




Modeling the Interaction of Sentence Processing and Eye-Movement Control in Reading

Maximilian Michael Rabe

Master of Science (University of Victoria, 2018)
Bachelor of Science (University of Potsdam, 2016)

 0000-0002-2556-5644

Department of Psychology
Faculty of Human Sciences
University of Potsdam

Dissertation submitted in September 2023 to the Faculty of Human Sciences at the
University of Potsdam in partial fulfillment of the requirements for the degree of
Doctor of Philosophy (Ph.D.) in Cognitive Science

This work is protected by copyright and/or related rights. You are free to use this work in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s).
<https://rightsstatements.org/page/InC/1.0/?language=en>

Date submitted: 2023-09-11

Date defended: 2024-01-19

Advisors:

Prof. Dr. Ralf Engbert, Department of Psychology

Prof. Dr. Shravan Vasishth, Department of Linguistics

Reviewers:

Prof. Dr. Ralf Engbert, Department of Psychology, University of Potsdam

Prof. Dr. Lynn Huestegge, Institute of Psychology, University of Wurzburg

Doctoral committee:

Prof. Dr. Milena Rabovsky, Department of Psychology (Chair of committee)

Prof. Dr. Ralf Engbert, Department of Psychology

Prof. Dr. Shravan Vasishth, Department of Linguistics

Prof. Natalie Boll-Avetisyan, Ph.D., Department of Linguistics

Dr. Jochen Laubrock, Department of Psychology

Published online on the

Publication Server of the University of Potsdam:

<https://doi.org/10.25932/publishup-62279>

<https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-622792>

Abstract

The evaluation of process-oriented cognitive theories through time-ordered observations is crucial for the advancement of cognitive science. The findings presented herein integrate insights from research on eye-movement control and sentence comprehension during reading, addressing challenges in modeling time-ordered data, statistical inference, and interindividual variability. Using kernel density estimation and a pseudo-marginal likelihood for fixation durations and locations, a likelihood implementation of the SWIFT model of eye-movement control during reading (Engbert et al., *Psychological Review*, 112, 2005, pp. 777–813) is proposed. Within the broader framework of data assimilation, Bayesian parameter inference with adaptive Markov Chain Monte Carlo techniques is facilitated for reliable model fitting. Across the different studies, this framework has shown to enable reliable parameter recovery from simulated data and prediction of experimental summary statistics. Despite its complexity, SWIFT can be fitted within a principled Bayesian workflow, capturing interindividual differences and modeling experimental effects on reading across different geometrical alterations of text. Based on these advancements, the integrated dynamical model SEAM is proposed, which combines eye-movement control, a traditionally psychological research area, and post-lexical language processing in the form of cue-based memory retrieval (Lewis and Vasishth, *Cognitive Science*, 29, 2005, pp. 375–419), typically the purview of psycholinguistics. This proof-of-concept integration marks a significant step forward in natural language comprehension during reading and suggests that the presented methodology can be useful to develop complex cognitive dynamical models that integrate processes at levels of perception, higher cognition, and (oculo-)motor control. These findings collectively advance process-oriented cognitive modeling and highlight the importance of Bayesian inference, individual differences, and interdisciplinary integration for a holistic understanding of reading processes. Implications for theory and methodology, including proposals for model comparison and hierarchical parameter inference, are briefly discussed.

Zusammenfassung

Die Evaluierung prozessorientierter kognitiver Theorien durch zeitlich geordnete Beobachtungen ist ein zentraler Baustein für die Weiterentwicklung der Kognitionswissenschaft. Die hier präsentierten Ergebnisse integrieren Erkenntnisse aus der Forschung zur Blickbewegungskontrolle und zur Satzverarbeitung beim Lesen und gehen dabei auf Herausforderungen bei der Modellierung von zeitlich geordneten Daten, statistischer Inferenz und interindividueller Variabilität ein. Unter Verwendung von Kerndichteschätzung und einer pseudo-marginalen Wahrscheinlichkeitsverteilung für Fixationsdauern und -orte wird eine Implementation für die Likelihood des SWIFT-Modells zur Blickbewegungskontrolle beim Lesen (Engbert et al., *Psychological Review*, 112, 2005, S. 777–813) eingeführt. Im breiteren Kontext der Datenassimilation wird Bayes'sche Parameterinferenz mit adaptiven Markov-Chain-Monte-Carlo-Techniken verwendet, um eine zuverlässige Modellanpassung zu ermöglichen. In verschiedenen Studien hat sich dieser methodische Rahmen als geeignet erwiesen, um zuverlässige Parameterrückgewinnung aus simulierten Daten und Vorhersage experimenteller Zusammenfassungen zu ermöglichen. Trotz dessen Komplexität kann SWIFT innerhalb eines fundierten Bayes'schen Workflows angepasst werden und macht daraufhin zuverlässige Vorhersagen für interindividuelle Unterschiede sowie die Modellierung experimenteller Effekte bei verschiedenen geometrischen Änderungen von Text. Basierend auf diesen Fortschritten wird das integrierte dynamische Modell SEAM eingeführt. Dieses kombiniert die Forschungsgebiete der traditionell psychologisch geprägten Blickbewegungskontrolle und der traditionell psycholinguistisch geprägten postlexikalischen Sprachverarbeitung in Form von cue-basiertem Gedächtnisabruf (Lewis und Vasishth, *Cognitive Science*, 29, 2005, S. 375–419). Der Nachweis der Durchführbarkeit solcher integrativer Modelle stellt einen bedeutenden Fortschritt bei der natürlichen Sprachverarbeitung beim Lesen dar und legt nahe, dass die vorgestellte Methodik nützlich sein kann, um komplexe kognitive dynamische Modelle zu entwickeln, die Prozesse auf den Ebenen der Wahrnehmung, höheren Kognition, und (okulo-)motorischen Kontrolle integrieren. Diese Erkenntnisse fördern insgesamt die prozessorientierte kognitive Modellierung und betonen die Bedeutung der Bayes'schen Inferenz, individueller Unterschiede und interdisziplinärer Integration für ein ganzheitliches Verständnis von Leseprozessen. Implikationen für Theorie und Methodologie, einschließlich Vorschlägen für Modellvergleich und hierarchische Parameterinferenz, werden kurz diskutiert.

Acknowledgments

This dissertation is the result of several years of collaborative effort and significant parts of it would have been impossible to achieve without the tireless and dedicated support from my advisors. My primary advisor Ralf Engbert introduced me into the field of complex cognitive modeling, particularly dynamical models of eye movements. Despite his many other academic duties, he ensured very close and productive supervision of this project throughout my studies. The many frequent and pleasant conversations with him as well as all my learnings from them were simply invaluable and helped me grow academically and personally. I also enjoyed all the extremely interesting and fruitful discussions with my secondary advisor Shravan Vasishth. From him, I learned close to everything I know about language processing and Bayesian statistics. Meetings with him never concluded without a solution or brilliant ideas for further research. Of course, I am also deeply grateful to Reinhold Kliegl for introducing me to psycholinguistic research in the first place (Masson et al., 2017), consigning me a comprehensive book collection upon his retirement from reading research, and including me in various exciting side projects along the way.

During my doctoral studies, I have been very blessed to collaborate closely with a couple of other doctoral students, all of whom have graduated with their own outstanding projects in the meantime. First of all, I would like to mention Daniela Mertzen, who planned and conducted all of the experiments in our research project (project B03 *Modelling the Interaction Between Eye-Movement Control and Parsing Processes* in SFB 1287). Her work culminated in an excellent dissertation, submitted and defended in 2022, and is also the empirical base for Chapter 4 of this dissertation. Together with her, Dario Paape has been very helpful preparing Daniela's experimental stimuli for modeling. Moreover, he has contributed many comments, thoughts, and concrete ideas for the integration of SWIFT and LV05 in SEAM (see Chapter 4). I also want to thank Stefan Seelig and Johan Chandra, the main collaborators on Chapters 2 and 3. Stefan was the main person in charge of implementing the likelihood for SWIFT and the majority of the SWIFT implementations used in Chapters 3 and 4 are based on his computer code. Johan Chandra was the main collaborator on Chapter 3, contributing the empirical base and relevant reference analyses for that work. All of aforementioned people have made unspeakable efforts to advance my dissertation project and I am very thankful for their collaboration.

