A Bayesian approach to dynamical modeling of eye-movement control in reading of normal, mirrored, and scrambled texts

Maximilian M. Rabe, Johan Chandra, André Krügel, Stefan A. Seelig, Shravan Vasishth, and Ralf Engbert
University of Potsdam

In eye-movement control during reading, advanced process-oriented models have been developed to reproduce behavioral data. So far, model complexity and large numbers of model parameters prevented rigorous statistical inference and modeling of interindividual differences. Here we propose a Bayesian approach to both problems for one representative computational model of sentence reading (SWIFT; Engbert, Nuthmann, Richter, & Kliegl, 2005). We used experimental data from 36 subjects who read text in a normal and one of four manipulated text layouts (e.g., mirrored and scrambled letters). The SWIFT model was fitted to subjects and experimental conditions individually to investigate between-subject variability. Based on posterior distributions of model parameters, fixation probabilities and durations are reliably recovered from simulated data and reproduced for withheld empirical data, at both the experimental condition and subject levels. A subsequent statistical analysis of model parameters across reading conditions generates model-driven explanations for observable effects between conditions.

Keywords: Reading, eye movement control, dynamical models, Bayesian inference, oculomotor control, individual differences

Introduction

Reading is an important everyday task that is characterized by high adaptivity. As a consequence, behavioral measures like reading rates or fixation durations vary strongly during silent vs. oral reading, reading of easy vs. difficult texts, and differ between proof-reading, mindless reading, or reading of scrambled texts. Such variations and adaptivity represent a key challenge for mathematical models of eye-movement control. Recent advances in Bayesian model inference for dynamical cognitive models (Schütt et al., 2017) provide the tools for rigorous evaluation of model generalizability. Here we investigate the generalizability of the SWIFT model (Engbert et al., 2005) from normal reading to several manipulations of the spatial layout of texts, i.e., text composed of words with mirrored, inverted, and scrambled letters, which are known to induce strong effects on reading performance (Kolers, 1976; Rayner, White, Johnson, & Liversedge, 2006).

During reading of normal texts, the reader generates 3 to 4 saccades per second; word processing occurs during fixations on different words with average durations in a range between 150 and 300 ms (Rayner, 1998). The number of fixations in a given text is of the same order as the number of words, however, some words do not receive a fixation (word skipping) while others are targeted multiple times. A secondary fixation of the same word as the currently fixated word is denoted as a refixation. However, the eye's scanpath is even more complicated, since some saccades go against the reading direction to previously inspected regions of text. Such regressions represent about 5 to 10% of the saccades in a typical text.

Two typical eye trajectories are presented in Figure 1. In both examples, forward saccades occur most frequently, as observed on average. As far as other fixation types are concerned, for example in the upper panel, the eyes generate refixations of the same word (e.g., fixations 8, 9 and 14), skip words between fixations 5 and 6 (the skipped word is the

Corresponding author. Correspondence should be addressed to Maximilian M. Rabe, maximilian.rabe@uni-potsdam.de. **Data.** All data and analyses reported in this article are made available online via the Open Science Framework (OSF, doi: 10.17605/osf.io/t9sbf). **Previous dissemination of results.** The manuscript was previously uploaded as a preprint to PsyArxiv (doi: 10.31234/osf.io/nw2pb). Empirical data in this manuscript were previously analyzed and discussed in Chandra, Krügel, and Engbert (2020). Simulations and modeling work were presented as a poster at the 60th Annual Meeting of the Psychonomic Society (Montréal, QC, Canada), at the 20th European Conference on Eye Movements (ECEM, Alicante, Spain), and at the 2019 Conference on Cognitive Computational Neuroscience (Berlin, Germany).

German conjunction “und”), and produce a regression from fixation 9 to 10.

Everyday circumstances require reading geometrically altered or manipulated words or sentences. As an example, reading transformed texts is necessary when reading mirror reflections of text on signs. Such manipulations do occur in everyday life and invoke drastic changes in reading patterns, even after training (Kolers, 1976; Kolers & Perkins, 1975). Another common type of text manipulation is intentional or unintentional scrambling of letters within words (Table 1).

A popular internet myth of the early 2000s claimed that reading sentences of words with scrambled letters were still readable and easy to understand. Rayner et al. (2006) investigated the statement and found that contrary to the claim, which was in fact not backed up by any scientific evidence, there is indeed a cognitive cost even though reading of such sentences is not greatly impaired.

In the present theoretical study, we fitted the SWIFT model to diverse reading patterns to evaluate whether the model can reproduce the variability between experimental conditions and baseline, following a principled workflow that improves model fit, inferences, and comparability (Schad, Betancourt, & Vasishth, 2020). In this approach, the likelihood function plays a key role as an objective optimization target for model fitting that was introduced in an earlier publication by Seelig et al. (2020). What is novel here is that we (i) run more extensive simulations using more free parameters, (ii) use a more powerful MHMC algorithm in the Bayesian framework, (iii) reproduce a more representative range of reading behaviors using the full covariance structure of the fitted posterior distributions, (iv) evaluate an experiment with 5 different reading conditions, and (v) develop an improved oculomotor model of saccadic landing positions. We will present subject-level results, so that observed patterns could be reproduced for each particular subject showing that between-subject variability can be captured by the model. Due to the principled Bayesian workflow (Schad, Vasishth, Hohenstein, & Kliegl, 2020), our approach includes (1) rigorous statistical inference, (2) an evaluation of goodness-of-fit for specific effects, and (3) explanations for findings via effects found in model parameters. All source code that was used for the analyses reported in this article is publicly available online.[1]

For the current work, we use eye-movement data from reading experiments on geometric alterations of text layout and scrambled-letter words. We expect this data-set to posit a challenge for dynamical reading models; mathematical models should be challenged to fit observed reading behavior across tasks, while readers should be challenged with respect to their performance. We also expect substantial interindividual differences; thus, the model should also be able to detect, reproduce, and explain the observed level of between-subject variability.

The Bayesian approach to dynamical cognitive models

Dynamical cognitive models represent a framework that permits the test of very specific hypotheses about cognitive processes underlying human behavior (Schütt et al., 2017), in particular when such models are investigated in a principled Bayesian workflow (Schad, Betancourt, & Vasishth, 2020). A strong test of dynamical models, however, requires time-ordered observations, such as eye movements, brain imaging, or single-cell recordings or other types of high-density behavioral data. As we will demonstrate, dynamical models are highly flexible and can implement processes for many observable dimensions, assuming that the same implemented processes can make predictions for all considered observables.

Generally, experimental data X_n for a dynamical model are sequences of n observed discrete instances $(x_1, \dots, x_n)$, expandable to an $n \times m$ matrix,

$$X_n = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \\ x_{n1} & & x_{nm} \end{pmatrix}, \tag{1}$$

where n is the number of time-ordered instances and m is the number of considered observables or measures. Typically, we assume that n is clearly greater than 1; for eye movements, n might be on the order of 10. Critically, each of these instances should provide data on all m measures.

A mathematical model with a computable likelihood function (as function of free model parameters and for a given dataset) can be fitted in a Bayesian inference framework if the necessary numerical implementation is efficient. In contrast to maximum-likelihood (MLE) or frequentist methods, Bayesian methods provide inference based on credible intervals for model parameters (see Schütt et al., 2017, for a dynamical cognitive model). The credible intervals relate to model plausibility and stability. To obtain a posterior probability distribution $P_M(\theta \mid X)$ for a model M specified by a set of parameters θ after observing data X, we first need to determine the likelihood $L_M(\theta \mid X)$ of the data X given some parameter set θ and the prior probability distribution $Q(\theta)$ over parameters θ, so that

$$P_M(\theta \mid X) \propto L_M(\theta \mid X) \cdot Q(\theta) . \tag{2}$$

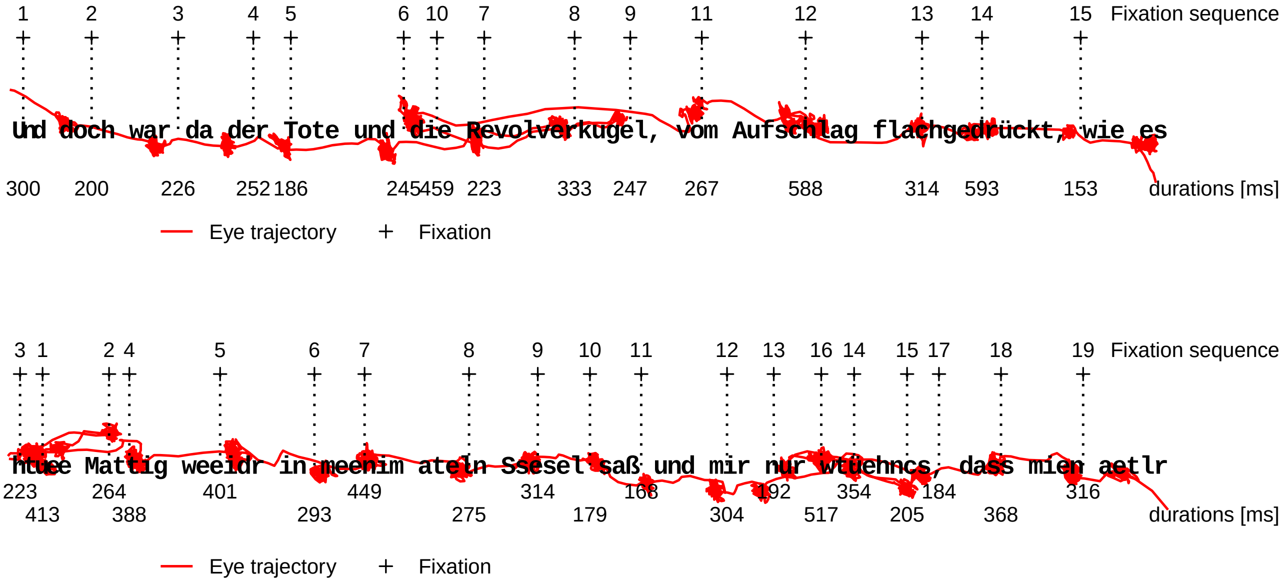
While the definition of $L_M(\theta \mid X)$ is typically objective and based on stringent mathematical formulation, the prior parameter distribution should ideally be based on domain expertise, which might include various forms of knowledge from cognitive to physiological processes.

In contrast to maximum likelihood estimation (MLE), which can quickly be overwhelmed by high dimensionality

[1] Analyses are available online at https://doi.org/10.17605/osf.io/t9sbf.

Figure 1

Typical eye trajectories during reading



Note. The upper panel represents the normal reading condition, whereas the lower panel represents an example of the scrambled reading condition.

Table 1

Reading conditions used as modeling targets

	Description	Order	Example						
N	Normal	LR	Jede	Sprache	der	Welt	besitzt	eine	Grammatik
mL	Mirrored letters	LR	Ⴑɘbɘ	Ƨqɿɒɔʜɘ	bɘɿ	Wɘlɟ	dɘƨiɟzɟ	ɘinɘ	Әɿɒmmɒɟiʞ
sL	Scrambled	LR	Jdee	Scrahpe	der	Wlet	bsizett	enie	Gmartimak
iW	Reversed letters	RL	edeJ	ehcarpS	red	tleW	tztiseb	enie	kitammarG
mW	Mirrored word	RL	ɘbɘႱ	ɘʜɔɒɿqƧ	ɿɘb	ɟlɘW	ɟzɟiƨɘb	ɘniɘ	ʞiɟɒmmɒɿӘ

Note. Letter order applies with regard to first and last letter of the word. *LR* = left-to-right, *RL* = right-to-left

(i.e., many free model parameters), the definition of a prior is what makes fitting complex models possible in the first place. This is because priors bound the parameter space to a computationally tangible subspace and avoid sampling of *a priori* unlikely model configurations. If domain expertise on model parameters is not readily available, uninformative priors with support on a wider range of values and weak maxima can be a sensible fallback option and tend to converge on similar solutions as MLE.

Bayesian parameter estimation enables us to infer statistically rigorous credible intervals for model parameters. Credible intervals can serve (i) to characterize different theoretical entities (i.e., subjects or items) and (ii) to account for variability induced due to the experimental manipulation. In order to permit Bayesian parameter inference, the model needs to provide a likelihood function $L_M(X_n \mid \theta)$ for time-ordered dataset X_n given some model configuration θ. The likelihood function is the product of the likelihood of all instances x_i of X_n, each conditional on model parameters θ and all previous instances $X_{i-1} = (x_1, \ldots, x_{i-1})$, i.e.,

$$L_M(X_n \mid \theta, \xi) = L_M(x_1 \mid \theta, \xi) \prod_{i=2}^{n} L_M(x_i \mid \theta, \xi, X_{i-1}) \ . \quad (3)$$

The additional variable ξ denotes internal degrees of freedom, which are stochastic states of saccade programming and word activation in the SWIFT model (Seelig et al., 2020). As a consequence, the likelihood is inherently stochastic and we will use an approximate *pseudo-marginal* likelihood $L_M(X_n \mid \theta, \xi)$ (Andrieu & Roberts, 2009) with internal degrees of freedom ξ.

Given a likelihood function and specified prior distributions, there exist different methods of sampling from the posterior distribution of model parameters. The most important numerical algorithm is the Metropolis-Hastings (MH) algorithm, which was developed by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and subsequently generalized by Hastings (1970). The class of *Metropolis-Hastings Monte Carlo* (MHMC) algorithms can become demanding in terms of computational resources, but requires less mathematical prerequisites (such as the definition of likelihood derivatives) than more advanced approaches. Therefore, the MHMC can be considered an adequate choice for complex models (Schütt et al., 2017), which is particularly true for models without an exact closed-form likelihood and stochastic internal degrees of freedom requiring a pseudo-marginal approach (Seelig et al., 2020).

