [16], Chateau = Chateau and Jared [36], Spieler Young and Old = Experiments 1 and 2 from Spieler and Balota [12], Treiman = Treiman et al. [38], Waters = Seidenberg and Waters [37], Yap = Yap and Balota [39].

https://doi.org/10.1371/journal.pcsy.0000074.t004

least somewhat atypical compared to a more standard reading context (see Reynolds et al. [48]). The error results where the small word only experiments had high error rates predicting the nonword experiments were also interesting, as discussed above.

The nonword results are also interesting when compared with the regression results. The regression equations fit the data with a relatively similar accuracy. However, the regression equations generalized more accurately across the experiments excluding the pseudohomophones where both CDP and the regression equations had poor generalization performance. Since the equation used to fit the nonword data was very simple (letter length, orthographic N), this suggests that CDP's performance on nonwords is made more difficult by the constraint that the same parameter set needs to support both word and nonword reading.

Whilst the difference between the regression equations on the nonwords might be seen as a positive for regression over CDP, the comparisons are somewhat difficult to interpret. This is because we used simple regression to examine the data and because length and orthographic neighborhood were correlated in the nonword experiments. This means that due to collinearity, a regression can potentially produce a reasonable fit, but the predicted strength of different factors may differ with different models. Thus, whilst the models might predict the data well, the factors the models predict are important may be quite different, hence making their interpretation difficult. This can be seen by examining the slopes of the length/orthographic N variables produced by the regression equations. For the Andrews experiment they were 29.21ms/-4.14ms and for the Seidenberg nonword database they were 23.76ms/-3.44ms. Thus, they were quite similar. Alternatively, the slopes from the other experiments often differed substantially from these (Jared X1: 25.2ms/1.13ms, Jared X2: 28.93ms/1.36ms, Jared X3: 18.99ms/.79ms, Jared X4: 14.40ms/.51ms; Rastle X1: -.44ms/-1.10ms; Rastle X2: 2.61ms/-.10ms; Chateau: 0.0ms/-5.24ms; Spieler New: 4.67ms/-.65ms; Spieler Old: 9.49ms/-.15ms; Treiman: 25.86ms/.70ms; Waters: 8.66ms/-1.0ms; Yap: 22.76ms/-2.18ms).

General discussion

Overall, the results showed that CDP performs well even when optimized on very small datasets. Its ability to generalize to other datasets was consistently strong. In contrast, regression models using standard psycholinguistic predictors performed worse except for the nonword-only experiments. This is particularly notable given that regression models do not have to generate phonological outputs, unlike CDP. Importantly, the parameters that influenced CDP's performance remained largely consistent when optimized on small compared to large datasets. Sloppy Parameter Analysis revealed no meaningful differences in which parameters were stiff or sloppy, indicating that the model's internal dynamics are robust to dataset size. These findings suggest that CDP does not overfit when trained on small item sets.

The results CDP produced are important not only because they show that computational cognitive models can be better than simple statistical models even when trained on only small datasets but also because they highlight the importance of incorporating at least some 'built-in' assumptions when modelling cognitive functions. By this we mean that CDP implements an explicit computational theory about the structure of the reading system that includes specific assumptions about the representations and processes used by skilled readers. These are assumed to be general, reflecting the idea that most readers share a common cognitive architecture (see Shallice [49] for a discussion). By using this architecture, all that needs to be done is to optimize the parameters rather than specify (or learn) the model structure. This is likely to be a major reason for the good overall performance of CDP: despite having many parameters, its predefined architecture narrows the space of possible behaviors in principled ways.

The results also allow us to predict different reading strategies that could be used depending on the task conditions (see Zorzi et al. [3] for a discussion with respect to CDP). Notably, comparison between the nonwords, words, and mixed lists revealed that the Minimum Naming Criterion parameter was consistently lower for nonword-only and mixed lists than for word-only lists. This parameter acts as a response criterion for when the final output of the model can be read out from the phoneme-level representation. CDP thus predicts that a lower response criterion is adopted in contexts involving

nonwords, as this helps reduce error rates more effectively than adjusting other parameters. This shows that the values of the parameters that produce optimal results in particular contexts cannot just be determined simply by taking the psycholinguistic statistics of the entire stimuli set and setting the values based on them. Instead, they emerge from CDP's computational architecture and how it processes information.