Besides the work on my primary research project, I also deeply enjoyed the daily interactions with other current and former lab members, including Daniel Backhaus, Anke Cajar, André Krügel, Sarah Risse, and Lisa Schwetlick. Thank you all for your helpful input at lab

meetings or just the casual chat over lunch at Golm's very best (and very only) university canteen. Of course, I would also like to thank everyone in administration and lab management, in particular Nicole Dungel, Petra Schienmann, and Michaela Schmitz, as well as many others, who made sure that everything from *Arbeitszeiterfassung* to *Zahlungsverfugung* went as smoothly as it did.

Last but not least, I am extremely grateful for the support I have received from family and friends and the many joyful moments I could experience with them. Among those people, I am without doubt most thankful for the endless support from my wife Sara and for our delightful daughter Lykka, who has joined the team somewhere between Chapters 3 and 4.

This work was supported by grants from Deutsche Forschungsgemeinschaft (DFG) to Ralf Engbert and Shravan Vasishth (SFB 1287 *Limits of Variability in Language*, project no. 317633480), and to Ralf Engbert and Sebastian Reich (SFB 1294 *Data Assimilation*, project no. 318763901). Computations were partly supported with high-performance computing resources provided by Norddeutscher Verbund fur Hoch- und Hochleistungsrechnen (HLRN, project no. bbx00001).

Contents

| | |
|--|-------------|
| Abstract | iii |
| Zusammenfassung | iv |
| Acknowledgments | v |
| Contents | vii |
| List of Figures | xi |
| List of Tables | xiii |
| 1 General introduction | 1 |
| 1.1 Eye movements in reading | 2 |
| 1.1.1 Eye movements | 3 |
| 1.1.2 Eye-movement statistics | 4 |
| 1.1.3 Reading-related statistics | 5 |
| 1.2 Experimental methods in eye-movement research | 6 |
| 1.3 Cognitive, linguistic, and oculomotor effects during reading | 7 |
| 1.3.1 Oculomotor effects | 8 |
| 1.3.2 Cognitive effects | 8 |
| 1.3.3 Psycholinguistic effects | 10 |
| 1.4 Computational models of reading | 13 |
| 1.4.1 Psychological reading models | 14 |
| 1.4.2 Linguistic reading models | 17 |
| 1.4.3 Contemporary integrated models | 18 |
| 1.5 Common shortcomings of contemporary reading models | 19 |
| 1.5.1 Explanatory deficiencies | 19 |
| 1.5.2 Material | 20 |
| 1.5.3 Parameter inference | 20 |
| 1.6 Methods of parameter inference | 21 |
| 1.6.1 Hand-picked values | 21 |
| 1.6.2 Parameter optimization | 22 |
| 1.6.3 Goodness-of-fit | 22 |

| | | |
|----------|--|-----------|
| 1.6.4 | Likelihood | 22 |
| 1.6.5 | Bayesian inference | 24 |
| 1.7 | Summary | 25 |
| 2 | Bayesian parameter estimation for the SWIFT model of eye-movement control during reading | 27 |
| 2.1 | Introduction | 27 |
| 2.1.1 | Eye-movement control during reading | 28 |
| 2.1.2 | The likelihood function for dynamical cognitive models | 30 |
| 2.2 | The SWIFT model of saccade generation during reading | 32 |
| 2.2.1 | Saccade target selection and temporal evolution of activations | 33 |
| 2.2.2 | Temporal control of saccades and foveal inhibition | 35 |
| 2.2.3 | Character-based visual processing | 36 |
| 2.2.4 | Word-based processing rate | 37 |
| 2.2.5 | Oculomotor assumptions | 38 |
| 2.2.6 | Modulation of the duration of the labile stage | 39 |
| 2.2.7 | Numerical simulation and model parameters | 40 |
| 2.3 | Likelihood function for the SWIFT model | 40 |
| 2.3.1 | Spatial likelihood | 42 |
| 2.3.2 | Temporal likelihood | 43 |
| 2.3.3 | Evaluation of the log-likelihood using single-parameter variations | 46 |
| 2.4 | Likelihood-based parameter inference using MCMC | 47 |
| 2.4.1 | Markov Chain Monte Carlo simulation for the SWIFT model | 47 |
| 2.4.2 | Parameter recovery using simulated data | 50 |
| 2.4.3 | Estimation of parameters based on experimental data | 50 |
| 2.4.4 | Interindividual differences and model parameters | 50 |
| 2.5 | Discussion | 55 |
| 3 | A Bayesian approach to dynamical modeling of eye-movement control in reading of normal, mirrored, and scrambled texts | 57 |
| 3.1 | Introduction | 57 |
| 3.1.1 | The Bayesian approach to dynamical cognitive models | 60 |
| 3.1.2 | Principled Bayesian workflow in model inference | 62 |
| 3.1.3 | Summary statistics | 63 |
| 3.2 | The SWIFT model of eye-movement control | 64 |
| 3.3 | The likelihood function for SWIFT | 65 |
| 3.4 | Computational methods | 68 |
| 3.5 | Experiment | 70 |
| 3.5.1 | Data preprocessing | 70 |

| | | |
|----------|---|------------|
| 3.6 | Results | 71 |
| 3.6.1 | Likelihood profiles | 71 |
| 3.6.2 | Parameter recovery | 72 |
| 3.6.3 | Experimental data: Summary statistics | 73 |
| 3.6.4 | Parameter estimates | 74 |
| 3.6.5 | Posterior predictive checks | 74 |
| 3.6.6 | Statistical evaluation of model parameters | 81 |
| 3.7 | Discussion | 84 |
| 4 | SEAM: An integrated activation-coupled model of sentence processing and eye movements in reading | 88 |
| 4.1 | Introduction | 89 |
| 4.1.1 | The activation-based model of sentence processing (Lewis & Vasishth, 2005) | 92 |
| 4.1.2 | The SWIFT model of eye-movement control (Engbert et al., 2005) | 95 |
| 4.1.3 | SEAM: Activation-based coupling of SWIFT and LV05 | 98 |
| 4.2 | Data availability | 103 |
| 4.3 | Experimental study (Mertzen et al., 2023) | 103 |
| 4.4 | Simulation study | 106 |
| 4.4.1 | Method | 109 |
| 4.4.2 | Results | 111 |
| 4.4.3 | Discussion | 126 |
| 4.5 | General discussion | 127 |
| 4.6 | Conclusion | 130 |
| 5 | General discussion | 131 |
| 5.1 | Summary | 131 |
| 5.2 | Statistical rigor through Bayesian inference | 132 |
| 5.3 | Individual differences in eye movements during reading | 133 |
| 5.4 | Integration of cognitive processes during reading | 133 |
| 5.5 | Effects of similarity-based interference | 134 |
| 5.6 | Accessibility of model implementations | 135 |
| 5.7 | Model comparison of complex models | 135 |
| 5.8 | Hierarchical Bayesian modeling | 136 |
| 5.9 | Conclusion | 138 |
| | References | 139 |
| A | Experimental data and sentence material | 158 |

| | |
|--|------------|
| B The SWIFT model: Some mathematical details | 159 |
| C Improved oculomotor assumptions | 161 |
| D SEAM model parameters | 165 |
| E Effects of memory interference on experimental and simulated regression probabilities | 167 |
| Author publications | 168 |
| Declaration of authorship | 171 |
| Eigenständigkeitserklärung | 171 |