In the MHMC methods, the sampler builds a *chain* in parameter space step by step. For each iteration, the sampler makes a proposal for a new parameter set based on its current state and evaluates whether the proposal provides a better fit than the previous one. If it does, it is accepted with certain probability. If not, it is rejected and stays with the previous proposal. Each accepted proposal represents a new sample from the posterior distribution and, therefore, the chain in its entirety will approach the desired *posterior* probability of the parameters.

Principled Bayesian workflow in model inference

In the following, several procedures are implemented to ensure computational faithfulness of model and sampling method, to evaluate the predictive power of the fitted model, and to make inferences to explain observed variability with assumed underlying model behavior. We adopted the principled Bayesian workflow discussed in Schad, Betancourt, and Vasishth (2020) to secure validity and reliability of our numerical inferences. The steps taken are as follows:

1. Definition of a generative model and derivation of an (approximate) likelihood function,
2. Check of the computational faithfulness of the model by inspecting likelihood profiles,
3. Prior predictive simulations,[2]
4. Test of the computational faithfulness of the sampling algorithm via parameter recovery,
5. Split of empirical datasets into fitting (train) and validation (test) datasets for cross-validation,
6. Analysis of posterior predictive checks on test datasets (cross-validation) and model predictions based on the generative model and fitted parameter values, and
7. Statistical evaluation of model parameters between experimental condition.

Summary statistics

In a successful mathematical model, simulated and empirical data will be in good agreement at the level of global summary statistics commonly reported in the literature. In our approach, summary statistics are not the primary target of model optimization, since the objective likelihood-based model fitting technique is neutral to the outcome at the level of specific summary statistics. Instead, summary statistics are applied for the comparison between withheld empirical data and data simulated with the generative after model after parameter fitting to evaluate goodness-of-fit (Roberts & Pashler, 2000). From this perspective, our approach might be looked upon as a case study for other models in eye-movement research in reading (e.g. Reichle, Pollatsek, & Rayner, 2012; Reilly & Radach, 2006; Snell, van Leipsig, Grainger, & Meeter, 2018). Related analyses in the principled Bayesian workflow are *prior* and *posterior predictive checks* discussed below.

Since our model aims at capturing and explaining both temporal and spatial aspects of eye movements in reading, it must be evaluated via spatial and temporal summary statistics. As discussed in the Introduction, saccades do not always move the eye's fixation point from word n to $n+1$; beyond such one-step saccades, there are word skippings, refixations, and regressions. Thus, a successful reading model should reproduce and predict fixation patterns, quantitatively described by fixation probabilities, i.e., the probability to fixate (or skip) a word in given context.

To investigate whether the model makes viable predictions, we evaluated first-pass fixation probabilities, which we defined as follows. The *single-fixation probability* is the proportion of times for a word to receive a fixation that is not followed by a refixation. Conversely, the *refixation probability* is the proportion of times for a word to receive at least one refixation. A word's *skipping probability* denotes whether it is fixated at all (i.e., skipped) in first-pass. Finally, the *(outgoing) regression probability* of a word is its probability to be fixated before a regressive saccade.

While fixation probabilities more closely relate to cognitive processing load in SWIFT, saccade lengths and landing positions are additionally modulated by low-level oculomotor processes (noise and biases occurring at the level of the motor implementation). We therefore also evaluate distributions of saccade lengths and within-word landing positions to verify that the oculomotor assumptions of the model are in line with the empirical data. This is particularly relevant for our investigation, since we expected to optimize statistics

[2] As we are currently using weakly informative priors, we are not reporting prior predictive checks in this paper. Future publications should incorporate expectations based on prior observations and theory and use more informative priors. These should then be evaluated using prior predictive checks.

via the modified Gamma-distributed saccade lengths (Appendix B).

In order to evaluate the goodness-of-fit at a temporal level of eye guidance, we compare simulated and empirical fixation duration measures. A word's *first-fixation duration* and *refixation duration* describe how long the eyes dwell on a word given that it is the first fixation or the second fixation (refixation) on that word, respectively. The *gaze duration* is the total time of all consecutive fixations on the same word given that it was the first time that word was encountered.

The SWIFT model of eye-movement control

The SWIFT model (Engbert et al., 2005; Seelig et al., 2020) is a dynamical cognitive model of eye-movement control during reading. The model can describe, explain, and predict temporal and spatial aspects that are commonly observed in eye trajectories recorded during natural reading. It is among several competitor models that aim at predicting and explaining similar eye movement statistics (e.g., see Reichle, Pollatsek, Fisher, & Rayner, 1998; Reilly & Radach, 2006; Snell et al., 2018). In Figure 2, a simulated eye trajectory as generated by SWIFT is presented.

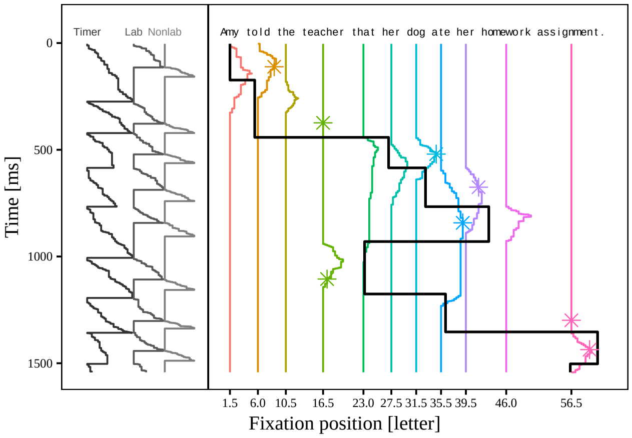
A core concept of the model is parallel processing of several words at a time. All words within the *processing span* around the current fixation location are processed in parallel (Engbert, Longtin, & Kliegl, 2002; Snell & Grainger, 2019). As long as a word's recognition is ongoing, its activation will rise up to a threshold that is modulated by word frequency and related model parameters. Once the threshold is reached, lexical processing is complete and post-lexical processing begins, which is reflected by decreasing activation. The word has been fully processed as soon as the activation returns to zero.

In SWIFT, saccade target selection is inherently stochastic. At any given time t, the probability to select a target word is computed from the relative word activations. If a word is more highly activated than any other word in the activation field, it is the most likely word to be selected as the next target. This also implies that words that are processed faster are on average less likely to be selected as saccade targets. This mechanism provides the basis for the generation of all types of saccades (including skippings, refixations, and regressions) from a single theoretical principle (Engbert et al., 2002).

The decision when to move the eyes is basically independent of the decision where to move the eyes (see Findlay & Walker, 1999). A cascade of random timers (gray lines in the left-hand panel of Figure 2) implement the temporal programming of saccades. A global saccade timer starts whenever the eyes settle on a fixation location. As soon as it reaches threshold, the labile saccade program begins. The saccade at this point can still be cancelled and target selection is still variable. It is not until the start of the non-labile phase that the saccade is inevitably programmed and a target has been selected. Once the non-labile saccade phase reaches its maximum value, the saccade is executed to the previously selected target.

Figure 2

An eye trajectory as simulated in SWIFT



Note. The thick black line is the simulated trajectory. Colored lines are word activations and gray lines on the left are saccade timer random walks, each as a function of time. Asterisks mark the points in time when the labile stage is complete and the target is selected.

Saccade execution is modulated by oculomotor errors (Engbert & Krügel, 2010; Krügel & Engbert, 2014). In fact, McConkie, Kerr, Reddix, and Zola (1988) proposed the *saccadic range error* (SRE) model, stating that the landing position is driven by systematic and random contributions, both of which depend on the distance between launch site and intended target. The systematic error describes saccade amplitudes as having an optimal expected value (mean), as close targets tend to be overshot and far targets undershot. The unsystematic error, sometimes termed *oculomotor noise*, also proposes a relationship between the variance of saccade amplitudes and the target distance, with amplitudes having a higher variance for more distant intended targets. In current versions of SWIFT, spatial aspects of saccade execution implement this model. Appendix A provides more mathematical details of key aspects of SWIFT, while Appendix B extends on the oculomotor assumptions.

The likelihood function for SWIFT

If model inference is done in a Bayesian framework, the computation of the likelihood for a given fixation sequence (such as the one shown in Figure 2) is required. While the concept of the likelihood function is well-established (see

Myung, 2003, for a tutorial), the calculations can be difficult. Alternatively, approximate versions of the likelihood function can be implemented (Palestro, Sederberg, Osth, Van Zandt, & Turner, 2018).

For generative models of eye movements in reading, data are given as sequences of fixations in an $n \times 4$ matrix F_n. In a sequence, each fixation $f_i = (k_i, l_i, T_i, s_i)$ is associated with the fixated word k_i, the landing position l_i within word k_i, the duration T_i of that fixation, and the duration of the consecutive saccade s_i,

$$F_n = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} k_1 & l_1 & T_1 & s_1 \\ \vdots & \vdots & \vdots & \vdots \\ k_n & l_n & T_n & s_n \end{pmatrix} \tag{4}$$

Recently, Seelig et al. (2020) have proposed and investigated an approximate likelihood function for the SWIFT model. In this approach, the likelihood of a fixation f_i is given as the combined spatial and temporal likelihood components, i.e.,

$$L_M\left(k_i, l_i, T_i \mid F_{i-1}, \theta, \xi\right) = P_{temp}\left(T_i \mid k_i, l_i, F_{i-1}, \theta, \xi\right) \cdot P_{spat}\left(k_i, l_i \mid F_{i-1}, \theta, \xi\right) \ , \tag{5}$$

where both spatial and temporal components are conditional on all preceding fixations F_{i-1}, model parameters θ and internal degrees of freedom ξ that generate model stochasticity.

The internal degrees of freedom ξ are due to the unknown states of the random walks governing target selection and saccade programming. This results in stochastic values that are obtained for multiple evaluations of the likelihood function. In principle, we could overcome the stochasticity via averaging, which is, however, computationally costly. Moreover, previous work indicated that stochasticity of the likelihood is effectively averaged out over the evaluations generating the Markov chain, if the likelihood of the previously accepted proposal is re-evaluated every time; this approach is is denoted as *pseudo-marginal likelihood* (Andrieu & Roberts, 2009). While the spatial likelihood P_{spat} is available in closed form and exact, it does depend on the stochastic word activations, thus, the pseudo-marginal approach is used here.

The spatial likelihood P_{spat} is further decomposed into the probability q to land on word k_i and letter l_i within word k_i after having selected word m with selection probability[3] π following the initiation of a saccade at time T_i (see Eq. 6). For an observed fixation i, it is unknown which word was the intended target word. Therefore, P_{spat} equals the probability of landing on (k_i, l_i), integrating the product of word targeting probability $\pi(m|.)$ and oculomotor error probability $q(k_i, l_i|m, .)$ by summation over all words m of the sentence $(m = 1, 2, ..., N_W)$, i.e.,

$$P_{spat}\left(k_i, l_i \mid T_i, F_{i-1}, \theta, \xi\right) = \sum_{m=1}^{N_W} \pi\left(m \mid T_i, F_{i-1}, \theta, \xi\right) \cdot q\left(k_i, l_i \mid m, F_{i-1}, \theta\right) \tag{6}$$

According to the SRE model of saccade amplitudes (McConkie et al., 1988), the systematic component ϵ_{sre}, Eq. 14, mainly shifts the mean landing position and the random component σ_{sre}, Eq. 15, modulates the variance of the distribution of landing positions (see Appendix B for mathematical details). While the selection probability $\pi(.)$ in SWIFT is driven by a time-dependent word activation field, the observable landing position, or its probability $q(.)$, depends on oculomotor process assumptions and only indirectly on the implementation of the word activation field.

The oculomotor assumptions, explicitly given by the probability $q(.)$, Eq. (6), strongly influence model performance, in particular, if difficult reading conditions with increased refixation and regression probabilities are investigated. Previous parameter estimations using the Gaussian saccade model (cf. Engbert et al., 2005) did not fit the shape of the bimodal saccade amplitude distributions satisfactorily, in particular for refixation with very short shifts of the gaze position (see Figure 3). The fit was particularly concerning for the reverse letter and mirrored words conditions investigated in this article. Therefore, we introduced an optimized oculomotor model within McConkie et al.'s (1988) framework that replaces normal distributions by Gamma distributions to improve model fits (for mathematical details see Appendix B).

Finally, for the temporal probability density P_{temp} in Eq. (6), exact computation is precluded by the complexity of the cascade of random timers. Here, the probability density can be approximated via kernel density estimation (Epanechnikov, 1969), an approach termed *probability density approximation* (Holmes, 2015; Palestro et al., 2018; Turner & Sederberg, 2014).

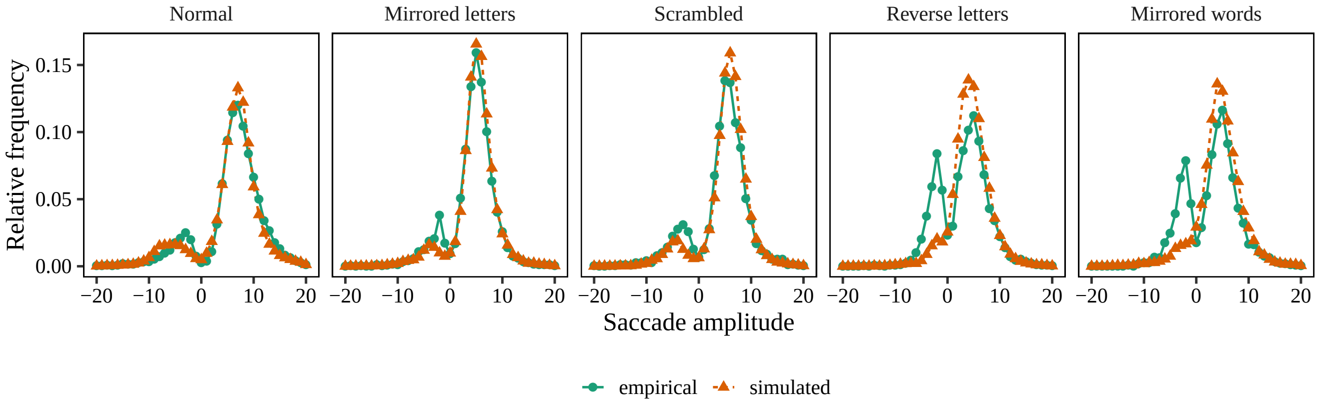
Computational methods

The modified SWIFT model was fitted independently to the available training datasets (see below). For data obtained from each subject in the normal as well as their respective manipulated reading condition, a vector of 15 free model parameters (see Table 2) was sampled using five Markov Chains Monte Carlo (MCMC) runs with 20,000 iterations each. This number of free parameters is, first of all, a computational challenge for numerical simulations, which could, however, be solved in our implementation, since corresponding computer code was implemented parallelization using OpenMP 3.0 in the C programming language.