The results of CDP are also compatible with a criterion shifting account of reading, rather than a change in the overall speed of activation build-up (see Lupker et al. [50]). While the Minimum Naming Criterion varied extensively across different experiments depending on the type of stimuli, the Global Activation Parameter, which controls the speed of activation build up across all representations, remained remarkably stable. Thus, even if it were possible to simulate the qualitative pattern of results by changing other parameters, the poorer quantitative fits between the actual data and CDP would suggest that such alternative strategies are unlikely.

Apart from the results of CDP, the results of the regression equations are interesting in their own right. In the experiments with words, regression generally performed worse than CDP. One reason for this may be because of the specific manipulations that were done in the experiments. For example, in the Jared [14] experiments, different types of consistency were examined but our regression models included only a single measure of consistency. This may have limited the regression model's ability to capture the full effect of the experimental manipulations, contributing to its poorer performance relative to CDP. However, not all variables used in the regression suffered from this problem, including word length, word frequency, and orthographic N. It is therefore unlikely that the weaker performance of regression models can be attributed solely to inconsistencies in how psycholinguistic predictors were defined.

Another possible reason for the worse performance of the regression models based on word datasets when compared to CDP could be the choice of psycholinguistic variables. There are many other predictors we could have used. However, word frequency, word length, consistency and orthographic N are arguably the most important variables in word recognition and naming experiments. It would also be possible to add further variables (e.g., age of acquisition, bigram frequency, concreteness, etc.), but the comparison with CDP would not be fair if they refer to processes or mechanisms that are outside the scope of the current version of CDP (e.g., semantics). It is also important to highlight that regression modelling has a major competitive advantage over CDP: fitting the data does not include the task of generating phonemes. Thus, it seems reasonable to suggest that the ability of CDP to predict the data is somewhat handicapped compared to regression.

One final notable anomaly in the regression results was observed with the two mixed datasets (Weekes and Ziegler), which included both words and nonwords. While regression equations fit these datasets reasonably well, their ability to generalize to other experiments was markedly poor—unlike CDP, which maintained good generalization performance. This is interesting because it suggests that the regression models may overfit more complex data, capturing superficial statistical patterns underlying cognitive processes. Given the increasing use of regression-based analyses in psycholinguistics (e.g., [39,51,52], exploring the extent to which these regression models generalize across datasets would be valuable to assess their validity, especially for weaker effects [45].

Overall, the results have a number of significant implications. First, they demonstrate that even small datasets contain sufficient information to meaningfully optimize CDP and produce good generalization performance. Second, the better performance of CDP over simple regression is likely to be due to its structure that explicitly implements the (assumed) cognitive architecture of the reading system. Third, the process of parameter optimization and SPA provides valuable insights into which components of the model drive behavior and how they vary across contexts. Finally, the results are important because they show that cognitive models can serve as more than just tools for prediction—they can also offer interpretable, mechanistic accounts of behavior. Rather than relying solely on summary statistics or regression coefficients, researchers may increasingly turn to parameter estimates from theoretically grounded models as a basis for explanation, even in small-scale experiments.

Such a modelling approach has already proven successful in other domains of cognitive science, such as the Two-Alternative Forced Choice task. With that task, the drift-diffusion model [53] has been widely used to extract meaningful parameters that reveal underlying cognitive processes—insights that traditional statistical methods cannot provide (e.g., [54–56]. With further research, there is no reason why parameters from reading models could not be used in a similar way to advance our understanding of language processing. In particular, our results pave the way for modelling individual reading profiles using small-size stimulus datasets, as are used in most empirical studies. This approach would provide explicit predictions about which parameters affect the reading performance of a given individual, with important implications for understanding reading difficulties. Indeed, personalized modelling could help identify the cognitive source of reading impairment, both in developmental and acquired dyslexia, and predict which types of remediation might be most effective for a given individual.

Supporting information

S1 Table. R values from regression equations including a frequency by consistency interaction taken from one dataset and used to predict the others.
(DOCX)

S1a Fig. Histograms of the *r* values comparing the actual and model data from the top 100 optimization runs in each experiment. *Note: Stimuli: Jaredx1-x4 = Experiments 1–4 of Jared (1), RastleS1-S2 = Experiments 1 and 2 of Rastle and Coltheart (2), Andrews = Andrews and Scarratt (3), Pseuds = McCann and Besner (4), Seid NWs = Seidenberg et al. (5), Weekes = Weekes (6), Ziegler = Ziegler et al. (7), Chateau = Chateau and Jared (8), Spieler Young and Old = Experiments 1 and 2 from Spieler and Balota (9), Treiman = Treiman et al. (10), Waters = Seidenberg and Waters (11), Yap = Yap and Balota (12).
(TIFF)