List of Figures

| | | |
|------|---|----|
| 2.1 | Sequence of fixations during reading | 29 |
| 2.2 | Simulation example for the SWIFT model | 34 |
| 2.3 | Oculomotor error model | 39 |
| 2.4 | Schematic illustrations of the generation of fixation durations for different types of fixations in SWIFT | 44 |
| 2.5 | Temporal, spatial and total log-likelihood profiles | 48 |
| 2.6 | Exemplary posterior distributions | 51 |
| 2.7 | Example posterior densities for single participants | 52 |
| 2.8 | Posterior distributions of 34 participants | 53 |
| 2.9 | Relationship between true parameters and estimated parameter values of generated data | 54 |
| 2.10 | Correlation of empirical and simulated fixation durations and probabilities | 55 |
| 3.1 | Typical eye trajectories during reading | 59 |
| 3.2 | An eye trajectory as simulated in SWIFT | 65 |
| 3.3 | Empirical and previously simulated Gaussian saccade amplitudes aggregated across all subjects in each experimental condition | 67 |
| 3.4 | Centered likelihood components for selected model parameters | 72 |
| 3.5 | Scatterplot of true and recovered parameters with 60% HPDIs | 73 |
| 3.6 | Posterior densities for all fitted model parameters | 75 |
| 3.7 | Comparison of simulated summary statistics when sampling from the posterior vs. using point estimates | 77 |
| 3.8 | Empirical and simulated spatial summary statistics (fixation probabilities) for different experimental conditions, aggregated across subjects, as a function of word length | 78 |
| 3.9 | Correlation between empirical (horizontal axis) and simulated (vertical axis) spatial summary statistics (fixation probabilities) | 78 |
| 3.10 | Empirical and simulated saccade amplitudes aggregated across all subjects in each experimental condition | 79 |
| 3.11 | Empirical and simulated landing positions for single fixations, first fixations, and second fixations | 79 |

| | | |
|------|--|-----|
| 3.12 | Empirical and simulated temporal summary statistics (fixation durations) for different experimental conditions, aggregated across subjects, as a function of word length | 82 |
| 3.13 | Correlation between empirical (horizontal axis) and simulated (vertical axis) temporal summary statistics (fixation durations) | 82 |
| 3.14 | Linear regression results for model parameters | 83 |
| 4.1 | Word activation in SWIFT | 99 |
| 4.2 | Word activation in SEAM | 101 |
| 4.3 | Experimental effects of Mertzen et al. (2023) | 105 |
| 4.4 | Example simulation in SWIFT | 107 |
| 4.5 | Example simulation in SEAM | 108 |
| 4.6 | Example profile log-likelihoods | 113 |
| 4.7 | Parameter recovery of SEAM parameters | 115 |
| 4.8 | Spatial and temporal summary statistics | 116 |
| 4.9 | Posterior distributions of estimated experimental effects | 117 |
| 4.10 | Distribution of absolute prediction errors for estimated experimental effects | 118 |
| 4.11 | Effect of subjecthood on the contributions of first-pass reading and rereading times to regression path durations | 120 |
| 4.12 | Conditional means of experimental and simulated regression probabilities | 121 |
| 4.13 | Conditional means of experimental and simulated regression durations | 122 |
| 4.14 | Effects of experimental condition on SEAM word activations at encoding of the critical verb | 123 |
| C.1 | Theoretical distribution of saccade amplitudes assuming a Gamma vs. Gaussian distribution | 162 |
| E | Effects of memory interference on experimental and simulated regression probabilities | 167 |

List of Tables

| | | |
|-----|---|-----|
| 2.1 | Stochastic transitions between adjoined states from $n = (n_1, n_2, \dots) \mapsto n' = (n'_1, n'_2, \dots)$ | 35 |
| 2.2 | Model parameters of the SWIFT model | 41 |
| 2.3 | Parameters of the SWIFT model considered in Bayesian estimation | 47 |
| 3.1 | Reading conditions used as modeling targets | 60 |
| 3.2 | Fitted SWIFT model parameters | 69 |
| 3.3 | Empirical means and standard errors in summary statistics aggregated across subjects | 74 |
| 3.4 | Change in MSE across subject-level summary statistics between posterior sampling and point estimates | 76 |
| 3.5 | Correlations across subjects for empirical vs. simulated summary statistics | 80 |
| 4.1 | Stochastic transitions between internal states from $n = (n_1, n_2, \dots) \mapsto n' = (n'_1, n'_2, \dots)$ | 97 |
| 4.2 | Summary of empirical vs. model estimates from SEAM and SWIFT of the subjecthood and animacy effects on regression path durations and first-pass regressions | 117 |
| D | SEAM model parameters | 165 |

Chapter 1

General Introduction

Reading is an extremely important skill, for most of us every single day in both personal and professional regards. Most knowledge, especially in academic contexts, is preserved and distributed in written form. That is why, perhaps unsurprisingly, reading skill is fundamental for academic success (e.g., Aerila & Merisuo-Storm, 2017; Barredo et al., 2022; Boakye, 2017; Meyer & Pietzner, 2022). Also in many other everyday situations, reading guides our actions and helps us navigate through society and our own personal development. We know, however, that reading literally does not come naturally; it is not an evolutionary trait but a cultural artifact. Nor is reading skill constant across the lifespan or task demands.

Understanding what processes are at play when we read has immense potential for both basic and applied research. For example, a deeper knowledge of reading can help us identify reasons for reading disabilities and develop targeted interventions to foster reading skill. Moreover, as reading draws on similar resources and mechanisms as other visual tasks, insights into the processes during reading will also be valuable in the understanding of visual cognition altogether.

The investigation of eye movements during reading has a long tradition in the cognitive sciences. Rayner (1978; 1998) identified four stages of this research domain: The first stage commenced with basic observations by the French ophthalmologist Javal in 1878. That era coined central concepts and terminology of eye movements such as *saccades* and *fixations* (see next Section for an introduction). Technical constraints quickly led to a second era of decreasing interest in eye-movement research. Given the high velocity and frequency of saccades during reading, observations under controlled experimental conditions were simply not possible and hence not further investigated. A third, and exponentially more productive, era began with the technological advances brought forward by eye-tracking devices and more sophisticated language models. Though, it was not until 2009 that Rayner identified a fourth era of sophisticated computational modeling of eye movements during reading. The abundance of experimental research in vision, cognition, and language had motivated highly sophisticated reading models, some of which are extremely successful in predicting and explaining challenging empirical phenomena. What most of these models have in common though is that they focus on single psychological *or* linguistic aspects of reading, which potentially limits their generalizability and scientific value.

How does a cognitive representation of the sentence emerge? How do cognition, higher language processing, and eye movements interact? Given the extremely fast and frequent

gaze shifts during reading, how much of that is directed by language processing, and how much is the result of noisy oculomotor execution? Without any doubt, during reading, a lot of processes take place, such as visual perception, attention allocation, language processing, and oculomotor movements. Agnostic to the concrete interaction between them, it is arguably important to consider all of them to offer a statistically reliable and satisfactory explanatory account of reading. After a brief chapter introducing basic concepts and contemporary reading research I present results from three modeling studies that bridge psychological and linguistic threads of reading research in SEAM, an integrated computational model within a principled Bayesian modeling framework.

1.1 Eye Movements in Reading

What happens when we read? Reading is the intake of information from abstracted language representations in visual or haptic form.¹ This intake and further comprehension presumably employs a large ensemble of processes, such as visual perception and lexical and/or phonological identification of fixated words, syntactic parsing, semantic integration, short-term memory storage, and eventually the planning and execution of the subsequent gaze shift.

Experimental research has shown that manipulating text or the reading task can systematically affect behavioral and neuropsychological measures such as eye movements, comprehension tasks, or neuroimaging signals, and therefore suggests that such processes must be involved during reading in some way. However, the isolated evidence for each of these individual processes does not provide much clarity about the concrete order of events in the processing cascade and how the individual processes are linked with each other. While it seems trivial that the eyes can only perceive what lies within their view, our prior experience with the currently read sentence or other texts in general can affect reading as early as word identification, which will in turn affect all processing that builds on word identification (see Section 1.3 for an introduction to cognitive, linguistic, and oculomotor effects on eye movements in reading).

However, there is intense debate in psychology and linguistics about the linking between those processes. How is the visual input entered into the processing cascade, what processes are involved in what order, and is there any backcoupling such that language processing demands modulate the attention shifts, resulting in targeted saccades, for example? To address those questions, it is necessary to take a closer look at eye movements in general and their application in reading research.

¹Due to the availability of technology (eye-tracking devices) and prior research, this dissertation focuses on visual reading.

1.1.1 Eye Movements

During the visual inspection of various natural and artificial stimuli (such as text), we can mainly observe short, ballistic eye movements, *saccades*, followed by longer periods of relative rest, *fixations*, which was first reported by French ophthalmologist Javal (1878) for reading and subsequently confirmed for vertebrates and other vision tasks (see Land, 2011, for a review). Saccadic eye movements are jerky radial movements of the ocular bulbs, resulting in a transitional shift of the fovea (or gaze) relative to the environment.