[3] The selection probabilities π are normalized so that $\sum_{m=1}^{N_W} \pi(m) = 1$.

Figure 3

Empirical and previously simulated Gaussian saccade amplitudes aggregated across all subjects in each experimental condition, including the normal reading condition



Note. For saccade amplitudes generated with improved oculomotor assumptions, see Fig. 10.

Another remark with respect to the number of free parameter seems necessary. We would like to argue that even with 15 free parameters considered here, the SWIFT model should still be perceived as a parsimonious model. We are aiming at reproducing a number of spatial and temporal observables (describing where and how long we fixate) from a single model fit across participants and tasks. Those observables will include three fixation probabilities and four fixation durations as functions of word length, saccade amplitude distributions, and within-word landing positions as a function of launch-site distance. Let us consider the case that all of those observables were analyzed statistically using multiple multivariate regression analyses, for example, this will likely require an approximate number of roughly 20 degrees of freedom. That would include two parameters per fixation probability and fixation duration (each with one intercept and linear slope), three for saccade amplitudes (shape, scale and proportion) and three for within-word landing positions (intercept, linear slope and quadratic slope). From this perspective, the SWIFT model would be more parsimonious in degrees of freedom and offer model parameters which are theoretically motivated and refer directly to specific processes assumed to be underlying reading behavior. Additionally, SWIFT offers explanation for more specific effects, discussed earlier by Engbert et al. (2005), such as the fixation-duration inverted optimal viewing position (IOVP; Nuthmann, Engbert, & Kliegl, 2005; Vitu, McConkie, Kerr, & O'Regan, 2001) or lag and successor effects as indicators for spatially distributed processing (Kliegl, Nuthmann, & Engbert, 2006).

Our Monte Carlo approach was numerically challenging, mainly due to higher dimensionality of the parameter space compared to the previous study (Seelig et al., 2020). After evaluation of different MHMC sampling algorithms, the algorithm tested most convincingly when fitting the SWIFT model was the $DREAM_{(ZS)}$ algorithm (Laloy & Vrugt, 2012; ter Braak & Vrugt, 2008; Vrugt et al., 2009), which we thus used for the present analyses. A modified version of *PyDREAM* (Shockley, 2019), a Python implementation of DREAM$_{(ZS)}$, was implemented in high-performance compute (HPC) facilities. Modifications of the implementation were motivated by the necessity to re-evaluate accepted proposals due to the stochasticity of the pseudo-marginal likelihood. The total computing time amounted to approx1̇0,000 core hours, scaling to 3.5 hours total run time on 72 independent parallel nodes with 40 cores per node.

Bayesian model fitting requires the definition of priors, which are probability distributions describing plausible parameter values. In the SWIFT model, we have expectations on the ranges of plausible parameter values but, due to lack of prior research, no informed knowledge how these expectations would be distributed within those ranges. Therefore, we used weakly informative truncated Gaussian priors with mean μ and standard deviation σ, truncated at $\mu - \sigma$ and $\mu + \sigma$, where the truncation points are equal to the respective range of plausible parameter values and μ to their respective mean (see Fig. 6). We chose truncated Gaussian priors over uniform priors to allow the model to converge on the center of the range of plausible parameter values in the case that the data do not constrain that parameter's marginal likelihood.

Table 2

Fitted SWIFT model parameters

Parameter	Description
β	Word frequency modulation of word difficulty
ω	Global decay during postlexical processing
δ	Processing span in letter spaces
η	Word length modulation of processing rate
t_{sac}	Mean duration of the saccade timer
M	Relative duration of the labile saccade stage for misplaced fixations (i.e., fixations landing on a non-intended target)
R	Relative duration of the labile saccade stage for well-placed (i.e., intended) refixations
$\tau_{n/l}$	Mean durations of the labile and non-labile saccade programs[a]
omn_1	Intercept term for random oculomotor noise[b]
omn_2	Slope term for random oculomotor noise[b]
sre_1	Intercept term for saccadic range error for forward fixations and skippings
$sre_1^{(RF)}$	Intercept term for saccadic range error for refixations[c]
$sre_2^{(FS)}$	Slope term for saccadic range error for forward fixations[c]
$sre_2^{(RF)}$	Slope term for saccadic range error for refixations[c]
$sre_2^{(SK)}$	Slope term for saccadic range error for skippings[c]

[a]Parameters τ_n (for the non-labile stage) and τ_l (for the labile stage) can be defined separately. We chose to couple the parameters so that $\frac{1}{2}\tau_l = \tau_n = \tau_{n/l}$. [b]Parameters omn_1 and omn_2 can be defined separately for each saccade type. All saccade types were assigned the value of the same coupled parameter. [c]Parameters sre_1 and sre_2 can be defined separately for each saccade type. We defined coupled parameters and chose the same value for the mentioned saccade types. For regressions, the parameters were set to $sre_1 = sre_2 = 0$ to disable saccadic range error.

Experiment

In order to demonstrate our approach and validate the model, we chose an experimental study[4] recently published by Chandra et al. (2020), in which experimental conditions were established to induce strong effects on oculomotor control. These effects provide a challenge to model generalizability (due to broad ranges of realized average fixation durations and fixation probabilities) and to interindividual differences.

From each of 36 participants in the experiment, eye trajectories were recorded in a normal reading condition (N) and in one of four manipulated reading conditions with manipulated visual layout. Each of the manipulated reading conditions altered the visual representation of the items by scrambling letters (sL), reversing letter order within the word (iW), mirroring the entire word (mW) or mirroring the individual letters within the word (mL). Table 1 contains example items for each of the experimental conditions. Chandra et al. (2020) showed that the manipulated reading conditions have significant and specific effects on reading, which vary considerably between participants.

Data preprocessing

From the initially recorded data, all trials including blinks were discarded. We used the velocity-based algorithm by Engbert, Sinn, Mergenthaler, and Trukenbrod (2015) to detect saccades and fixations in the raw data. We removed single fixations with durations below 40 ms, landing outside the text rectangle, or shorter than one character space. Trials were cut off after either of the last two words of the item had been fixated, keeping subsequent refixations if any and keeping the full sequence if those words were not fixated at all. Ultimately, trials were excluded if they contained fixations with durations greater than 99.5% of all fixation durations in that experimental condition. We thus excluded trials with fixation durations over 900 ms for normal reading (N), 1605 ms for mirrored letters (mL), 1892 ms for scrambled letters (sL), 2518 ms for inverted words (iW), and 3170 ms for mirrored words (mW).

For each subject in each condition, remaining datasets were split into a fitting (training) and validation (test) dataset. Trials within each dataset were shuffled, keeping the sequence of fixations within a trial intact, and the split criterion incrementally shifted trial by trial until 70% of all fixations for that subject in that condition fell under the criterion. Those were marked as the training dataset to which SWIFT should be fitted. The remaining 30% were marked as the test dataset. This ensured that for each subject and condition there was an approximately equal ratio of data for fitting and for model validation.

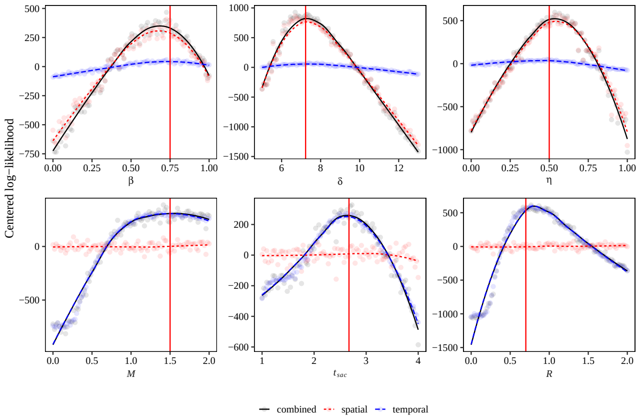
Results

We investigated the SWIFT model for a range of reading tasks using an advanced method for parameter inference. Consequently, our results refer to both methodological and reading-related aspects. First, we present likelihood profiles to demonstrate the validity of the likelihood function. Next, we simulate data using the SWIFT model and investigate parameter recovery based on our methods to check the identifiability of model parameters. Second, we present summary statistics for the experimental results (Chandra et al., 2020) and apply SWIFT parameter inference to the corresponding fixations sequences in the training data set. Posterior predictive checks are obtained for the posteriors on model parameters applied to the test data. Finally, we present a statisti-

[4]The experimental data are available at `https://osf.io/bmvrx/`

Figure 4

Centered likelihood components for selected model parameters



Note. Each point represents one likelihood evaluation. Different colors are the different likelihood components (spatial and temporal) and their product (combined likelihood). Red vertical lines are true parameter values of the simulated dataset. Note that likelihood evaluations (dots) are stochastic. Smooth lines are included in this figure only to enhance visibility of the underlying trends.

cal analysis of model parameter estimates across participants and experimental conditions.

Likelihood profiles

We start our analyses with a numerical test of the likelihood function. Likelihood profiles are generated by varying only one of the model parameters along an informative interval while holding constant all other parameters. The resulting likelihood curvature should be dominated by the effect of the varied parameter. For a simulated dataset with known (true) parameter values, we evaluate a spatial and a temporal likelihood component (L_{spat} and L_{temp}, respectively, Eq. 5) as well as the combined likelihood (L_M), which is exclusively used for further fitting purposes. As shown below, likelihood profiles (i) peak at the true value, (ii) have identical maxima for simulated data, and (iii) show selective influences on the spatial vs. temporal component for parameters that are designed to have predominantly spatial vs. temporal effects. Severe divergence would necessitate a revision of the likelihood function, which is not the case here (see Figure 4).

Parameter recovery

While the inspection of likelihood profiles validates the likelihood itself, a parameter recovery study additionally validates the sampling procedure, which is another necessary precondition for fitting the model to empirical data. We generated 48 datasets for which the selected parameter values (i.e., the "true" values for the recovery analysis) were randomly and independently sampled from the chosen prior distributions; we assumed uncorrelated parameters for this analysis.

We fitted the model to each generated dataset, using the same priors. Subsequently, we calculated 60% highest posterior density intervals (HPDIs) for each parameter and dataset in order to evaluate whether the true value was recovered, i.e., included in the credible interval. As can be seen in Figure 5, most true values are recovered reliably. Parameters β, η, $\log_{10}\omega$, and δ appear to have a somewhat systematic bias, possibly due to their interactions with other model parameters. Fitted parameter values in biased regions should therefore be interpreted with caution. Overall, however, these results lend support to the computational faithfulness of the model and the method of statistical inference. We therefore proceed to fitting the model to the empirical datasets.

Experimental data: Summary statistics

In Table 3, we report summary statistics derived from the experimental data published by Chandra et al. (2020). The manipulated reading conditions are associated with significantly different patterns in fixation probabilities and durations compared to the normal reading condition. Moreover, high standard errors (in parentheses) suggest high between-subject variability overall, in particular, in the manipulated reading conditions. Thus, the experimental data pose a challenge for our mathematical model.

With regard to accuracy in response to the comprehension questions asked after each session, subjects answered an average of 2.58 ($SE = 0.115$) out of three correctly in the normal reading (N) condition. In the manipulated reading conditions, those were 2.44 ($SE = 0.176$) for mirrored letters, 2.11 ($SE = 0.261$) for scrambled reading, 2.89 ($SE = 0.111$) for reversed letters, and 2.78 ($SE = 0.147$) for mirrored words. A linear regression analysis indicated none of these as statistically different from the accuracy observed in the normal reading condition.