S1b Fig. Histograms of the *r* values of Pearson correlations generated by taking the reaction times of the top 20 optimization runs and using the correlation values from all possible pairings. * Note: Stimuli: Jaredx1-x4 = Experiments 1–4 of Jared(1), RastleS1-S2 = Experiments 1 and 2 of Rastle and Coltheart (2), Andrews = Andrews and Scarratt (3), Pseuds = McCann and Besner (4), Seid NWs = Seidenberg et al. (5), Weekes = Weekes (6), Ziegler = Ziegler et al. (7), Chateau = Chateau and Jared (8), Spieler Young and Old = Experiments 1 and 2 from Spieler and Balota (9), Treiman = Treiman et al. (10), Waters = Seidenberg and Waters (11), Yap = Yap and Balota (12).
(TIFF)

S1c Fig. Histograms of the RMSE values generated by taking the reaction times of the top 20 optimization runs of each experiment and using the RMSE values from all possible pairings. *Note: Stimuli: Jaredx1-x4 = Experiments 1–4 of Jared (1), RastleS1-S2 = Experiments 1 and 2 of Rastle and Coltheart (2), Andrews = Andrews and Scarratt (3), Pseuds = McCann and Besner (4), Seid NWs = Seidenberg et al. (5), Weekes = Weekes (6), Ziegler = Ziegler et al. (7), Chateau = Chateau and Jared (8), Spieler Young and Old = Experiments 1 and 2 from Spieler and Balota (9), Treiman = Treiman et al. (10), Waters = Seidenberg and Waters (11), Yap = Yap and Balota (12).
(TIFF)

S2 Fig. Histograms of the parameter distributions from 100 optimization runs on each dataset. The low and high values of the X axis are taken from the minimum and maximum values in Table 1. The largest value of the Y axis is taken from the bin with the highest count. * Note: Stimuli: Jaredx1-x4 = Experiments 1–4 of Jared (1), RastleS1-S2 = Experiments 1 and 2 of Rastle and Coltheart (2), Andrews = Andrews and Scarratt (3), Pseuds = McCann and Besner (4), Seid = Seidenberg et al. (5), Weekes = Weekes (6), Ziegler = Ziegler et al. (7), Chateau = Chateau and Jared (8), Spieler Young and Old = Experiments 1 and 2 from Spieler and Balota (9), Treiman = Treiman et al. (10), Waters = Seidenberg and Waters (11), Yap = Yap and Balota (12). Parameters: ff = feedforward, fb = feedback, Ex – Excitatory, In = Inhibitory,

Feat = Feature, Let = Letter, LatIn = lateral inhibition, OL = Orthographic Lexicon, PL = Phonological Lexicon, St = Stress, Stress_Over = stress over criterion, Let_Scan = letter scanning time, Act = activation, FreqMod = Frequency modifier, Min_Naming = Minimum Naming criterion, Let_Over = Letter Over criterion.
(TIFF)

S3 Fig. Histograms of the SD values from the small experiments and large datasets.
(TIFF)

S4 Fig. Histograms of log eigenvalues from the small experiments and large datasets.
(TIFF)

S5 Fig. Histograms of the values of the parameter ranking scores from the small experiments and large datasets.
(TIFF)

Financial disclosure statement

This research was supported by the Australian Research Council (Grant DP210100936 to CP; https://www.arc.gov.au/ and the Institute of Convergence ILCB (France 2030, ANR-16-CONV-0002 to JZ; https://www.ilcb.fr/. MZ is supported by the Italian Ministry of Health (Ricerca Corrente to IRCCS San Camillo Hospital; https://www.sancamilloscientifico.it/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests statement

The authors have declared that no competing interests exist.

Additional information

Correspondence with respect to this article can be sent to ConradPerry@gmail.com.

Author contributions

Conceptualization: Conrad Perry, Marco Zorzi, Johannes Ziegler.

Formal analysis: Conrad Perry.

Funding acquisition: Conrad Perry.

Methodology: Marco Zorzi, Marco Zorzi, Johannes Ziegler.

Project administration: Conrad Perry, Johannes Ziegler.

Software: Conrad Perry.

Supervision: Marco Zorzi, Johannes Ziegler.

Validation: Conrad Perry, Marco Zorzi, Johannes Ziegler.

Visualization: Conrad Perry.

Writing – original draft: Conrad Perry, Marco Zorzi, Johannes Ziegler.