Although the term “fixation” suggests that the eyes be motionless during those periods, they never really are, as von Helmholtz (1896) noted shortly after. The largest and fastest type of movements during fixations (i.e., *fixational eye movements*) are *microsaccades*. One of the most likely functions of microsaccades is to prevent retinal fatigue, i.e. the perceptual fading of the visual stimulus, as our visual system is primed to detect changes in the environment (Ditchburn & Ginsborg, 1952; Martinez-Conde & Macknik, 2011; Martinez-Conde et al., 2009). Similar to saccades, microsaccades are mostly linear, ballistic movements, though of much smaller amplitude. Other fixational eye movements are the short, meandering *drifts* between microsaccades and superimposed oscillatory *tremor*.²

Reading research usually focuses on saccades and collapses all fixational eye movements as pseudo-stationary fixations. In a typical reading task, depending on context, language, reader, and publication, fixations have average durations of about 150–300 ms and saccades take about 20–40 ms, which results in alternating sequences of about three to four fixations and saccades per second. Given the very short duration and long distances of saccades, they are executed at very high velocities of about 400° visual angle³ per second (see Gilchrist, 2011, for a review).

In addition to counteracting visual fading, the eyes must frequently move across visual stimuli because the surface area of the retina with sufficiently high acuity for detailed stimuli such as words, the *fovea*, is quite limited to about 2° visual angle (Land, 2011). For example, a segment of the stimulus greater than 1.4 cm, viewed at a distance of 40 cm, already exceeds 2° visual angle and cannot be processed at high acuity without additional cognitive inference, a executing a refixation on a different subregion of the same stimulus, or compromising visual input quality. This would already apply to many words in print, for example, but also to larger objects in scene viewing. So, in order to allow for at least the most relevant segments of the stimulus to be visually processed at adequate acuity, the eyes have to move

²For a comprehensive characterization of microsaccades, drift, and tremor, see Martinez-Conde et al. (2009).

³Visual angle is a spatial measure relative to the spherical surface of the retina. The geometric relation of size s , distance d and visual angle ν is $\nu = 2 \arctan \frac{s}{2d}$. Therefore, a larger object occupies more visual angle than a smaller object, given equal distance. Likewise, an object at higher distance occupies less visual angle than a closer object, given equal size. As a “rule of thumb”, the width of the thumbnail held at a distance of an arm’s length occupies about 1° visual angle.

across the stimulus, depending on various factors of stimulus, context, and observer, in order to take in as much detailed visual information as possible.

Assuming that saccades are at least under some volitional control, using the *double-step paradigm*, Becker and Jürgens (1979) could demonstrate that saccades are “programmed” in two stages and that at least the direction of a saccade is inevitably programmed once a randomly varying point-of-no-return during a fixation is reached. This suggests that, even though information processing may continue at all times, target selection has to take place some time before the execution of the saccade, i.e. during the fixation immediately preceding the saccade.

Due to the high velocity of saccades, which would smear the retinal image, visual perception during saccades is suppressed. Instead, intake of visual information is more effective when the image is relatively stationary on the retina (e.g., see Ross et al., 2001).⁴ During a fixation, i.e. during the periods of relative rest after the covert attention shift and before the initiation of the upcoming saccade, the visual stimulus is available to visual perception and subsequent higher cognitive processing. Furthermore, if saccade execution is in any way modulated by cognition, oculomotor processes such as saccade target selection (i.e., deciding whether/when/where to move next) would have to be carried out during that time as well.

1.1.2 Eye-Movement Statistics

The execution of saccades, by definition, determines the sequence of fixations. Insofar, statistics based on placement and timing of saccades/fixations should be understood as outcomes of the same system, focusing on different phases of the eyes’ trajectory. Saccades are movements, which is why they can be characterized by their duration, their *velocity profiles*, their *direction* or angle, and spatial relations between their *launch* and *landing sites* (such as *saccade amplitudes* or *landing locations*). Given that information intake during saccades is severely limited, the duration of a fixation, being the time between two saccadic eye movements, may indicate information processing and saccade planning demands. So, if a fixation takes significantly longer under some specified condition, it may be inferred that this condition poses additional difficulty for the processing of the visual information and/or planning of the upcoming saccade. Moreover, the locations of fixations may be aggregated in *fixation maps* or *fixation probabilities*, the latter of which are of particular interest for reading research.⁵

⁴A normal exception to saccadic eye movements, besides some ocular pathologies, is the *smooth pursuit* of a moving stimulus. However, even then the eyes only move in order to keep the stimulus stationary relative to the retina.

⁵A fixation probability is typically the probability for some defined spatial region to receive a fixation during some defined time period (such as the trial), $P(\text{fixation} \mid \text{location})$. Fixation maps, on the other hand, are conditional maps of fixation locations across the visual stimulus, $P(\text{location} \mid \text{fixation})$.

1.1.3 Reading-Related Statistics

In the context of reading research, eye movements have been established as an extremely useful observable. Reasons include that eye trackers have much higher temporal and spatial resolution and precision than behavioral measures but are more affordable than neuroimaging (such as EEG or fMRI). Moreover, as texts are inherently artificial stimuli, compared to other domains of vision research, it is relatively straightforward to generate, manipulate, or annotate stimuli, and to analyze the eye movements in response to the presented materials.

Sentences in alphabetical languages are relatively standardized (or standardizable) stimuli such that they consist of visually delimited linguistic subunits, *words*. They are often presented in a single line on the screen to avoid line jumps, which naturally divides the screen into potentially meaningful discrete target locations. When single-line displays are used, we can also disregard vertical movements and only analyze horizontal gaze shifts. This simplifies the analysis by enabling us to categorize all saccades into *forward*, *skipping*, *regressive*, or *refixating saccades*, depending on the launch site (previous fixation location). In typical reading tasks in German and English (Kliegl et al., 2004; Rayner, 1998), forward saccades are most common (approx. 50%), followed by refixations and skipplings (each approx. 20%), and regressions (approx. 10%).

Most commonly, reading research considers *fixation durations* and *fixation probabilities*, which are derived from the observed fixation sequences. Generally speaking, they are the conditional mean duration of a fixation or the conditional probability to observe a fixation on some region or word. When calculating these statistics, most frequently, only *first-pass*⁶ fixations are considered, unless specifically noted. Additionally, especially when considering oculomotor execution, it is also possible to study the distribution of saccade amplitudes, possibly conditional on saccade type.

This allows us to calculate a number of more or less gold-standard summary statistics, which are commonly reported in the experimental and modeling literature. Fixation durations are thought of as a proxy for processing demand during the fixation. *First fixation durations* (FFD) are first-pass fixations of the initial fixation on a word, excluding any consecutive (re-)fixations. *Gaze durations* (GD) or *first-pass reading times* (FPRT), include all consecutive first-pass fixations (including the initial fixation and refixations) on the same word until the gaze shifts away from the word. *Skipping durations* (SD) are first-pass fixation durations on word following a skipping saccade. *Go-past duration* (GPD) or *regression-path duration* (RPD) is the sum of all gaze durations from first fixating a word until (but excluding) the eyes have left to the right of the word or region (or the trial ends), which includes all refixations and regressions immediately following first-pass reading. *Re-reading times* or

⁶For left-to-right languages, a fixation is considered first-pass if the eyes have not fixated the current word or any word to the right of the current word previously in the fixation sequence. Refixations immediately following a first-pass fixation are also considered first-pass.

second-pass/n-pass fixation durations are fixation durations on words after first pass, which includes words previously fixated or skipped. *Total viewing time* (TVT) or *total reading time* (TRT) is the sum of all gaze durations on a word or region, including but not limited to first-pass fixations.

Besides these temporal statistics, fixation probabilities can be thought of as a means for the spatial distribution of attention allocation, since it is commonly assumed that words are skipped or regressed more often if they pose lower or, respectively, higher processing demand. Like fixation durations, they are conditional on the type of saccade. In that sense, *(first-pass) fixation probability*, *regression probability*, *skipping probability*, or *refixation probability* are the probability that a word or region has been fixated, regressed to, skipped, or refixated, respectively, during first-pass reading. Besides fixation probabilities, researchers concerned with oculomotor execution may also be interested in the distribution of *saccade amplitudes* (usually in reference to the launch site) or *within-word landing positions* (usually in reference to the word center).