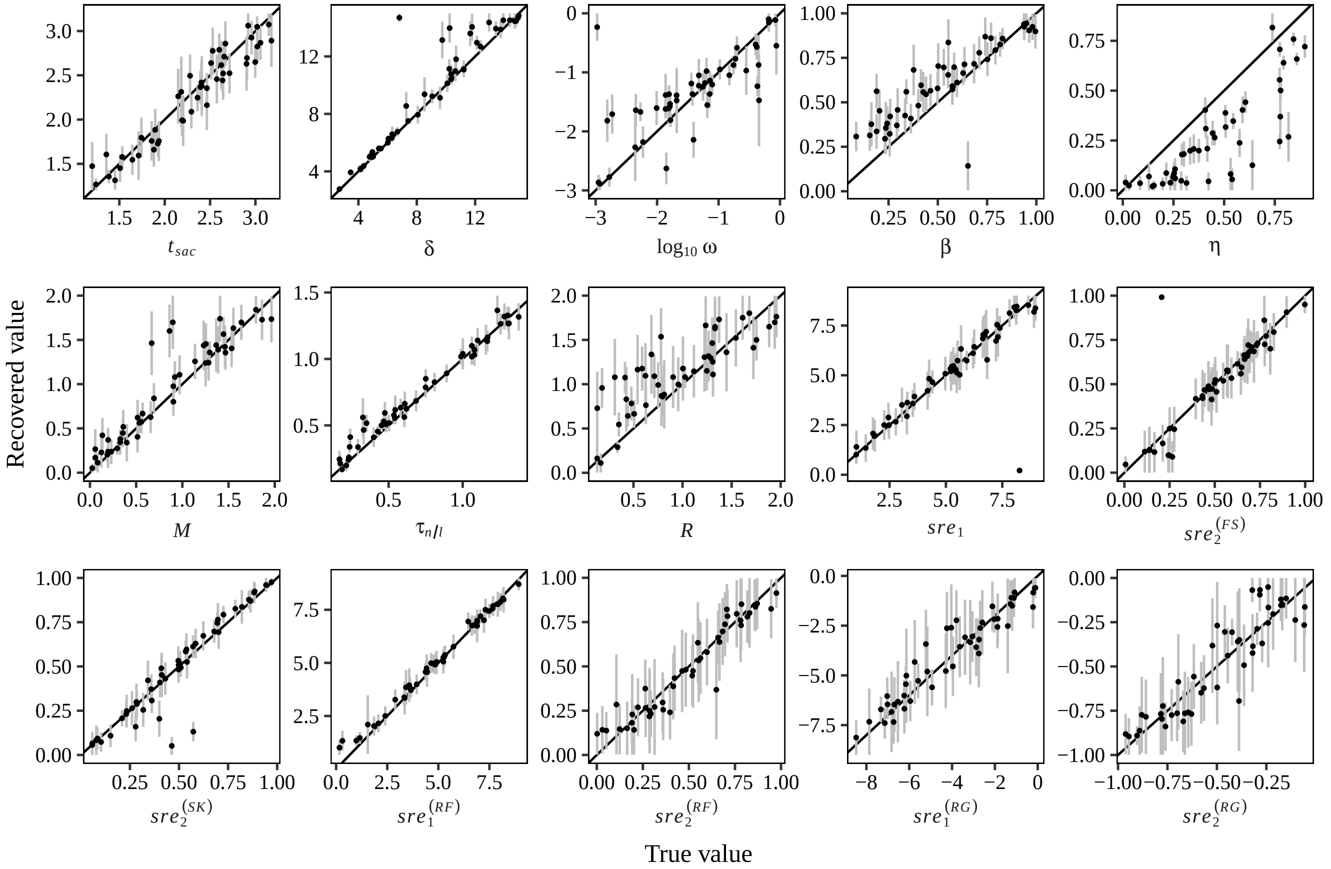
Parameter estimates

For every participant of the experiment, posteriors were generated using MCMC sampling in both the normal and a manipulated reading conditions. In Figure 6, all samples were aggregated across subjects for the five experimental conditions. The corresponding distributions indicate how the posteriors deviate, on average, between experimental conditions. It appears that some model parameters (e.g., $sre_1^{(RF)}$) converge on similar values, while others (e.g., δ) differ quite substantially between experimental conditions.

While the likelihood-based Bayesian inference provides an objective approach to statistical inference on model parameters, it is important to note that the convergence of parameters to specific posterior distributions does not prove the

Figure 5

Scatterplot of true and recovered parameters with 60% HPDIs (vertical grey lines) across simulated data sets



Note. The diagonal line indicates identity, i.e., credibile intervals touching the diagonal include the true value.

model's adequacy in terms of experimentally observed effects. Therefore, the numerical computation of posteriors needs to be combined with an analysis of the model behavior with respect to relevant characteristics of fixation sequences.

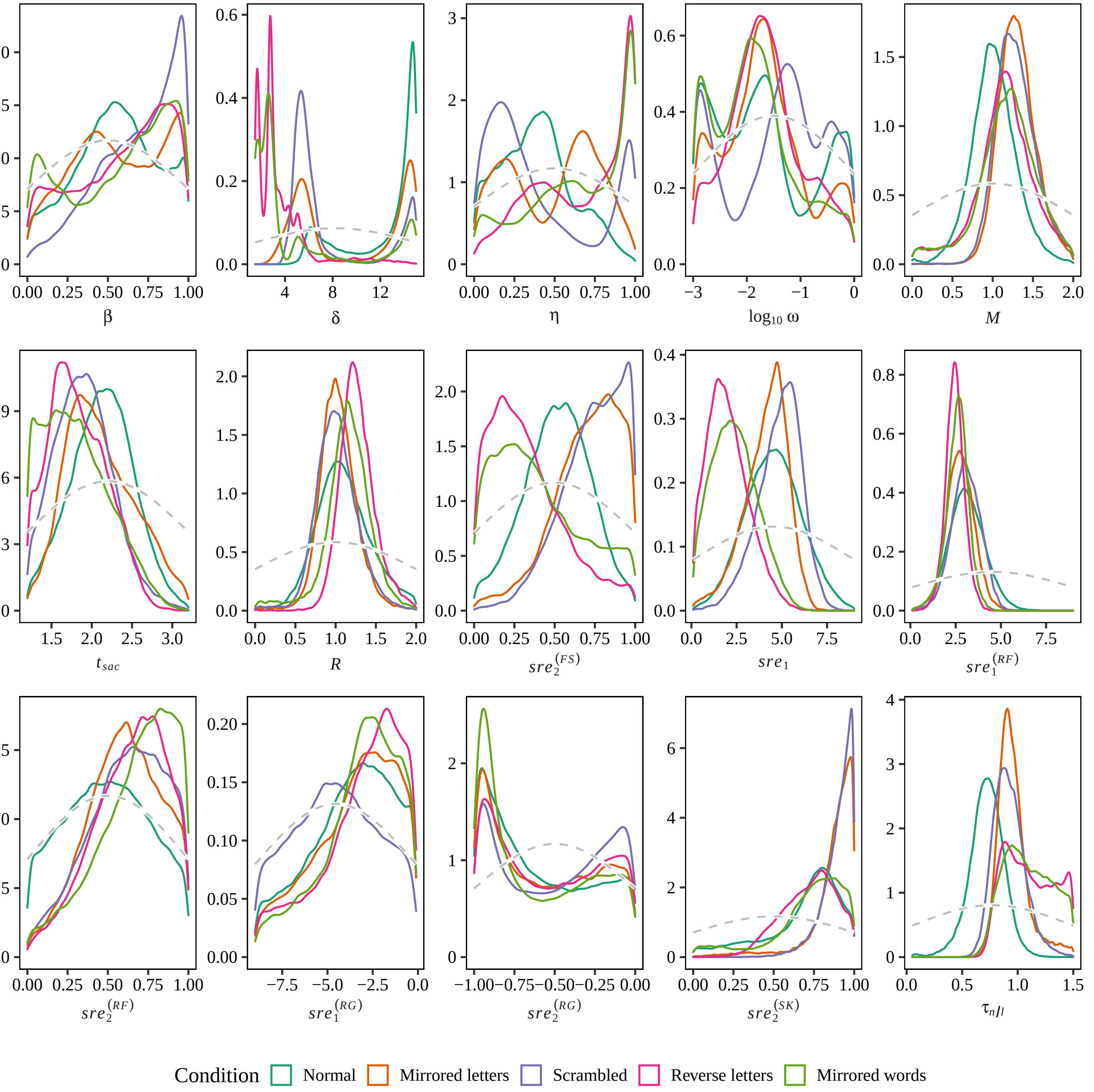
Posterior predictive checks

Next we validate that the estimated parameters drive the model's behavior into psychologically plausible regimes and, thus, provide an explanation for reading behavior across experimental conditions. These *posterior predictive checks* can be accomplished by cross-validation. Having fitted the model to a portion of the data (training dataset) only, we can compare summary statistics of derived model simulations to the remaining experimental data (i.e., test or validation datasets).

For each subject, we obtained empirical summary statistics from the observed (empirical) eye trajectories of the test dataset and simulated summary statistics from eye trajectories generated for the same trials. Instead of using point estimates for the validation checks, we randomly sampled parameter configurations from the posterior parameter distributions. For each subject and condition, 20 distinct parameter configurations θ were randomly sampled from the respective posterior distribution, i.e., the fitted posterior for that subject in that condition. For each sampled θ, fixation sequences were generated for the trials previously withheld from fitting. Simulated summary statistics were derived as the average of each respective summary statistic across simulated datasets for each respective subject and condition. We employed this technique in order account for the full covariance structure of the parameter distributions and thus the full range of plausible model behavior. As can be seen in Table 4 and Figure 7, the mean squared error across subject-level summary statistics is considerably reduced for most of the combinations of dependent variables and conditions. As a result, simulated quantities more closely approximate the empirical summary statistics when sampling parameter combinations from the full posterior than when using point estimates for each parameter.

Figure 6

Posterior densities for all fitted model parameters



Note. Colored lines represent the different experimental conditions. Each line is aggregated across all subjects in that condition, i.e. $N = 36$ for the baseline, normal-reading condition (N) and $N = 9$ for the other four conditions. Dashed gray lines are prior distributions, which were identical for all subjects and conditions.

Table 3

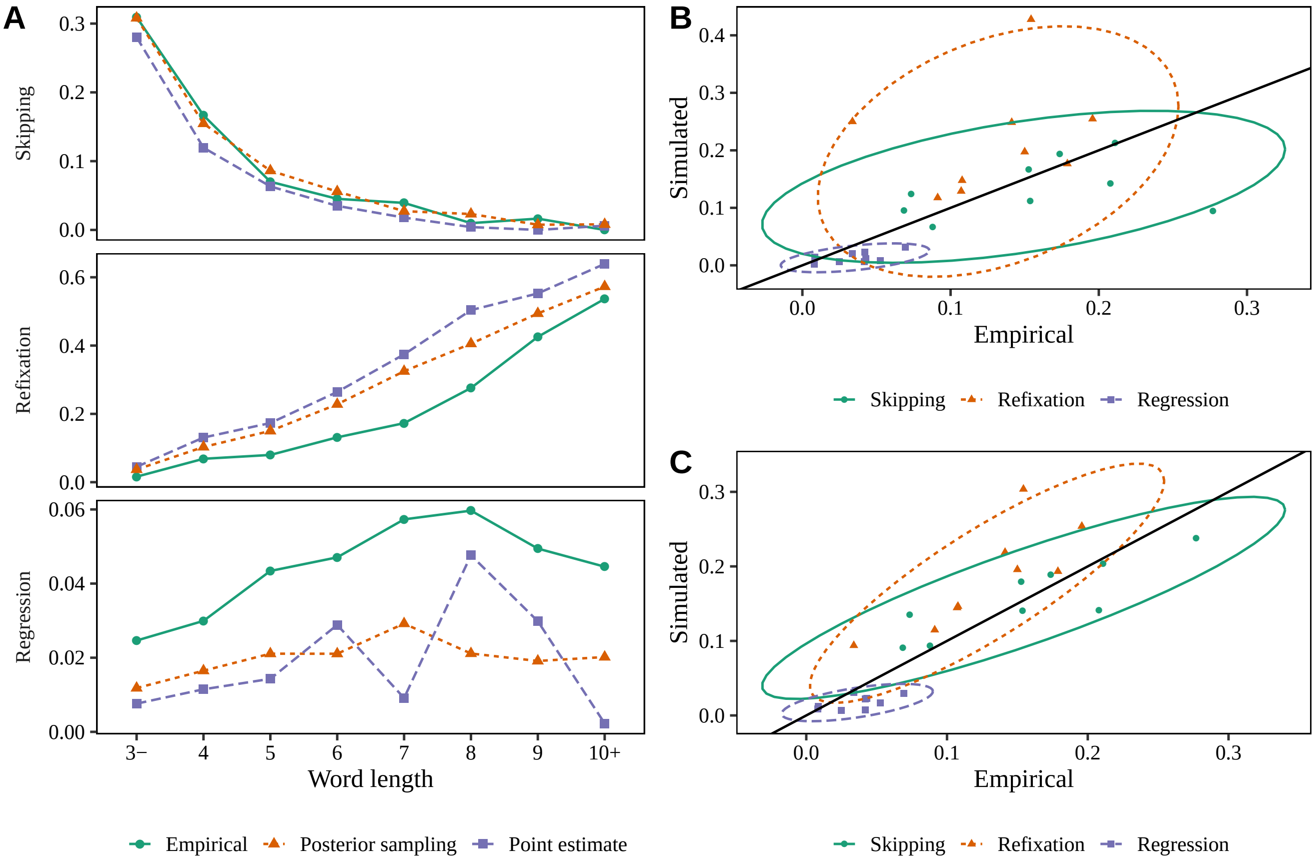
Empirical means and standard errors in summary statistics aggregated across subjects

	Normal	Mirrored letters	Scrambled	Reversed letters	Mirrored words
Fixation probabilities					
Regression	.035 (.004)	.036 (.007)	.040 (.009)	.029 (.006)	.057 (.013)
Refixation	.097 (.006)	.203 (.018)	.148 (.020)	.246 (.044)	.240 (.039)
Skipping	.267 (.011)	.156 (.024)	.181 (.018)	.079 (.019)	.113 (.023)
Fixation durations					
Gaze	287.4 (5.7)	455.5 (31.7)	414.7 (30.2)	843.0 (67.5)	885.4 (68.1)
First-fix.	251.1 (5.6)	313.8 (20.8)	289.1 (19.2)	538.9 (57.4)	546.6 (73.2)
Refixation	224.7 (6.0)	311.3 (21.9)	315.6 (23.9)	504.5 (59.1)	520.6 (50.6)
Single-fix.	248.2 (5.7)	313.5 (21.1)	287.4 (18.3)	540.5 (57.4)	539.6 (68.3)

Note. Estimates are means of fixation probability or duration subject means with standard errors in parentheses.

Figure 7

Comparison of simulated summary statistics when sampling from the posterior vs. using point estimates



Note. Panel A shows summary statistics as a function of word length. Panel B and C show between-subject variability for point estimates and posterior sampling, respectively. All panels refer to validation results for the mirrored-letters condition (mL).