Writing – review & editing: Conrad Perry, Marco Zorzi, Johannes Ziegler.

References

1. Seidenberg MS. Computational models of reading. In: Spivey M, McRae K, Joanisse M, editors. The Cambridge Handbook of Psycholoinguistics. Cambridge University Press; 2012. p. 186–203.
2. Perry C, Evertz R, Zorzi M, Ziegler JC. Understanding the complexity of computational models through optimization and sloppy parameter analyses: The case of the Connectionist Dual-Process Model. J Mem Lang. 2024;134:104468. https://doi.org/10.1016/j.jml.2023.104468

3. Zorzi M, Houghton G, Butterworth B. The Development of Spelling-Sound Relationships in a Model of Phonological Reading. Lang Cogn Process. 1998;13(2–3):337–71. https://doi.org/10.1080/016909698386555
4. Zorzi M, Houghton G, Butterworth B. Two routes or one in reading aloud? A connectionist dual-process model. J Experim Psychol Hum Percept Perform. 1998;24(4):1131–61. https://doi.org/10.1037/0096-1523.24.4.1131
5. Perry C, Ziegler JC, Zorzi M. Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. Psychol Rev. 2007;114(2):273–315. https://doi.org/10.1037/0033-295X.114.2.273 PMID: 17500628
6. Perry C, Ziegler JC, Zorzi M. Beyond single syllables: large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. Cogn Psychol. 2010;61(2):106–51. https://doi.org/10.1016/j.cogpsych.2010.04.001 PMID: 20510406
7. Perry C, Ziegler JC, Zorzi M. A computational and empirical investigation of graphemes in reading. Cogn Sci. 2013;37(5):800–28. https://doi.org/10.1111/cogs.12030 PMID: 23489148
8. Perry C, Ziegler JC, Zorzi M. CDP++.Italian: modelling sublexical and supralexical inconsistency in a shallow orthography. PLoS One. 2014;9(4):e94291. https://doi.org/10.1371/journal.pone.0094291 PMID: 24740261
9. Perry C, Ziegler JC, Zorzi M. When silent letters say more than a thousand words: An implementation and evaluation of CDP++ in French. J Mem Lang. 2014;72:98–115. https://doi.org/10.1016/j.jml.2014.01.003
10. Coltheart M, Rastle K, Perry C, Langdon R, Ziegler J. DRC: a dual route cascaded model of visual word recognition and reading aloud. Psychol Rev. 2001;108(1):204–56. https://doi.org/10.1037/0033-295x.108.1.204 PMID: 11212628
11. Plaut DC, McClelland JL, Seidenberg MS, Patterson K. Understanding normal and impaired word reading: computational principles in quasi-regular domains. Psychol Rev. 1996;103(1):56–115. https://doi.org/10.1037/0033-295x.103.1.56 PMID: 8650300
12. Spieler DH, Balota DA. Bringing Computational Models of Word Naming Down to the Item Level. Psychol Sci. 1997;8(6):411–6. https://doi.org/10.1111/j.1467-9280.1997.tb00453.x
13. Balota DA, Yap MJ, Cortese MJ, Hutchison KA, Kessler B, Loftis B, et al. The English Lexicon Project. Behav Res Methods. 2007;39(3):445–59. https://doi.org/10.3758/bf03193014 PMID: 17958156
14. Jared D. Spelling-Sound Consistency and Regularity Effects in Word Naming. J Mem Lang. 2002;46(4):723–50. https://doi.org/10.1006/jmla.2001.2827