1.2 Experimental Methods in Eye-Movement Research

Cognitive and linguistic theories, especially process-oriented ones, attempt to explain how underlying cognitive or linguistic processes generate behavior in response to some (language) stimulus. The general aim of a behavioral experiment is to test beliefs or predictions from those theories about the relation between stimuli and resulting behavior by testing whether the controlled experimental manipulation of a presented stimulus coincides with different behavior. If we find effects across trials (i.e., experimental variations of stimuli being associated with consistent changes in behavioral patterns), we can conclude that there is some systematic influence of the stimulus on the observed behavior. Variability or inconsistencies are surprisingly often considered a nuisance or “measurement error” and not further considered. Nevertheless, even if the human participant was the only mediator between stimulus and response, that systematic influence may still include perception, higher processing, and execution of the response, all of which are covert, likely vary between individuals, may partially overlap or interact, and can contribute to the form and latency of the observable response. Across the many different experimental paradigms in behavioral sciences, it is therefore considered vital to attempt to reduce the contribution of any systematic influence that is outside the scope of the research question. Often, this is carried out at the levels of the proper experiment but also via statistical techniques.

In the context of reading research, a typical experiment entails the analysis of the effects (and sometimes the variability) of eye movements in response to experimental manipulation of the presented text. As for other behavioral experiments, the goal then is to test theories about how humans perceive, process, and respond to the exposition with written language.

The response of interest during reading are often saccadic eye movements, which are hypothesized to reflect overt shifts of visual attention and therefore permit the investigation of the spatial and temporal dimensions of the perception and cognitive processing of the stimuli as well as the oculomotor execution of saccades.

During the first two eras of reading research (see Introduction above; Rayner, 1978; Rayner, 1998), eye movements were directly observed by researchers. This very superficial observation was of course inappropriate for the many subtle eye movements that occur within even small timeframes. It was not until the mid-1900s that eye movements, particularly during reading and image viewing, were systematically and non-intrusively recorded by optical eye-tracking devices (e.g., see Yarbus, 2013). Eye trackers differ with regard to their spatial and temporal resolutions. Modern scientific eye trackers record gaze positions at rates of up to 2000 Hz, where sampling rates below 1000 Hz are impractical for the investigation of fixational eye movements. Although technical specifications have changed, the optical eye tracker is still the most common apparatus for studying eye movements in reading to this day.

Technically, most optical eye trackers emit (invisible) infrared light, which is reflected by the cornea. The distance vector between the center of the iris and the corneal reflection can be determined to estimate the fixation location on the screen. As the corneal reflex differs greatly between individuals, devices, experimental settings, etc., calibration of the eye tracker is necessary before (and sometimes during) each experimental session. Likewise, data quality critically depends on proper calibration.

1.3 Cognitive, Linguistic, and Oculomotor Effects During Reading

There is ongoing debate regarding the extent to which eye movement behavior is affected by low-level oculomotor factors versus higher-level cognitive processes (Kliegl et al., 2006). Some researchers argue that eye movements are primarily guided by visual or oculomotor factors (e.g., see Adeli et al., 2016; Vitu et al., 1995), while others emphasize the role of cognitive and/or linguistic processes (e.g., see Just & Carpenter, 1980; Reichle et al., 2010). Over the previous decades, a lot of research has been accumulated that demonstrates the relevance of cognitive, (psycho-)linguistic, and oculomotor aspects of sentence reading and the view has emerged that neither low-level nor high-level processes alone can account for all variability and complexity of eye movements in reading (Eskenazi & Folk, 2016; Kliegl et al., 2006). In the following, I will summarize the most relevant experimental effects that a complete theory of reading should ideally account for, or that theory-driven models should be able to predict.⁷

⁷For brevity and since they are not uniquely applicable in reading research, purely perceptual effects (e.g., stimulus quality) are not considered in this overview.

1.3.1 Oculomotor Effects

Arguably, the most direct influence on eye movements in reading (i.e., the spatial dimensions of saccades and resulting fixation locations) comes from *oculomotor processes*, i.e. the execution of eye movements itself. Across many different studies and contexts, and regardless of the linguistic qualities of the stimuli, experimental effects have been established in the literature that show very stable effects independent of the actual displayed text. These include *preferred viewing/landing position* (PVP/PLP; Rayner, 1979) and *(inverted) optimal viewing* ([I]OVP; O'Regan, 1990; Vitu et al., 2001) effects.

Preferred viewing position. The PVP effect is a statistical bias for the eyes to land on locations at a relatively consistent distance from the launch site. Assuming the center of a word to be the “optimal viewing position”, the PVP manifests as a tendency to overshoot close word centers and overshoot farther ones. Upon statistical analysis of targets at different eccentricities, McConkie et al. (1988) demonstrated that saccade amplitudes follow a pattern of systematic bias due to the *saccadic range error*, and unsystematic *oculomotor noise*, still a very popular approach to model oculomotor error in reading models (e.g., see Engbert et al., 2005; Reichle et al., 1998).

(Inverted) optimal viewing position. Besides this bias or preference to execute saccades at certain amplitudes, it has also consistently been shown that the initial fixation location within a word can have a strong influence on fixation durations (IOVP) and refixation probabilities (OVP). Specifically, when the eyes initially land on non-central word locations, compared to landing on the word center, it is more likely that the saccade is executed slightly earlier and that a refixation follows. Both have been interpreted as evidence that the center of a word is the optimal viewing position, since it will be perceived more efficiently by the fovea. Consequently, earlier saccades and more likely refixations may be countermeasures for correcting inefficiently placed initial (or “misplaced”) fixations.⁸

1.3.2 Cognitive Effects

Cognitive effects on eye movements in reading have been extensively studied in cognitive psychology. These include but are not limited to effects of *attention*, *working memory*, and *parafoveal processing*.

Attention. Attentional processes play a crucial role in guiding eye movements and allocating visual processing resources to relevant information in the text. For example, *inhibition of return* (IOR; Posner & Cohen, 1984), a prominent effect of visual attention, positing that it takes more time to return to a previously attended location than to a novel fixation location, has also been demonstrated in reading for both adults and children (e.g., see Chasteen

⁸The error-correction hypothesis, however, is hardly possible to test experimentally because saccade targets, i.e. the intended landing positions, are not observable using eye tracking.

& Pratt, 1999; Eskenazi & Folk, 2016; Parker et al., 2020; Rayner et al., 2003; Slattery & Parker, 2019). Another convincing demonstration of the role of attention in reading is *mindless reading*, i.e. when readers do not fully attend to comprehension of the text being read. Mindless readers tend to make more and longer fixations (Luke & Henderson, 2013; Reichle et al., 2010), especially when fixations are close to the word center (i.e., the optimal viewing position, see Nuthmann & Engbert, 2009; Nuthmann et al., 2007). Effects of attention on reading are, however, difficult to isolate from other effects, as we can assume that it is a necessary precursor for efficient linguistic processing.

Working memory. Another important concept of cognitive psychology, *working memory*, has been found to be associated with eye movement characteristics in reading tasks. Higher working memory scores have been linked to longer saccades during reading and longer fixations during scene viewing (Luke et al., 2018). Other studies have shown that working memory load affects eye movement behavior, with a shortage of working memory resources leading to deficits in attentional control and eye movement coordination (e.g., Azuma et al., 2014). This suggests that working memory plays a role in the allocation of attention and the planning of eye movements during reading and visual exploration.

Often considered part of working memory, *executive functions*, which encompass cognitive processes such as inhibition and cognitive flexibility, have also been found to influence eye movements during reading. Reduced reading comprehension has been associated with executive function deficits in both dyslexic and healthy young readers (Georgiou & Das, 2016; Locascio et al., 2010). In Parkinson's disease, a neurological disorder with significant impact on working memory, executive dysfunction has been linked to alterations in reading speed and eye movement patterns (Stock et al., 2020).

Parafoveal processing. While it is undisputed that processing of the foveal information affects reading, there is ongoing debate about the theoretical embedding of *parafoveal processing*. The existence of parafoveal-on-foveal effects would speak for the parallel (vs. serial) processing of words and challenge the more constrained assumption of serial processing (see also Section 1.4 for an overview of serial and parallel processing models), at least for models with strict coupling of attention and gaze.⁹ For example, there is evidence that semantically related parafoveal words can facilitate the processing of foveal words, leading to fewer errors and faster naming times (Rusich et al., 2020). Conversely, when the perceptual or processing span is limited by foveal processing load or a gaze-contingent boundary paradigm, parafoveal processing is affected, resulting in effects on saccade targeting and attenuated preview benefit (e.g., see Henderson & Ferreira, 1990; Risse, 2014; Risse & Seelig, 2019; Vasilev & Angele, 2017; Zhang et al., 2019).