Table 4

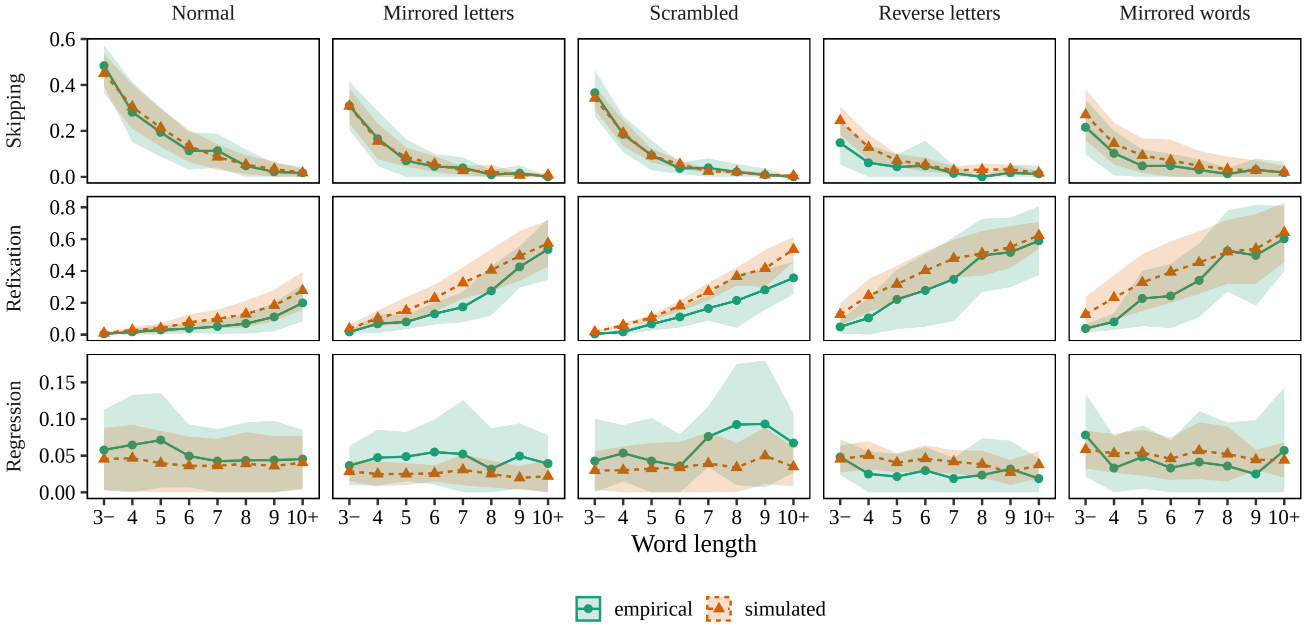
Change in MSE across subject-level summary statistics between posterior sampling and point estimates

	Normal	Mirrored letters	Scrambled	Reverse letters	Mirrored words
Fixation probabilities					
Skipping	−48.0%	−73.7%	−11.0%	+21.7%	−64.8%
Refixation	−40.6%	−68.3%	−37.2%	−14.7%	−32.4%
Regression	−39.6%	−21.3%	−35.6%	−45.0%	−18.1%
Fixation durations					
First-fix.	+36.7%	+17.7%	−60.5%	−89.7%	−99.2%
Refixation	−46.0%	+59.9%	−13.7%	−83.3%	−98.8%
Gaze	−12.5%	−80.2%	−66.2%	−89.8%	−97.6%
Single-fix.	+17.3%	+15.2%	−62.5%	−89.8%	−99.1%

Note. Negative percentages are reductions of the mean squared error (MSE) when using posterior sampling relative to the MSE when using point estimates. Positive percentages are increases.

Figure 8

Empirical and simulated spatial summary statistics *(fixation probabilities) for different experimental conditions, aggregated across subjects, as a function of word length*



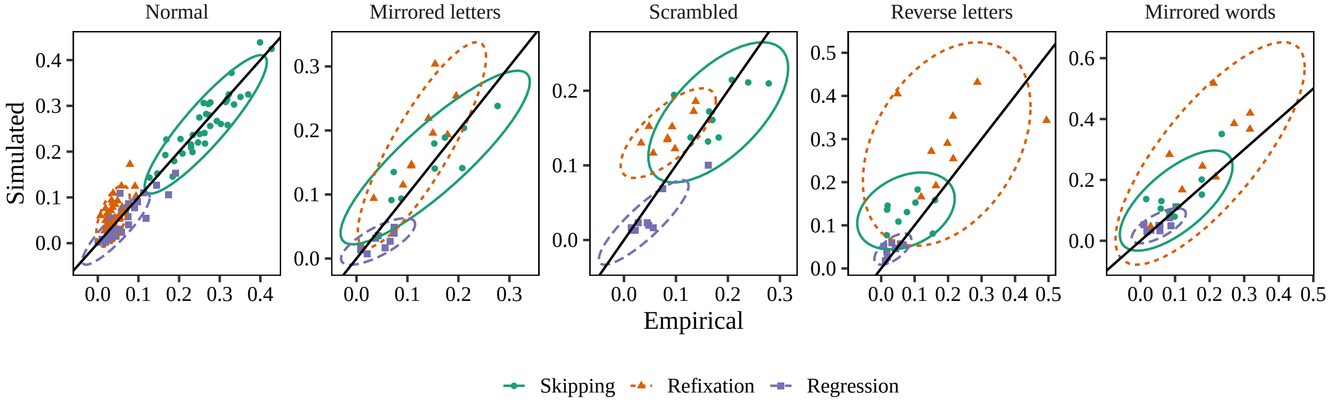
Note. Upper and lower bound of the shaded area is one standard deviation around the plotted mean, measuring the variability in subject means.

Spatial summary statistics. For the evaluation of spatial aspects of model validation, we analyze fixation probabilities. In most summary statistics, SWIFT can reliably reproduce different reading characteristics. Especially notable are skipping and refixation probabilities in all experimental conditions (see Figure 8), including word length effects on those at a global level. There is, however, still some divergence with regard to regression probabilities, namely that the models predicts too few regressions, in particular, in the mirrored letter and scrambled letter conditions.

To analyze that the model captures and reproduces between-subject variability, we used scatterplots and correlation analyses of summary statistics across subjects between simulated and experimental data. A significant correlation can be interpreted as statistical evidence that the approach was successful with regard to the respective summary statistic. According to this criterion, spatial summary statistics are reliably reproduced for the set of participants. As can be seen in Figure 9 and Table 5, most conditions, the averages across subjects (ellipsis midpoints) correlate closely across statis-

Figure 9

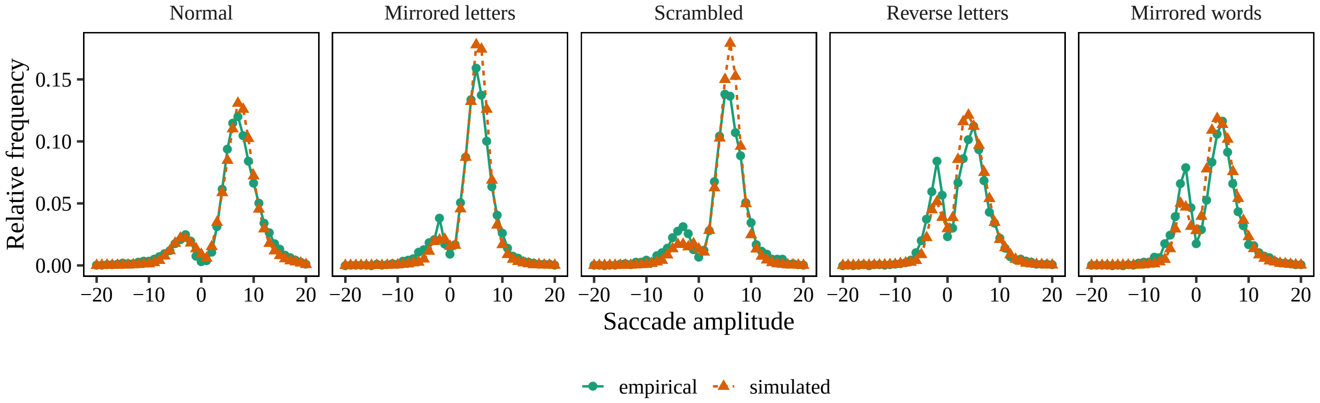
Correlation between empirical (horizontal axis) and simulated (vertical axis) spatial summary statistics *(fixation probabilities)*



Note. Each subject is represented by one dot in each color in the respective experimental condition (panel).

Figure 10

Empirical and Gamma-distributed simulated saccade amplitudes aggregated across all subjects in each experimental condition, including the normal reading condition



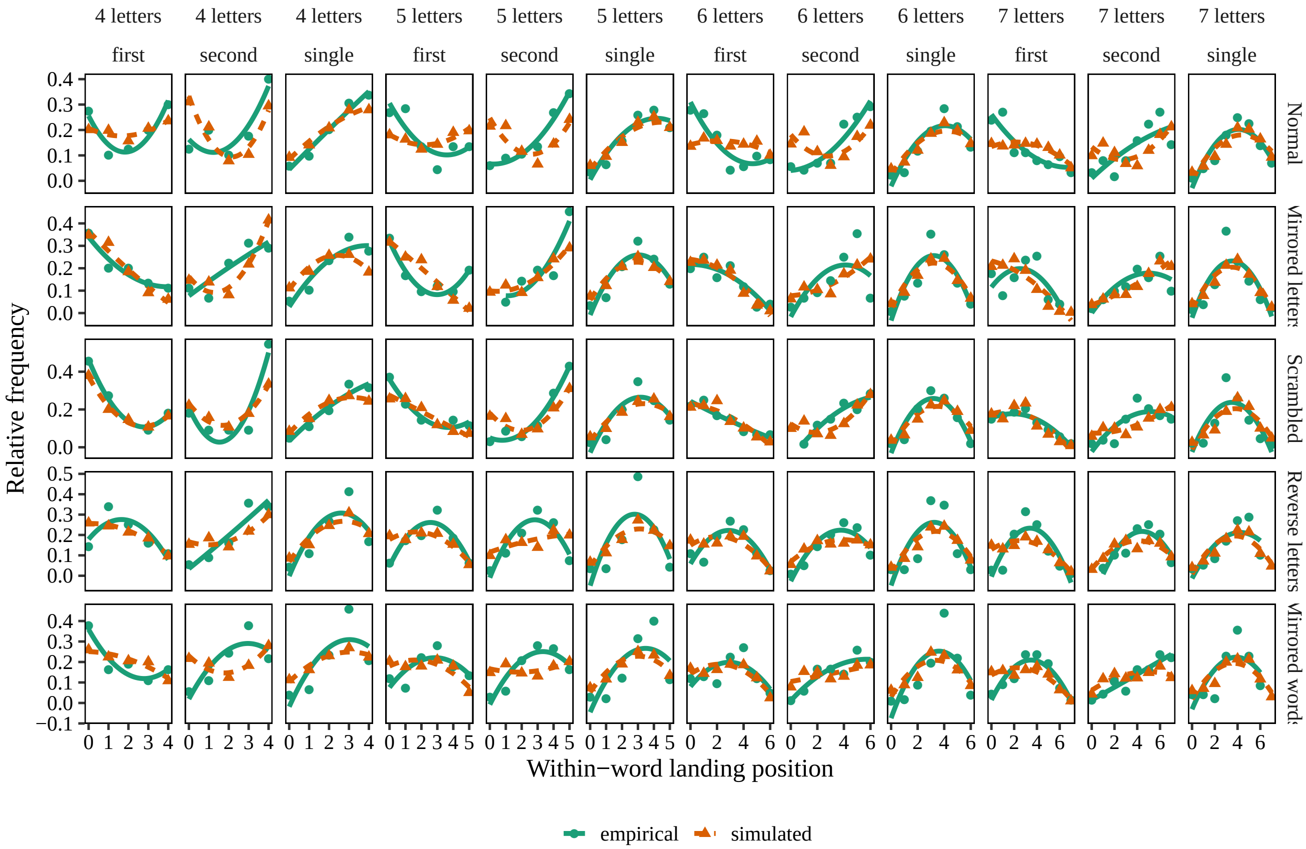
tics and the subject-level variance (covariance within each ellipsis) is captured very reliably.

Moreover, results for saccade amplitudes are clearly important (see Figure 10), supporting the notion of Gamma-distributed saccade length distributions. In contrast to the previous Gaussian saccade amplitudes (see Figure 3), the bimodality of the distribution is clearly visible for all experimental conditions. Interestingly, the model can even capture differences between experimental conditions, with saccade amplitudes being more widely spread in the right-to-left reading conditions compared to the baseline or other left-to-right conditions. Figure B1 in Appendix B also shows a comparison to previously Gaussian distributed saccade amplitudes, which fit the data less satisfactorily. As depicted in Figure 11, the model can also capture and reproduce word length effects on within-word landing positions.

Temporal summary statistics. Similarly for temporal summary statistics, global averages and slight word-length effects are reproduced quite reliably for the test datasets in fixation durations. As for spatial summary statistics, there is some divergence for the right-to-left conditions (reverse letters and mirrored words, see Figure 12). When compared at a by-subject level (see Figure 13 and Table 5), it is clear that for most conditions, the model can again successfully replicate different temporal reading measures. This can most clearly be seen for fixation durations in the normal reading condition (N).

Figure 11

Empirical and simulated landing positions for single fixations, first fixations, and second fixations



Note. Data are aggregated across all subjects in each experimental condition (row). Lines represent quadratic regression fits of the displayed aggregated data points.

Statistical evaluation of model parameters

The modeling of interindividual differences permits a new analysis for cognitive models of eye-movement control, since we are able to observe the specific responses of participants to experimental conditions. We carried out a multiple multivariate linear regression analysis (see Figure 14) for model parameters to statistically infer how and which aspects of the reading manipulations caused which type of change in reading pattern.