15. Andrews S, Scarratt DR. Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds? J Experim Psychol Hum Percept Perform. 1998;24(4):1052–86. https://doi.org/10.1037/0096-1523.24.4.1052
16. Ziegler JC, Perry C, Jacobs AM, Braun M. Identical words are read differently in different languages. Psychol Sci. 2001;12(5):379–84. https://doi.org/10.1111/1467-9280.00370 PMID: 11554670
17. Zorzi M. The connectionist dual process (CDP) approach to modelling reading aloud. Europ J Cogn Psychol. 2010;22(5):836–60. https://doi.org/10.1080/09541440903435621
18. Perry C. Graphemic parsing and the basic orthographic syllable structure. Lang Cogn Process. 2013;28(3):355–76. https://doi.org/10.1080/01690965.2011.641386
19. Balota DA, Spieler DH. The Utility of Item-Level Analyses in Model Evaluation: A Reply to Seidenberg and Plaut. Psychol Sci. 1998;9(3):238–40. https://doi.org/10.1111/1467-9280.00047
20. Perry C, Zorzi M, Ziegler JC. Understanding Dyslexia Through Personalized Large-Scale Computational Models. Psychol Sci. 2019;30(3):386–95. https://doi.org/10.1177/0956797618823540 PMID: 30730792
21. Gutenkunst RN, Casey FP, Waterfall JJ, Myers CR, Sethna JP. Extracting falsifiable predictions from sloppy models. Ann N Y Acad Sci. 2007;1115:203–11. https://doi.org/10.1196/annals.1407.003 PMID: 17925353
22. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol. 2007;3(10):1871–8. https://doi.org/10.1371/journal.pcbi.0030189 PMID: 17922568
23. Hartoyo A, Cadusch PJ, Liley DTJ, Hicks DG. Parameter estimation and identifiability in a neural population model for electro-cortical activity. PLoS Comput Biol. 2019;15(5):e1006694. https://doi.org/10.1371/journal.pcbi.1006694 PMID: 31145724
24. Kardynska M, Smieja J, Naumowicz A, Janus P, Widlak P, Kimmel M, editors. Sloppy/Stiff parameters rankings in sensitivity analysis of signaling pathways. International Conference on Bioinformatics Models, Methods and Algorithms. 2016.
25. Stone GO, Van Orden GC. Strategic control of processing in word recognition. J Exp Psychol Hum Percept Perform. 1993;19(4):744–74. https://doi.org/10.1037//0096-1523.19.4.744 PMID: 8409857
26. Kinoshita S, Lupker SJ. Priming and attentional control of lexical and sublexical pathways in naming: a reevaluation. J Exp Psychol Learn Mem Cogn. 2023;29(3):405–15. https://doi.org/10.1037/0278-7393.29.3.405 PMID: 12776751
27. Reynolds M, Besner D. Basic processes in reading: a critical review of pseudohomophone effects in reading aloud and a new computational account. Psychon Bull Rev. 2005;12(4):622–46. https://doi.org/10.3758/bf03196752 PMID: 16447376
28. Zevin JD, Balota DA. Priming and attentional control of lexical and sublexical pathways during naming. J Exp Psychol Learn Mem Cogn. 2000;26(1):121–35. https://doi.org/10.1037//0278-7393.26.1.121 PMID: 10682293
29. Ziegler JC, Perry C, Zorzi M. Learning to Read and Dyslexia: From Theory to Intervention Through Personalized Computational Models. Curr Dir Psychol Sci. 2020;29(3):293–300. https://doi.org/10.1177/0963721420915873 PMID: 32655213