⁹There are serial-attention shift models, where attention lags behind fixation location, which is the case for E-Z Reader (Reichle et al., 1998, see Section 1.4) and derivative models. These can account for some parafoveal-on-foveal effects without challenging the strict sequential attention allocation.

1.3.3 Psycholinguistic Effects

Finally, in addition to the oculomotor and cognitive effects on eye movements in reading, there has been research on the psychological and linguistic processes that must take place in order to understand written language. Interfacing with the lower-level cognitive aspects described in the previous subsection, they are necessary to construct a mental representation of the structure and content of the text, although there is disagreement about the linking between them and the resulting relation to the observed eye movements. The most relevant aspects include lexical (incl. phonological, orthographic, and morphological), syntactic, and semantic effects, all of which are briefly introduced in the following.

Lexical effects. According to Kliegl et al. (2006), the “big three” lexical variables with significant effects on eye movements in reading (particularly fixation durations) are *corpus frequency*, *word length*, and *predictability*. Even though, when considering popular text corpora, these measures are significantly correlated (e.g., see Balota et al., 2007; Brysbaert & New, 2009; Heister et al., 2011; van Heuven et al., 2014) and therefore not completely independent, they all have significant individual contributions to eye movements, even when controlling for the respective others. Furthermore, *orthography*, *phonology*, and *morphology* have been shown to significantly affect eye-movement statistics.

Frequency. The *word frequency* or *corpus frequency* is a measure of how frequently a word occurs in natural language, or more specifically within a given text corpus. The frequency is simply determined by counting its occurrences in a text corpus. It has a significant impact on eye-movement behavior during reading. For example, words with high corpus frequencies receive shorter fixations, and are skipped and refixated more often compared to low-frequency words (Inhoff & Rayner, 1986; Kliegl et al., 2004; Rayner & Duffy, 1986). This suggests that highly frequent words require less processing because readers have more experience with them, facilitating lexical access in foveal and parafoveal vision.

Word length. The physical variable *word length* has a maybe more trivial effect: When words are longer, they take up more area of the fovea or, for longer words, even exceed the fovea. Therefore, longer words are fixated longer, receive more refixations, and are skipped more often than shorter words (Kliegl et al., 2004). Likewise, shorter words allow for more foveal and parafoveal processing of surrounding information, which has been shown to affect following saccade amplitudes and fixation durations (e.g., see Morris et al., 1990; O’Regan, 1979; Rayner, 1979).

Predictability. Although the corpus frequency of a word can be seen as a general proxy for the average exposure to a word, and hence of its identifiability, the same word can be more or less predictable under different conditions or in different contexts. The *predictability* of a word is the probability of a word being correctly guessed when absent. Predictability norms must be collected experimentally because they can be expected to differ between contexts,

task demands, and readers. When controlling for word length and corpus frequency, effects of predictability are additive and can have qualitatively similar and quantitatively stronger effects on eye-movement measures compared to word frequency (Kliegl et al., 2004; Rayner et al., 2004), also confirmed by results of Kretzschmar et al. (2015) who found that only predictability, but not corpus frequency, has an effect on N400 amplitudes in event-related potentials (ERPs).

Orthography. Another factor that contributes to the identifiability of a word is its orthography, or more specifically its *orthographic similarity* (Andrews, 1997; Coltheart et al., 1977). This refers to the number of words in a lexicon that can be constructed from a word by replacing, adding, or deleting a single letter. Although orthographic similarity has consistently been found to facilitate lexical decision (see Andrews, 1997, for a review), it has been found to increase fixation durations in reading. Consequently, words are fixated longer if they are orthographically similar to primes (such as preceding words; e.g., see Paterson et al., 2009).

Research on the N400 amplitudes in ERPs suggest that high orthographic neighborhood size increases processing demands (Holcomb et al., 2002) but only when attention is not shared with concurrent tasks (Rabovsky et al., 2019). Effects of orthographic similarity are especially strong if words have neighbors with significantly higher or lower frequency, which results in a spillover of that neighbor's frequency effects on the target word (Perea & Pollatsek, 1998). Relatedly, there is some evidence by a follow-up study by Pollatsek et al. (1999), that readers tend to make more regressive saccades to the target, presumably in order to correct the erroneous identification.

Phonology. Even if orthography affects eye movements in reading and processing of written language, there is still no consensus about whether that influence is direct or mediated by additional serial or parallel processing such as for the phonology of a word. A classical example is a study by McCutchen and Perfetti (1982), who showed that silent reading of tongue twisters tends to be slower than reading of control sentences. Though replications and follow-up research had mixed results (e.g., Daneman et al., 1995), phonology at least has the potential to modulate effects of semantics and orthography (e.g., Rayner et al., 1998). Moreover, when controlling for orthographic and other features of words, Inhoff and Topolski (1994) and Sereno and Rayner (2000) found that fixation durations were increased when words were phonologically irregular, such as *pint*, the so-called *spelling-to-sound regularity effect*.

Morphology. There are different views regarding the role of the morphological representation of a word. While some researchers attribute morphemes a linking function between form and meaning (Marslen-Wilson et al., 1994), others would attribute it a role at the level of orthographic processing (Rastle & Davis, 2008). A very common method in the research of morphology effects in reading is to prefix, suffix, or compound words and compare eye

movement statistics on these morphemes compared to different morphemes or the word root. Unsurprisingly, since the morphemes are longer than their root, this increases gaze durations, refixation probabilities etc. However, after controlling for word length, there are still some interesting findings, such as in Finnish, where the frequencies of the individual lexemes of compound words have been shown to have independent contributions to gaze durations, but non-additive effects for longer compounds (Pollatsek, Hyönä, & Bertram, 2000; Pollatsek, Tan, & Rayner, 2000), where the overall frequency of the compound seems to drive the frequency effect, which the authors interpreted as evidence for a dual-route word identification process for compound words.

Syntactic effects. In addition to lexical effects of a word, the syntactic structure of a sentence can affect eye movements, too. The effects range from spatially demarcated effects to more global adjustments of eye-movement control (Huestegge & Bocianski, 2010). While some phenomena such as garden-pathing can have severe effects on eye movements, most effects of sentence processing (e.g., Jäger et al., 2020) are generally of smaller magnitude compared to lexical effects (Boston et al., 2008).

One reason for the generally small magnitude of syntactic effects on eye movements may be the popular rationale of the *dominant interpretation*, i.e. the reader extracts syntactic information about the sentence as it is being read and sticks to the dominant parse. Effects of language processing difficulty on eye movements are mostly only then observed once new information challenges the dominant interpretation (Frazier & Rayner, 1987; Rayner et al., 1983). If such misanalysis due to syntactic category ambiguity occurs, however, it may be more costly for the reader to reconstruct the syntactic representation than to resolve a lexical-semantic ambiguity (Jones et al., 2012).

Another piece of evidence in support of the dominant interpretation hypothesis are garden-path sentences. In several studies (e.g., Frazier & Rayner, 1982; Rayner et al., 1983), it has been demonstrated that reading a sentence like *The old house their young* will lead to significantly increased reading times, in particular on regions that challenge the previously dominant parse. According to the authors, in the example, the reader will establish a syntactic representation with *old* as an adjective and *house* as a noun, as in *The old house was demolished*. However, when the reader encounters *their*, this dominant interpretation is challenged. Instead, *old* must be interpreted as a noun and *house* as a verb, as in *The old [people] house their young [children]*.