As linear regressions were conducted by model parameter, to control for multiple testing, p-values were corrected according to Šidák (1967), denoted by p_S. In order to test how specific characteristics of the experimental manipulations had an effect on model parameters, we tested four null hypotheses, from which we derived a contrast matrix for regression analysis using the *hypr* package (Rabe, Vasishth, Hohenstein, Kliegl, & Schad, 2020; Schad, Vasishth, et al., 2020) in the R programming language. The tested null hypotheses are given as

$$\begin{aligned}
H_{0_1} &: \quad \mu_{mL} = \mu_N \\
H_{0_2} &: \quad \mu_{iW} = \mu_N \\
H_{0_3} &: \quad \mu_{mW} = \mu_N + (\mu_{mL} - \mu_N) + (\mu_{iW} - \mu_N) \\
H_{0_4} &: \quad \mu_{sL} = \mu_{mL} \ ,
\end{aligned}$$

where each null hypothesis relates to one contrast in a linear regression model. H_{0_1} and H_{0_2} test the effects of letter flipping (mL, mirrored letters condition) and word inversion (iW, reverse letters condition), respectively, with regard to the baseline. H_{0_3} tests whether the mirrored words condition (mW), which combines characteristics of letter flipping and word inversion, is different from an addition of the effects of the letter-flipping (mL) and word inversion (iW) conditions to the baseline. As scrambled reading only shares reading direction (i.e., whether letter sequences have been inverted or not) with the letter-flipping condition (mL) but no other characteristics with any other condition other than the baseline,

Figure 12

Empirical and simulated temporal summary statistics *(fixation durations) for different experimental conditions, aggregated across subjects, as a function of word length.*

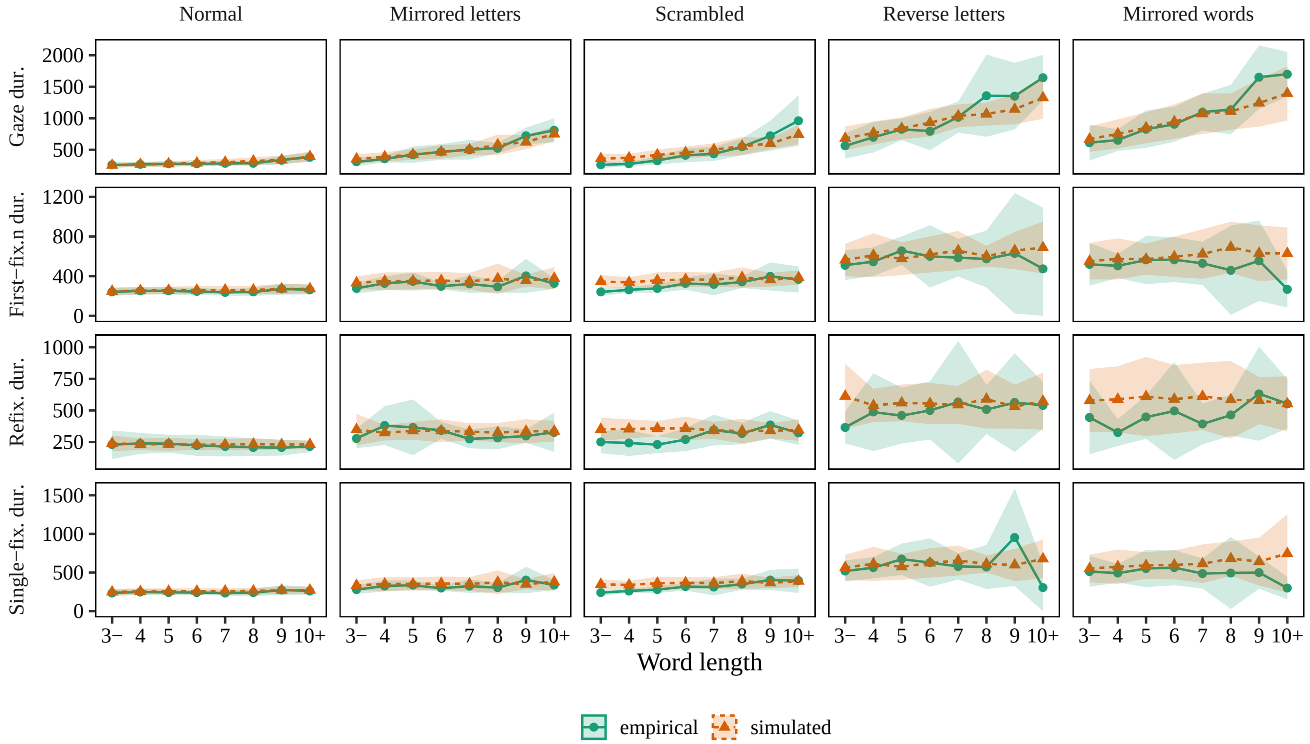
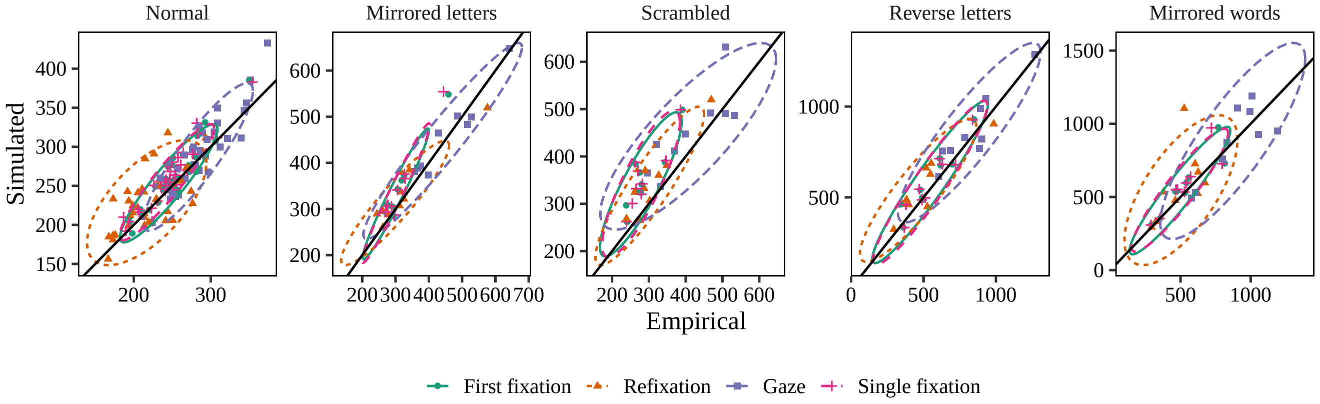


Figure 13

Correlation between empirical (horizontal axis) and simulated (vertical axis) temporal summary statistics *(fixation durations)*



Note. Each subject is represented by one dot in each color in the respective experimental condition (panel).

Table 5

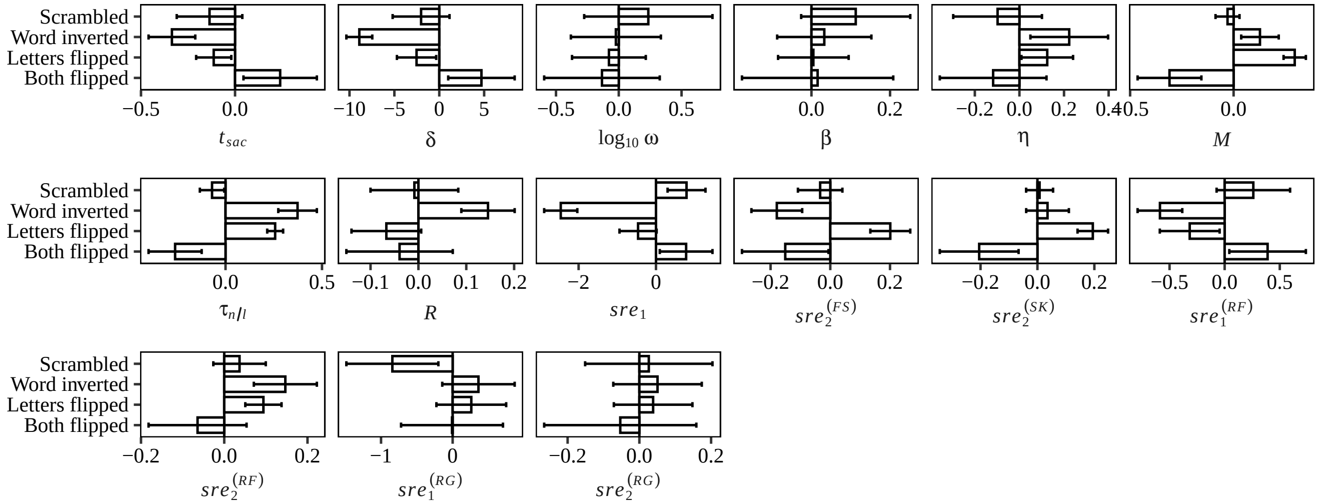
Correlations across subjects for empirical vs. simulated summary statistics

	Normal	Mirrored letters	Scrambled	Reverse letters	Mirrored words
Fixation probabilities					
Skipping	**0.91** (.001)	**0.86** (.004)	0.52 (.148)	0.20 (.605)	0.76 (.018)
Refixation	**0.65** (.001)	**0.81** (.009)	0.70 (.038)	0.27 (.485)	0.73 (.026)
Regression	**0.88** (.001)	0.74 (.023)	**0.93** (.001)	0.59 (.094)	0.70 (.037)
Fixation durations					
First-fix.	**0.94** (.001)	**0.98** (.001)	**0.89** (.002)	**0.95** (.001)	**0.92** (.001)
Refixation	**0.63** (.001)	**0.97** (.001)	**0.93** (.001)	**0.91** (.001)	0.54 (.134)
Gaze	**0.90** (.001)	**0.96** (.001)	**0.81** (.009)	**0.91** (.001)	**0.81** (.009)
Single-fix.	**0.94** (.001)	**0.98** (.001)	**0.87** (.003)	**0.95** (.001)	**0.90** (.002)

Note. Estimates are two-sided Pearson correlation coefficients with p-values in parentheses (bold font for $p < 0.01$).

Figure 14

Linear regression results for model parameters



Note. Horizontal bars are estimated coefficients as effects vs. the normal reading condition (intercept). Error bars are uncorrected 95% confidence intervals around the estimated effects. The baseline is tested against zero, while contrasts are tested against the baseline.

H_{04} was formulated to test whether the effects on model parameters of scrambled reading are statistically distinct from the mirrored letters condition (mL).

Effects of inverting words. The inversion of the sequences of letters within words is associated with a narrower processing span δ ($b = -8.91, p_S < 0.001$) compared to normal reading. A reduced processing span is psychologically plausible because of the higher visual difficulty. The reduced processing span is associated with smaller optimal saccade amplitudes for forward fixations and skippings sre_1 ($b = -2.46, p_S < 0.001$) as well as refixations ($sre_1^{(RF)}$, $b = -0.59, p_S < 0.003$), which contribute to the reduced average saccade length.

Moreover, saccade execution is less sensitive to the actual target distance for refixations $sre_2^{(RF)}$ ($b = 0.15, p_S < 0.044$) compared to refixations in normal reading but more sensitive for forward saccades $sre_2^{(FS)}$ ($b = -0.18, p_S < 0.030$). This indicates more well-placed forward fixations and fewer well-placed refixations than in the normal reading condition. In default of knowing which empirical fixation is well-placed or not, such a pattern is difficult to test experimentally. Nevertheless, it could indicate that the certainty about a word's location before it is fixated is higher than in normal reading, possibly due to the many refixations. However, once it has been fixated, the difficulty of the manipulation decreases the certainty for the optimal within-word target location below the level observed in normal reading.

With regard to the timing of saccades, the global timer t_{sac} is shorter than in the baseline ($b = -0.34, p_S < 0.002$), which in itself would cause more frequent saccades. However, longer labile and non-labile saccade programs $\tau_{n/l}$ ($b = 0.37, p_S < 0.001$) can counteract this effect, as the global timer reaching threshold during the labile saccade stage can cancel the saccade and actually cause a longer fixation duration. In addition, the global timer is significantly slower than in the normal reading condition (parameter R, $b = 0.15, p_S < 0.007$), leading to longer refixation durations in relation to baseline fixation durations.

Effects of letter flipping. Analogous to inverting letter sequences, the horizontal flipping of letters at their respective (normal or inverted) location is also associated with longer labile and non-labile saccade programs $\tau_{n/l}$ ($b = 0.26, p_S < 0.001$). However, effects on δ, sre_1, $sre_1^{(RF)}$ or t_{sac} are not significant.

In contrast to the inversion of letter sequences, flipping letters, however, causes the global saccade timer to slow down after misplaced fixations (parameter M, $b = 0.30, p_S < 0.001$). Potentially related to this is the less precise execution of saccades to intended forward fixations, skippings, or refixations, as suggested by greater SRE slopes and thus reduced oculomotor control, $sre_2^{(FS)}$ ($b = 0.20, p_S < 0.002$), $sre_2^{(SK)}$ ($b = 0.19, p_S < 0.001$), and $sre_2^{(RF)}$ ($b = 0.09, p_S < 0.017$), respectively.

Interactive effects. When reversed letter sequence and mirrored letters are combined, i.e., the word is mirrored as a whole rather than by letters individually, most model parameters are affected additively, given that the interaction terms are not statistically significant. In two timing parameters, however, there were significant interaction effects. Significant interactions in M ($b = -0.31, p_S < 0.010$) and $\tau_{n/l}$ ($b = -0.26, p_S < 0.020$) effectively cancel out the magnitude of the effect of mirrored letters on those parameters. This result might indicate that the presence of reversed letter order overrides the effects of mirroring letters in terms of saccade timing.

Effects of scrambling words. None of the effects of scrambled letters on the model parameters was significant. Given the null hypothesis comparing against letter flipping, this means that there is no statistical evidence for scrambling letters being different from letter flipping with regard to SWIFT model parameters.

Discussion

Following a principled Bayesian workflow, we fitted the SWIFT model (Engbert et al., 2005) in a new version with oculomotor improvements to experimental data from 36 subjects who read text in a baseline (control) condition and in four different reading conditions with manipulated text layout.