30. Ziegler JC, Perry C, Zorzi M. Understanding normal and impaired reading development through personalized large-scale neurocomputational model. Cambridge Handbook Cogn Devel. 2022. p. 554–65. https://doi.org/10.1017/978110839983
31. Kennedy J, Eberhart R. Particle Swarm Optimization. In: Proceedings of ICNN'95 - International Conference on Neural Networks. 1995. https://doi.org/10.1109/ICNN.1995.488968
32. Rastle K, Coltheart M. Lexical and Nonlexical Print-to-Sound Translation of Disyllabic Words and Nonwords. J Mem Lang. 2000;42(3):342–64. https://doi.org/10.1006/jmla.1999.2687
33. McCann RS, Besner D. Reading pseudohomophones: Implications for models of pronunciation assembly and the locus of word-frequency effects in naming. J Experim Psychol Hum Percept Perform. 1987;13(1):14–24. https://doi.org/10.1037/0096-1523.13.1.14
34. Weekes BS. Differential Effects of Number of Letters on Word and Nonword Naming Latency. Quarter J Experim Psychol. 1997;50(2):439–56. https://doi.org/10.1080/713755710
35. Perry C, Evertz R, Zorzi M, Ziegler JC. Understanding the complexity of computational models through optimization and sloppy parameter analyses: The case of the Connectionist Dual-Process Model. J Mem Lang. 2024;134:104468. https://doi.org/10.1016/j.jml.2023.104468
36. Chateau D, Jared D. Spelling–sound consistency effects in disyllabic word naming. J Mem Lang. 2003;48(2):255–80. https://doi.org/10.1016/s0749-596x(02)00521-1
37. Seidenberg MS, Waters GS. Word recognition and naming: A mega study [Abstract]. Bull Psychon Soc. 1989.
38. Treiman R, Mullennix J, Bijeljac-Babic R, Richmond-Welty ED. The special role of rimes in the description, use, and acquisition of English orthography. J Exp Psychol Gen. 1995;124(2):107–36. https://doi.org/10.1037//0096-3445.124.2.107 PMID: 7782735
39. Yap MJ, Balota DA. Visual word recognition of multisyllabic words. J Memory Lang. 2009;60(4):502–29. https://doi.org/10.1016/j.jml.2009.02.001
40. Seidenberg MS, Plaut DC, Petersen AS, McClelland JL, McRae K. Nonword pronunciation and models of word recognition. J Exp Psychol Hum Percept Perform. 1994;20(6):1177–96. https://doi.org/10.1037//0096-1523.20.6.1177 PMID: 7844510
41. Machta BB, Chachra R, Transtrum MK, Sethna JP. Parameter space compression underlies emergent theories and predictive models. Science. 2013;342(6158):604–7. https://doi.org/10.1126/science.1238723 PMID: 24179222
42. Panas D, Amin H, Maccione A, Muthmann O, van Rossum M, Berdondini L, et al. Sloppiness in spontaneously active neuronal networks. J Neurosci. 2015;35(22):8480–92. https://doi.org/10.1523/JNEUROSCI.4421-14.2015 PMID: 26041916
43. Ponce-Alvarez A, Mochol G, Hermoso-Mendizabal A, de la Rocha J, Deco G. Cortical state transitions and stimulus response evolve along stiff and sloppy parameter dimensions, respectively. Elife. 2020;9:e53268. https://doi.org/10.7554/eLife.53268 PMID: 32181740
44. Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. J Chem Phys. 2015;143(1):010901. https://doi.org/10.1063/1.4923066 PMID: 26156455
45. Perry C. Using Monte-Carlo simulation to test predictions about the time-course of semantic and lexical access in reading. PLoS One. 2024;19(4):e0296874. https://doi.org/10.1371/journal.pone.0296874 PMID: 38564586
46. Balota DA, Cortese MJ, Sergent-Marshall SD, Spieler DH, Yap M. Visual word recognition of single-syllable words. J Exp Psychol Gen. 2004;133(2):283–316. https://doi.org/10.1037/0096-3445.133.2.283 PMID: 15149254
47. Keuleers E, Lacey P, Rastle K, Brysbaert M. The British lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. Behav Res Methods. 2012;44:287–304. https://doi.org/10.3758/s13428-010-0020-5
48. Reynolds M, Besner D, Coltheart M. Reading aloud: new evidence for contextual control over the breadth of lexical activation. Mem Cogn. 2011;39(7):1332–47. https://doi.org/10.3758/s13421-011-0095-y PMID: 21830161
49. Shallice T. Cognitive neuropsychology and its vicissitudes: The fate of Caramazza's axioms. Cogn Neuropsychol. 2015;32(7–8):385–411. https://doi.org/10.1080/02643294.2015.1131677 PMID: 27355606
50. Lupker S. Mixing costs and mixing benefits in naming words, pictures, and sums. J Mem Lang. 2003;49(4):556–75. https://doi.org/10.1016/s0749-596x(03)00094-9
51. Cortese MJ, Khanna MM. Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: an analysis of 2,342 words. Q J Exp Psychol. 2007;60(8):1072–82. https://doi.org/10.1080/17470210701315467 PMID: 17654392
52. Khanna MM, Cortese MJ. How well imageability, concreteness, perceptual strength, and action strength predict recognition memory, lexical decision, and reading aloud performance. Mem Cogn. 2021;29(5):622–36. https://doi.org/10.1080/09658211.2021.1924789 PMID: 33971794
53. Ratcliff R. A theory of memory retrieval. Psychol Rev. 1978;85(2):59–108. https://doi.org/10.1037/0033-295x.85.2.59
54. Ratcliff R, Rouder JN. Modeling Response Times for Two-Choice Decisions. Psychol Sci. 1998;9(5):347–56. https://doi.org/10.1111/1467-9280.00067
55. Milosavljevic M, Malmaud J, Huth A, Koch C, Rangel A. The Drift Diffusion Model can account for the accuracy and reaction time of value-based choices under high and low time pressure. Judgm Decis Mak. 2010;5(6):437–49. https://doi.org/10.1017/s1930297500001285
56. Pedersen ML, Frank MJ, Biele G. The drift diffusion model as the choice rule in reinforcement learning. Psychon Bull Rev. 2017;24(4):1234–51. https://doi.org/10.3758/s13423-016-1199-y PMID: 27966103