The role of memory in sentence processing has been proposed very early in psycholinguistics by Miller and colleagues (e.g., Miller, 1962; Miller & Isard, 1964). Especially, when readers process very long sentences, it may be that processing words later in the sentence require the retrieval of dependents read very early in the sentence. For example, a sentence like *The mouse that the cat that the dog chased saw ate quietly* could be difficult to interpret, especially because three nouns (*mouse*, *cat*, and *dog*) must be matched with three verbs

and when the first verb is read, all three subjects are still unmatched. Presumably, this can cause *memory interference* and thereby pose additional cognitive demands when processing the respective verbs (Gibson, 1998, 2000; Lewis et al., 2006). However, most evidence for memory interference has not been collected in the context of eye movements and there, evidence is inconclusive: Even though readers do take more time to read and make more regressions in more complex sentences involving difficult memory retrievals (e.g., Gordon et al., 2006; Jäger et al., 2015; Lee et al., 2007; Mertzen et al., 2023), regressions are far less common in those trials than a strong eye-mind coupling would predict (Just & Carpenter, 1980), suggesting a much more complicated influence of memory retrieval on eye movements in reading.

Semantic effects. In two experiments, Rayner et al. (1983) showed that reading times were more influenced by syntactic processing demands than by semantic processing but the latter had a stronger influence on the final interpretation of the sentence. In the following decades, there has not been very convincing evidence in favor of semantic effects on eye movements in reading (e.g., see Luke & Henderson, 2016; Rayner & Morris, 1992). Though some others suggested that semantic effects potentially manifest later during reading, not during first-pass reading, which is often the main focus of reading research (Weiss et al., 2018) and could therefore simply be a methodological problem. Further research beyond first-pass and single-sentence reading is necessary in order to evaluate this hypothesis.

1.4 Computational Models of Reading

The accumulation of experimental evidence has motivated the development of different theories about the reading process, some of which have been implemented in testable computational models. Computational models have emerged as a powerful tool to test theories across different scientific disciplines. By deriving a model from the theory, the modeler has to transfer theoretical assumptions into an explicit form (Fum et al., 2007), making the theory as a whole testable. It could even be argued that a theory without a model implementation is useless because its assumptions are unfalsifiable and thus challenge the scientific value of the theory. A computational model is typically a computer implementation of a model, making it possible to run efficient simulations of the model outcome (i.e., the behavioral response for cognitive models).

As typically in the behavioral sciences, reading patterns are often analyzed by focusing on the commonalities derived from aggregating data across trials. While the rationale usually is to decrease the “noise” caused by the inherent stochasticity, it also removes any sequential information within trials, that could otherwise be of explanatory value, for both the inference method and theory development. This is especially critical for eye movement research, where stimuli are (visually) perceived and the behavior of interest (eye movements) is generated by

the same system, i.e. the eye. Any observable eye movements are hence possibly influenced by the sequence of preceding eye movements on the same trial. As a consequence, process-oriented computational models considering the complex dynamics of fixation sequences have been particularly successful in reading research. However, there are also models with a stronger focus on modeling processing demand and less so the eye movements, particularly in linguistics.

1.4.1 Psychological Reading Models

Psychological models often have a strong focus on cognitive domains with a predominantly psychological research tradition, such as attention, vision, and decision-making. A main distinction between contemporary psychological models of reading is the assumption of serial vs. parallel processing of words around foveal vision, grouping them into *serial-attention shift* (SAS) models and *parallel graded attention* (PGA) models. This distinction has been widely accepted (for reviews, see Engbert & Kliegl, 2011; Reichle, 2011) but of course oversimplifies the characteristics of these models. Here, I will summarize key aspects of *SWIFT* (Engbert et al., 2005), *Glenmore* (Reilly & Radach, 2002; Reilly & Radach, 2006), *E-Z Reader* (Reichle et al., 1998), *OBI-Reader* (Snell et al., 2018), and the *superior colliculus model* (Adeli et al., 2016).

Superior colliculus (SC) model (Adeli et al., 2016). Adeli et al. (2016) were the first to present a reading model, which solely depends on oculomotor processes and works without a lexicon, i.e. without knowledge of lexical, orthographic, syntactic, or semantic information. Based on the central role in vision of the superior colliculus (SC), the image-based processing model accumulates early visual input signals in order to generate saccadic eye movements, mimicking basic neural circuitry of the SC. As visual input enters the system, the model computes a luminance-contrast saliency map, connected to an oculomotor map, representing ocular motoneuron populations. The most active region of the resulting motor map determines the saccade. Given that oculomotor effects are the most reliable and sizable effects in eye movements during reading, it is not surprising that the model performs well with regard to eye movement statistics, especially word-length effects on skipping probabilities, PVP effects, saccade amplitudes, and IOVP effects (Adeli et al., 2016). However, given the lack of more detailed cognitive and linguistic processing, and the strongly constrained focus on oculomotor effects, there has not been published any evidence that the model is able to predict higher-level linguistic processing effects such as garden-path effects. Moreover it does not consider temporal aspects of eye movements, such as fixation/saccade durations or saccade velocities.

E-Z Reader (Reichle et al., 1998). In 1998, Reichle et al. published the first implementation of the E-Z Reader model family, which has been augmented over the decades

many times (e.g., Reichle, 2011; Reichle et al., 1999, 2003, 2012), and was the first model that aimed at providing a complete account for eye movements during reading. It considers perceptual, cognitive, and oculomotor processes during sentence reading. Siding with the predominant view of the experimental and modeling literature at the time, E-Z Reader assumes serial (non-parallel) processing of words. Compared to other serial-attention shift models such as *Reader* (Just & Carpenter, 1980), which predict a strong link between mind and eye, however, E-Z Reader allows for more variability in the execution of eye movements during reading.

All processes within E-Z Reader are discrete states, each of which, upon initiation, is assigned a specific fixed or stochastic duration. In its initial configuration, the model encompasses three oculomotor processing stages: (1) a labile stage (M_1), (2) a non-labile stage (M_2), and (3) a saccade-execution stage (M_3). Word processing within the model also encompasses three sequential stages: (1) attention allocation, followed by the lexical stages of (2) familiarity checking (L_1), and (3) lexical completion (L_2). The conceptualization of these two lexical processing stages draws inspiration from the dual-process theory of recognition, as articulated by Reichle (2011).

Oculomotor processes and word processing can operate in parallel but motor processes must be triggered by word processing. When a word is attended to, it undergoes two consecutive lexical processing stages, L_1 and L_2 . When the familiarity check L_1 has concluded, the oculomotor programming cascade starts (M_1) in parallel to the lexical completion stage (L_2). Attention is shifted to the next word $n + 1$ once L_2 completes, which could be before or after the execution of the saccade. Once M_1 concludes, the programmed saccade is executed. If for some reason (such as in parafoveal processing) L_1 on a new attention target completes before a concurrent M_1 , the saccade is cancelled and reprogrammed to a new saccade target ($n + 1$ in reference to the word whose L_1 completed). Even though attention (and the saccade target) is serially shifted to the right, there is additional stochasticity in the eye movements due to an implementation of the McConkie et al. (1988) model of saccadic range error and oculomotor noise.

Despite some deliberate simplifications of oculomotor control and cognition, the overarching objective of the model has been demonstrated to replicate many of the fundamental effects observed during reading with a minimal set of underlying assumptions. This includes but is not limited to parafoveal-on-foveal, frequency, word-length, predictability, saccade-amplitude effects. A noteworthy limitation of the model, which is due to the strictly incremental attention shift, is that the model does not predict long-range regressions, which is problematic when fitting experimental data, where such saccades do occur.

Glenmore (Reilly & Radach, 2006). The Glenmore model was originally published by Reilly and Radach (2006). Its essential components encompass a visual input module, a word processing module that permits parallel processing of multiple words, a central fixa-

tion point, and a saccade generator responsible for generating saccadic eye movements. The input vector represents the current perceptual span and encodes the visual arrangement. During reading, information is passed to the saliency map and a linguistic processing module, which implements operations at both the letter and word levels using an interactive activation model. Activation values are computed as a cumulation of (a) bottom-up visual activation originating from the input units and (b) top-down letter and word activation. The temporal dynamics of activation at the word level depend on competitive inhibition from adjacent words and the frequency of fixated words. Additionally, words exert feedback onto the letter level through inhibitory connections, thereby retarding the decay of letter units.

SWIFT (Engbert et al., 2005). Engbert et al. (2005) introduced the SWIFT model of eye-movement control in reading, which was mainly motivated by the lack of alternative models allowing for parallel word processing. In fact, parallel processing in SWIFT is not a constraint of the model architecture but a result of parameter configuration. It can produce eye movements under the assumption of (approximately) serial processing but Engbert et al. (2005) demonstrated that the parallel configuration provides a better fit to experimental data.