Our approach is fundamentally based on a recently proposed likelihood function for the SWIFT model (Seelig et al., 2020), which is a prerequisite for Bayesian inference. This is a major advance compared to earlier parameter fitting based on *ad-hoc* discriminating statistics, which were mainly taken over from experimental research and not theoretically motivated (Engbert et al., 2005). The lack of objective statistical treatment is characteristic for the field of dynamical eye-movement modeling. For example, the E-Z Reader model (Reichle et al., 1998) has been investigated in the context of different reading and non-reading tasks (Pollatsek, Reichle, & Rayner, 2006; Reichle et al., 2012), however, without objective statistical parameter inference. Therefore, the latter results can be interpreted as viability tests rather than statistically approved evidence. In our current approach, however, models are fitted on the basis of an objective likelihood and summary statistics are not used as optimization targets but as model validation criteria after parameter fitting.

We demonstrated that model parameters could be estimated reliably—even after splitting data into training and test data. While interindividual differences are an important topic in eye movement research during reading, so far dynamical cognitive models could not be fitted to individual datasets. Therefore, our results suggest that the Bayesian approach will strengthen cognitive modeling of eye-movement control to include the prediction of interindividual differences.

As a first step, we investigated computational faithfulness of the model by examining likelihood profiles and recovering known (true) parameter values from simulated data. The results indicated that the likelihood and sampling algorithm converges reliably for almost all model parameter and thus yielded plausible credible intervals. Recovery studies for model parameters represent a substantial progress to the field of cognitive modeling of eye-movement control (cf. Engbert et al., 2005).

Next, the model was fitted to individual data in the predefined training dataset. To investigate whether the estimated parameters can in fact account for the observed behavior, we simulated eye trajectories for the withheld test subsets and compared summary statistics between empirical and simulated data. The presented temporal and spatial summary statistics (fixation durations and fixation probabilities, respectively) indicate a convincing model fit to the data. In particular, in the normal reading condition and those with normal reading direction (letter flipping, mL, and scrambled letters, sL), the model was shown to predict empirical fixation durations and probabilities very reliably, across groups and subjects.

An important improvement of the current computational approach relates to a balance between underlying cognitive and oculomotor processing. While earlier computational models were in a first step ignoring oculomotor processes (e.g. Engbert et al., 2002; Reichle et al., 1998) and later extended to include oculomotor variability (Engbert et al., 2005; Reichle, Rayner, & Pollatsek, 1999), our approach is fully integrating oculomotor and cognitive models on the level of parameter inference. This might be a promising approach to future integration of further processes, e.g., word recognition (Snell et al., 2018) or higher-level language processing (Reichle, Warren, & McConnell, 2009). We suppose that such an integration will improve the predictive and explanatory power in various facets of the model dynamics, in particular with regard to regressive saccades, as those may be partly triggered by top-down linguistic processes (Engelmann, Vasishth, Engbert, & Kliegl, 2013) in addition to baseline regressions observed even during scanning of meaningless strings (Nuthmann & Engbert, 2009).

In general, we observed that the high reliability is partly achieved by simulating behavior for different parameter configurations sampled from the fitted posterior distributions rather than using only point estimates. This approach makes use of the distributional properties of the fitted model parameters such as their covariance structure. Consequently, parameter configurations that were used for simulating fixation sequences were in their entirety more representative of the range of explainable behavior under the model assumptions.

Given that the model can capture the differences in summary statistics between reading conditions and that all model parameters are theoretically motivated, the differences in model parameters between experimental conditions can help explain why reading behavior differs between those. Essentially, this approach is similar to statistical models such as regression models in which the parameters are effects on the dependent variable. In this approach, however, the parameters are directly related to the assumed underlying cognitive processes and their variability.

Our results also provide specific insights into the reading patterns for manipulated text layouts. In an analysis of model parameters between experimental conditions, we observed statistically significant changes in model parameters that indicate distinct adaptations to processing demands as well as temporal control of fixation duration and oculomotor errors. Inverting letter sequences is associated with a significant reduction of the processing span, which is a psychologically plausible adaptation that leads to a reduced average saccade length and is related to other, more specific changes. This prediction could be tested in experiments using the moving window paradigm (see Starr & Rayner, 2001, for an overview). Similarly, letter flipping slows the saccade timer and produces a number of other effects, which can be mainly associated with an increased processing difficulty and heightened uncertainty about word locations. Our results also indicated two significant interactions of letter flipping and reversed letter sequences (i.e., flipping the word as a whole) on model parameters, suggesting that the presence of both manipulations may lead to the effects of letter flipping being overridden by the effects of reversed letters. Interestingly, the well-known scrambled-letter manipulation is largely similar to the letter-flipping condition or at least not significantly different.

For future modeling work, it is important to note that we have not yet taken advantage of hierarchical modeling techniques. We expect that a hierarchical Bayesian approach will noticeably improve model fits, especially for cases in which less data was available due to exclusions etc. In addition, as hierarchical models are fitted to all subjects in concert, it would be possible to reduce degrees of freedom by limiting the number of parameters varying between subjects. Due to the stochasticity of the likelihood function, however, numerical MCMC algorithms are related to a subset of MCMC methods. For example, gradient-based MCMC methods such as Hamiltonian Monte Carlo (HMC) are precluded in the current model formulation (Seelig et al., 2020).

In the scope of this research, we make predictions for data the model has not been fitted to as part of the model validation procedure. Future research should evaluate how reliable predictions are for unseen experimental conditions and subjects, e.g., by first predicting parameters based on pooled inferences of a subject's behavior in other conditions and/or other subjects' behavior in the condition to be predicted and subsequent model simulations for validation. Our regression analyses could in principle be used to predict model parameter values for a subject and/or condition and these should subsequently be used to simulate trials, from which summary statistics can be derived and compared to withheld data. The successful posterior predictive checks and other validity checks suggest that this is generally possible. However, it should be noted that our fitted and "predicted" data originate from each respective same subject and condition.

To conclude, we presented results from an improved version of the SWIFT model, evaluated against a challenging data set, and fitted along a Bayesian workflow. The Bayesian approach turned out to be sensitive enough to reproduce effects at the level of individual subjects and across a set of strong experimental manipulations of text layout. Point estimates of model parameters over the set of subjects provided theory-driven qualitative and quantitative explanations for variability in reading behavior as induced by experimental manipulations. This approach can in principle be used with other dynamical cognitive models (Schütt et al., 2017) and provides a basis for model comparisons within and between different models and theories.

Acknowledgements

This work was funded by Deutsche Forschungsgemeinschaft via Collaborative Research Center (SFB) 1287, project B03 (project no. 317633480). R. E. and S. A. S. received additional support by SFB 1294, project B03 (project no. 318763901). We acknowledge support by Norddeutscher Verbund für Hoch- und Höchstleistungsrechnen (HLRN, project no. bbx00001) for providing high-performance computing resources that contributed to the research results reported in this work.

References

Andrieu, C., & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, *37*(2), 697–725. doi: 10.1214/07-aos574

Chandra, J., Krügel, A., & Engbert, R. (2020). Modulation of oculomotor control during reading of mirrored and inverted texts. *Scientific Reports*, *10*, 4210. doi: 10.1038/s41598-020-60833-6

Engbert, R., & Kliegl, R. (2001, aug). Mathematical models of eye movements in reading: A possible role for autonomous saccades. *Biological Cybernetics*, *85*(2), 77–87. doi: 10.1007/PL00008001

Engbert, R., & Krügel, A. (2010). Readers use Bayesian estimation for eye movement control. *Psychological Science*, *21*(3), 366–371. doi: 10.1177/0956797610362060

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, *42*(5), 621–636. doi: 10.1016/S0042-6989(01)00301-7

Engbert, R., & Nuthmann, A. (2008). Self-consistent estimation of mislocated fixations during reading. *PLoS One*, *3*(2), e1534. doi: 10.1371/journal.pone.0001534

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813. doi: 10.1037/0033-295X.112.4.777

Engbert, R., Sinn, P., Mergenthaler, K., & Trukenbrod, H. (2015). *Microsaccade toolbox.* Retrieved from `http://read.psych.uni-potsdam.de/`

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Topics in Cognitive Science*, *5*, 452–474. doi: 10.1111/tops.12026

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, *14*(1), 153–158. doi: 10.1137/1114019

Findlay, J. M., & Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences*, *22*(4), 661–674. doi: 10.1017/S0140525X99002150

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109. doi: 10.1093/biomet/57.1.97

Holmes, W. R. (2015). A practical guide to the probability density approximation (PDA) with improved implementation and error characterization. *Journal of Mathematical Psychology*, *68*, 13–24. doi: 10.1016/j.jmp.2015.08.006

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12–35. doi: 10.1037/0096-3445.135.1.12

Kolers, P. A. (1976). Reading a year later. *Journal of Experimental Psychology*, *2*, 554–565. doi: 10.1037/0278-7393.2.5.554

Kolers, P. A., & Perkins, D. N. (1975). Spatial and ordinal components of form perception and literacy. *Cognitive Psychology*, *7*(2), 228–267. doi: 10.1016/0010-0285(75)90011-0

Krügel, A., & Engbert, R. (2014). A model of saccadic landing positions in reading under the influence of sensory noise. *Visual Cognition*, *22*(3-4), 334–353. doi: 10.1080/13506285.2014.894166

Laloy, E., & Vrugt, J. A. (2012). High-dimensional posterior exploration of hydrologic models using multiple-try dream(zs) and high-performance computing. *Water Resources Research*, *48*(1). doi: 10.1029/2011WR010608

Luce, R. D., & Raiffa, H. (1989). *Games and decisions: Introduction and critical survey*. Courier Corporation.

McConkie, G., Kerr, P., Reddix, M., & Zola, D. (1988). Eye movement control during reading: I. The location of initial eye fixations on words. *Vision Research*, *28*(10), 1107–1118. doi: 10.1016/0042-6989(88)90137-X

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. doi: 10.1063/1.1699114

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. doi: 10.1016/S0022-2496(02)00028-7

Nuthmann, A., & Engbert, R. (2009). Mindless reading revisited: An analysis based on the SWIFT model of eye-movement control. *Vision Research*, *49*(3), 322–336.

Nuthmann, A., Engbert, R., & Kliegl, R. (2005). Mislocated fixations during reading and the inverted optimal viewing position effect. *Vision Research*, *45*(17), 2201–2217. doi: 10.1016/j.visres.2005.02.014

Palestro, J. J., Sederberg, P. B., Osth, A. F., Van Zandt, T., & Turner, B. M. (2018). *Likelihood-free methods for cognitive science*. Springer.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, *52*(1), 1–56. doi: 10.1016/j.cogpsych.2005.06.001

Rabe, M. M., Vasishth, S., Hohenstein, S., Kliegl, R., & Schad, D. J. (2020). hypr: An R package for hypothesis-driven contrast coding. *The Journal of Open Source Software*, *5*, 2134. Retrieved from `https://CRAN.R-project.org/package=hypr` doi: 10.21105/joss.02134

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372–422. doi: 10.1037/0033-2909.124.3.372

Rayner, K., White, S. J., Johnson, R. L., & Liversedge, S. P. (2006). Raeding wrods with jubmled lettres. *Psychological Science*, *17*(3), 192–193. doi: 10.1111/j.1467-9280.2006.01684.x

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2012). Using E-Z Reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye–mind link. *Psychological Review*, *119*(1), 155. doi: 10.1037/a0026473

Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the EZ reader model. *Vision Research*, *39*(26), 4403–4411. doi: 10.1016/S0042-6989(99)00152-2

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21. doi: 10.3758/PBR.16.1.1

Reilly, R., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research*, *7*, 34–55. doi: 10.1016/j.cogsys.2005.07.006

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. doi: 10.1037/0033-295X.107.2.358

Schad, D. J., Betancourt, M., & Vasishth, S. (2020). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*. doi: 10.1037/met0000275

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. doi: 10.1016/j.jml.2019.104038

Schütt, H. H., Rothkegel, L. O., Trukenbrod, H. A., Reich, S., Wichmann, F. A., & Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, *124*(4), 505–524. doi: 10.1037/rev0000068

Seelig, S. A., Rabe, M. M., Malem-Shinitski, N., Risse, S., Reich, S., & Engbert, R. (2020). Bayesian parameter estimation for the SWIFT model of eye-movement control during reading. *Journal of Mathematical Psychology*, *95*. doi: 10.1016/j.jmp.2019.102313

Shockley, E. (2019). *PyDREAM.* Retrieved from `https://github.com/LoLab-VU/PyDREAM`

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*, 626–633. doi: 10.1080/01621459.1967.10482935

Snell, J., & Grainger, J. (2019). Readers are parallel processors. *Trends in Cognitive Sciences*, *23*(7), 537–546. doi: 10.1016/j.tics.2019.04.006

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018, nov). OB1-Reader: A model of word recognition and eye movements in text reading. *Psychological Review*, *125*(6), 969–984. doi: 10.1037/rev0000119

Starr, M. S., & Rayner, K. (2001). Eye movements during reading: Some current controversies. *Trends in Cognitive Sciences*, *5*(4), 156–163. doi: 10.1016/S1364-6613(00)01619-3

ter Braak, C. J. F., & Vrugt, J. A. (2008). Differential evolution Markov chain with snooker updater and fewer chains. *Statistics and Computing*, *18*(4), 435–446. doi: 10.1007/s11222-008-9104-9

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, *21*(2), 227–250. doi: 10.3758/s13423-013-0530-0

Vitu, F., McConkie, G. W., Kerr, P., & O'Regan, J. K. (2001). Fixation location effects on fixation durations during reading: An inverted optimal viewing position effect. *Vision Research*, *41*(25-26), 3513–3533. doi: 10.1016/S0042-6989(01)00166-3

Vrugt, J. A., ter Braak, C., Diks, C., Robinson, B. A., Hyman, J. M., & Higdon, D. (2009). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, *10*(3). doi: 10.1515/ijnsns.2009.10.3.273

Appendix A
The SWIFT model: Some mathematical details

From its first proposal, SWIFT (Engbert et al., 2002) incorporated two basically independent mechanisms for target selection and saccade timing, which are integrated through word activations. Word activations keep track of word processing, but also control target selection probabilities and modulate saccade timing. The state of the model (cf. Seelig et al., 2020) at time t is given as $n = (n_1, n_2, ..., n_{4+N_W})$ where $n_1, ..., n_4$ are saccade timers and $n_5, ...n_{4+N_W}$ are word activations with N_W as the number of words in a given sentence. Word activations rise during word recognition and fall during postlexical processing, where all n_i are discrete states, so that the internal state of SWIFT is a continuous-time, discrete state random walk. The temporal evolution of states is given by a master equation, which can be simulated efficiently on a computer (Seelig et al., 2020).

Words that fall within a processing span centered at the current gaze location are processed in parallel (Snell & Grainger, 2019). Processing starts at the letter level. We denote the eccentricity of letter j in word i (i.e., the distance from the current gaze position) by $\epsilon_{ij}(t)$. The width of the processing span is given by δ letter spaces to the left and to the right of the current fixation position. Using an inverse parabolic (asymmetric) processing function, a letter at eccentricity ϵ receives a processing rate

$$\lambda(\epsilon) = \lambda_0 \cdot \begin{cases} 1 - \epsilon^2/\delta^2 \,, & \text{for} \qquad |\delta| \leq \epsilon \\ 0 \,, & \text{otherwise} \end{cases} \,, \tag{7}$$

with $\lambda_0 = 3/4\delta$ a normalization constant. For the simulations in the current study, we are assuming a symmetric processing span given by Eq. (7); for a version with an asymmetric processing span extended to δ_L to left and to δ_R to right see Engbert et al. (2005); Seelig et al. (2020).

Next, the word-level processing rate $\Lambda_i(t)$ for word i is computed by

$$\Lambda_i(t) = L_i^{-\eta} \sum_{j=1}^{L_i} \lambda(\epsilon_{ij}(t)) \,, \tag{8}$$

where L_i is the word length of word i in number of letters and parameter η is a word-length exponent.

During processing, a word's activation increases with rate $\Lambda_i(t)$ to a word-frequency dependent maximum and decreases until activation returns to zero. During the decreasing part, word activations also decay with rate ω to account for memory leakage.

In SWIFT, a saccade is programmed to target a single word. Whenever a saccade target needs to be determined at time t, the target is selected according to a dynamic word activation field $a_m(t)$, with targeting probability $\pi(m,t)$ for word m given by relative activation, i.e.,

$$\pi(m,t) = \frac{a_m(t)}{\sum_{j=1}^{N_W} a_j(t)} \;, \tag{9}$$

which is implementing Luce's choice rule (Luce & Raiffa, 1989).

In SWIFT, saccades are generated by random timing (see also Engbert & Kliegl, 2001) that is modulated by foveal word activation (i.e., activation $a_k(t)$ of the fixated word k at time t) with strength given by parameter h (for details see Seelig et al., 2020).

Appendix B
Improved oculomotor assumptions

Oculomotor assumptions are critical for mathematical models of eye-movement control. For example, oculomotor noise generates about 10 to 15% of mislocated fixations (Engbert & Nuthmann, 2008; Krügel & Engbert, 2014; Nuthmann et al., 2005) as suggested by earlier work (McConkie et al., 1988).

For simplicity, most oculomotor models were based on normally distributed errors (Engbert & Nuthmann, 2008; McConkie et al., 1988). However, it should be noted that a normal distribution of saccade lengths will assign a non-zero likelihood to zero-length saccades (see Figure B1), in particular, in the case of refixations as their mean is often not significantly different from $d = 0$. Gamma distributions, however, specifically exclude values of zero, which means that a saccade length of $d = 0$ violates the model, i.e., it is assigned a likelihood of $P_{spat}\,(d = 0) = 0$ and will thus never stay at the exact same location after initiating a saccade, independent of the intended saccade target. In line with these assumptions, we propose a modified version of SWIFT which implements Gamma-distributed rather than normally distributed saccade lengths. Figure B1 compares the theoretical distributions of saccade amplitudes following a Gamma vså Gaussian distribution.

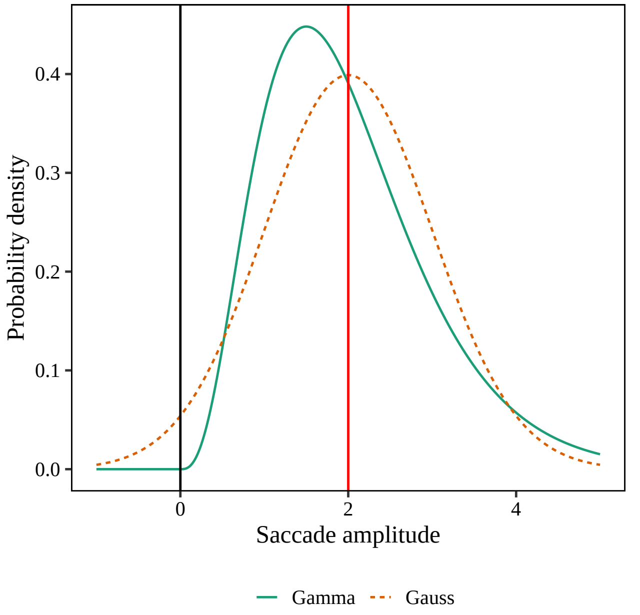
The likelihood (probability density function, PDF) $f(x)$ and cumulative density function (CDF) $F(x)$ of a Gamma distributed variable $x \in X$ are defined as follows, where $\Gamma(\alpha)$ is the gamma function and $\gamma(\alpha, \beta x)$ is the lower incomplete gamma function. The likelihood and CDF of a truncated Gamma distribution are normalized through division by the CDF of the upper bound, i.e.,

$$f(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{10}$$

$$F(x;\alpha,\beta) = \int_0^x f(u;\alpha,\beta)du = \frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)} \tag{11}$$

Figure B1

Theoretical distribution of saccade amplitudes assuming a Gamma vs. Gamma distribution



Note. Both distributions have a theoretical expected value (mean) of $\mathbb{E}[X] = 2.0$ and variance of $\mathrm{Var}[X] = 1.0$ originating from the gaze position $x_0 = 0$.

The saccade length $d \in D$ is a random variable that describes the one-dimensional spatial difference between two fixation locations. In reading research, it is often normalized to represent letter units, such that a saccade length of $d = 1.0$ describes a movement to the right by one letter width, whereas negative values denote movements to the left. In SWIFT, it has an expected value (mean) and variance of

$$\mathbb{E}\,[D] = v_m + \epsilon_{sre} - x_{i-1} \tag{12}$$

$$\mathrm{Var}\,[D] = \sigma_{sre}^2 \;, \tag{13}$$

where x_{i-1} is the launch site [5] and v_m is the target word center of word m. ϵ_{sre} and σ_{sre} are further decomposed into a fixed intercept and distance-dependent slope term where

$$\epsilon_{sre} = sre_1 - sre_2 \cdot (v_m - x_{i-1}) \tag{14}$$

$$\sigma_{sre} = omn_1 + omn_2 \cdot |v_m - x_{i-1}| \;, \tag{15}$$

which is in line with McConkie et al. (1988) and previous versions of SWIFT (Engbert et al., 2005; Seelig et al., 2020).

[5] Note that any within-word fixation location can be translated to a global (sentence-level) gaze position $x_i = l_i + \sum_{m=1}^{k_i-1}(1 + L_m)$, which is the cumulative letter position starting at the first letter of the first word, and vice versa. The global notation x_i in favor of (k_i, l_i) simplifies the computation of the spatial likelihood without any loss of precision.

Fixed and distance-dependent contributions to σ_{sre} are simply additive. As the expected value of the saccade amplitude $\mathbb{E}[D]$ is the sum of target distance ($v_m - x_{i-1}$) and ϵ_{sre}, saccade execution is more sensitive to the actual target distance for values of sre_2 closer to 0 and less sensitive for values closer to 1.

In our current work, we have changed the underlying distribution from Gaussian to Gamma with identical means and variances. The modification does not introduce any additional model parameters. Nor does it change the interpretation of existing model parameters with respect to the effect on mean (sre_1 and sre_2) and variance (omn_1 and omn_2) of saccade amplitudes. Note that the expected value of the saccade length is corrected to a half letter space if it occurs to be smaller (see Eqs. 17, 18), so that the expected value is always in the direction of the respective intended target.

$$d = x_i - x_{i-1} \tag{16}$$

$$\mathbb{E}_F[D] = \max(v_m + \epsilon_{sre} - x_{i-1}, 0.5) \tag{17}$$

$$\mathbb{E}_B[D] = \max(x_{i-1} - v_m - \epsilon_{sre}, 0.5) \tag{18}$$

$$\mathrm{Var}[D] = \sigma^2_{sre} \tag{19}$$

Depending on the relative location of the gaze position x_i and the center v_m of word m, the parameters of the Gamma distribution are chosen to be:

$$\alpha_{(.)} = \frac{(\mathbb{E}_{(.)}[D])^2}{\mathrm{Var}[D]} \tag{20}$$

$$\beta_{(.)} = \frac{\left|\mathbb{E}_{(.)}[D]\right|}{\mathrm{Var}[D]} \tag{21}$$

After the target k_i has been selected (cf. Seelig et al., 2020), the landing position x_i is determined by the sum of the launch site x_{i-1} and the saccade amplitude d where $d < 0$ is a saccade directed to the left and $d > 0$ is a saccade directed to the right. The saccade length d always follows a truncated Gamma distribution Γ_T in either direction. For forward saccades, the distributional parameters are determined by α_F and β_F with $d \in (0, x_{max} - x_{i-1})$, i.e. a landing position between x_{i-1} and x_{max}. For backward saccades, the distributional parameters are determined by α_B and β_B with $d \in (-x_{i-1}, 0)$, i.e. a landing position between 0 and x_{i-1}.

For forward fixations ($k_i = k_{i-1} + 1$), skippings ($k_i > k_{i-1} + 1$), and forward refixations ($k_i = k_{i-1} \wedge z > s$), $d \in D$ is Gamma-distributed with the tail to the right of the launch site. For regressions ($k_i < k_{i-1}$) and backward refixations ($k_i = k_{i-1} \wedge z \leq s$), $d \in D$ is Gamma-distributed with the tail to the left of the launch site:

$$D \sim \begin{cases} \Gamma_T(\alpha_F, \beta_F, x_{max} - x_{i-1}) & \text{for } k_i > k_{i-1} \vee \\ & (k_i = k_{i-1} \wedge z > s) \\ -\Gamma_T(\alpha_B, \beta_B, x_{i-1}) & \text{otherwise} \end{cases} \tag{22}$$

For refixations, the saccade length follows a weighted mixture distribution, composed of a positive Gamma distribution Γ_T with weight $1 - R$ and a negative Gamma distribution $-\Gamma_T$ with weight R, where R is the relative position of x_{i-1} within the word with 0.0 being the leftmost (including trailing whitespace) and 1.0 being the rightmost relative position. Therefore, a backward refixation following $-\Gamma_T$ is most likely for launch sites on the right word boundary and forward refixations are most likely for launch sites on the left word boundary. Thus, a backward refixation is executed if a uniformly distributed random number z is greater than the CDF of the backward refixation saccade length distribution (s, see Eq. 24). If not, a forward refixation is executed. R (see Eq. 23) depends on the previous fixation location x_{i-1}, the location of the right border of the previously fixated word $b_{k_{i-1}-1}$ and the length of that word $L_{k_{i-1}}$.

$$R = \frac{x_{i-1} - b_{k_{i-1}-1}}{L_{k_{i-1}} + 1}; R \in [0, 1] \tag{23}$$

$$s = \frac{R \cdot F(x_{i-1}; \alpha_B, \beta_B)}{(1 - R) \cdot F(x_{max} - x_{i-1}; \alpha_F, \beta_F) + R \cdot F(x_{i-1}; \alpha_B, \beta_B)} \tag{24}$$

The probability q of a landing position x_i differs between planned saccade directions. It always depends on the target m and the launch site x_{i-1}. For forward fixations and skippings (i.e., forward saccades), the likelihood is q_F (see Eq. 26). For regressions, the likelihood is determined with q_B (see Eq. 27). For refixations, the likelihood is the weighted sum of forward and backward saccade likelihoods, q_R (see Eq. 28).

$$\begin{aligned} q(k_i, l_i \mid m, F_{i-1}, \theta) &= q(x_i \mid m, x_{i-1}, \theta) \\ &= \begin{cases} q_F(x_i \mid m, x_{i-1}, \theta), & \text{for } k_i > k_{i-1} \\ q_B(x_i \mid m, x_{i-1}, \theta), & \text{for } k_i < k_{i-1} \\ q_R(x_i \mid m, x_{i-1}, \theta), & \text{for } k_i = k_{i-1} \end{cases} \end{aligned} \tag{25}$$

$$q_F(x_i \mid m, x_{i-1}, \theta) = \frac{f(x_i - x_{i-1}; \alpha_F, \beta_F)}{F(x_{max} - x_{i-1}; \alpha_F, \beta_F)} \tag{26}$$

$$q_B(x_i \mid m, x_{i-1}, \theta) = \frac{f(x_{i-1} - x_i; \alpha_B, \beta_B)}{F(x_{i-1}; \alpha_B, \beta_B)} \tag{27}$$

$$\begin{aligned} &q_R(x_i \mid m, x_{i-1}, \theta) = \\ &\frac{R \cdot f(x_i - x_{i-1}; \alpha_F, \beta_F) + (1 - R) \cdot f(x_{i-1} - x_i; \alpha_B, \beta_B)}{(1 - R) \cdot F(x_{max} - x_{i-1}; \alpha_F, \beta_F) + R \cdot F(x_{i-1}; \alpha_B, \beta_B)} \end{aligned} \tag{28}$$