In the model, each word of the sentence is associated with an activation value, which changes over time from 0 to some threshold during the *lexical processing* stage and from threshold back to 0 during the *post-lexical processing* stage. Depending on the concrete model implementation, either the threshold or the processing rate of each word is modulated by its corpus frequency and word predictability, resulting in less frequent or less predictive words requiring more processing time than more frequent or more predictable words. Only words that are within the *processing span*, which is centered around the current fixation location at a given time, are processed.

A saccade timer cascade controls the programming and execution of saccades. At first, a global timer starts, which in turn starts the labile saccade programming stage as soon as the global timer reaches threshold. When the labile timer reaches its threshold, a saccade is programmed to a word target probabilistically determined on the basis of relative word activations at that time, i.e. every word can be selected as target at any point in time but the word with the highest activation at the time the labile saccade stage concludes, very likely wins. The programmed saccade, however, is not executed until the non-labile saccade stage, following the labile stage, concludes. Finally, the saccade is executed and a new global timer is started. The global timer may be restarted if a saccade ends before the global timer reaches threshold. Likewise, the labile timer may be restarted if the global timer reaches threshold before the labile timer does. Consequently, the saccade programming can be canceled if the labile stage is prevented from reaching threshold. It is important to note that the saccade timer cascade and word activation field are mostly independent. There are only slight interactions by means of foveal load on the inhibition of the global timer. Otherwise, saccade timing is relatively independent of target selection.

OB1-Reader (Snell et al., 2018). Another model with particular focus on the interaction of word recognition and eye-movement control in sentence reading is *OB1-Reader* (Snell et al., 2018). The model can reliably recognize words in text and reproduce orthographic effects such as orthographic neighborhood size. It incorporates a spatiotopic sentence-level representation, allowing for parallel processing of multiple words. The model's attentional distribution adapts to the skill of the reader and the difficulty of the text, increasing after successful recognition and decreasing after failure. OB1-Reader accounts for various reading phenomena, including word length, frequency, and predictability effects, as well as orthographic parafoveal-on-foveal effects (Snell et al., 2018).

1.4.2 Linguistic Reading Models

Unlike psychologically motivated models of reading, linguistic approaches tend to focus on processing demands in the form of effects on fixation durations. Therefore, they often neglect spatial aspects of reading and are rarely applied to sequences of fixations and saccades. They are, however, frequently applied to isolated fixation durations or reading times from self-paced reading experiments. There, the two theories of *memory retrieval* (Lewis & Vasishth, 2005) and *surprisal* (Levy, 2008) have been particularly successful in predicting and explaining language processing demand.

Cue-based memory retrieval (Lewis & Vasishth, 2005). According to the ACT-R-based (Anderson & Lebiere, 1998) cue-based memory retrieval theory by Lewis and Vasishth (2005), when words are processed, their linguistic features (e.g., syntactic category, locality, etc.) are encoded in memory. If the reader subsequently encounters a word that requires the attachment of a dependent, such as a verb requiring a subject, a cue-based retrieval is triggered. Words in memory whose features match the retrieval cues are potentially retrieved. The degree to which memory features and retrieval cues match, determines the memory activation strength and consequent retrieval latency. The retrieval stops as soon as any word has been matched, which will often be the word with the best feature-cue match. However, if many words share the same features, the *fan effect* slows down processing of all these words, resulting in slower overall processing of these words and potentially more stochasticity and misretrievals. The model has been demonstrated to reproduce and explain effects of sentence length, garden-path reanalysis, and structural complexity (Lewis & Vasishth, 2005). While the model provides a very successful approach to syntax effects of sentence reading, it is not frequently applied to fixation sequences. This is mainly due to the fact that the model only predicts processing times and not saccadic eye movements.

Surprisal (Levy, 2008). The *surprisal* theory by Levy (2008) proposes a simple information-theoretic characterization of processing difficulty as the processing demand incurred by resource reallocation during probabilistic sentence comprehension. It assumes

that readers draw on a probabilistic grammar, which allocates a probability to all well-formed sentence structures that can follow from a given processed sentence portion. Mathematically, surprisal is the negative log-probability of expecting a word in a given context, resulting in high values for lower probabilities and a theoretical lower bound of zero for a completely unambiguous word. The greater the distance between the predicted word and the target word, the higher the surprisal and consequently the processing difficulty. Surprisal has been shown to predict fixation durations and skipping probabilities in various eye-tracking studies (e.g., see Ankenier et al., 2018; Balling & Kizach, 2017; Boston et al., 2008).

1.4.3 Contemporary Integrated Models

Some researchers have identified that psychological and linguistic threads of research should be combined to provide a satisfactory account of eye movements during reading. Instead of reinventing the wheel, however, researchers should embrace the success of the individual models with regard to their specific domain. That is why the first attempts to develop integrated reading models have taken advantage of established models, which were then extended. Most importantly, these integrated models include Dotlačil (2018) and Über-Reader (Reichle, 2021).

ACT-R-based reader (Dotlačil, 2018). Dotlačil (2018, 2021) published a noteworthy implementation of an integrated model of eye movements in reading, which builds on EMMA (Salvucci, 2001) within the ACT-R architecture as the eye movement module. Dotlačil included an account for syntax processing using the Lewis and Vasishth (2005) model of cue-based memory retrieval and demonstrated that model parameters can be reliably fitted within a Bayesian inference framework. Due to the EMMA-based oculomotor module, as discussed in Engelmann et al. (2013), the model comes with the severe limitation that no long-range regressions are within the predictive and explanatory scope of the model.

Über-Reader (Reichle, 2021). Based on the E-Z Reader architecture, Reichle (2021) and Veldre et al. (2020) introduced a reading model that is designed to predict and explain various additional aspects of sentence reading, such as detailed word identification, syntax parsing, discourse, and semantics. Even though, in his book, Reichle (2021) provided some high-level details about Über-Reader (such as the Lewis and Vasishth (2005) memory-retrieval model), the model lacks an accessible computer implementation and sufficient mathematical details for an independent implementation, it remains unclear to this date how the model can actually generate quantitative predictions for experimental data and how it compares to competitor models. In line with the shortcomings discussed in the previous subsection, this currently precludes the model from further consideration and systematic

model comparison.

1.5 Common Shortcomings of Contemporary Reading Models

As mentioned before, there has been a long-standing debate, mostly between psycholinguists and cognitive psychologists, about the link between eye-movement control and language processing. This is demonstrated by the focus on either in the development of the models mentioned above. Indisputably, there must be a bottom-up influence of eye-movement control on language processing because textual information can only be processed if at least partially perceived. So if for some reason the eyes move in a way that hinders visual identification of the words of a sentence and the true content cannot be reliably inferred, then language processing is severely compromised. But what happens when processing difficulty occurs in higher cognition or language comprehension? Are there likewise top-down mechanisms that enable the reader to execute targeted saccades to resolve the difficulty or is it all just chance? After all, the possibility that a processing difficulty is resolved after fixating the region in question does not necessarily mean that it was the processing difficulty that caused the saccade in the first place. Previous models were unable to address this question by leaving out either psychological or linguistic aspects of sentence reading, or neglecting some of the observable eye movements (in particular long-range regressions).

1.5.1 Explanatory Deficiencies

The most common critique that applies to all models introduced in the previous section is that they are explanatory deficiencies. Many models focus on a limited set of aspects, such as oculomotor effects (e.g., Adeli et al., 2016), word recognition (e.g., Snell et al., 2018), or memory retrieval (e.g., Lewis & Vasishth, 2005), while neglecting other processes. This approach may be useful to explain sufficiently large isolated effects of the respective domains on eye movements but if an explanation requires substantial oversimplification and neglect of otherwise significant contributors to the same outcome, the models may provide inaccurate explanations or predictions, limiting their value for research. For example, if a purely oculomotor model and a purely memory-based model both predict a regression (or activation) of a word in a sentence, which explanation is closer to the truth? Combining the two theoretical accounts in different configurations in a single model with a single outcome could provide a much more detailed and accurate account for the observed behavior, such as detailed qualitative estimates for the relative contributions of the competing accounts.

Moreover, a model should always be able to predict an observation under some model configuration. If not, it can only provide predictions and explanations for a subset of all possible outcomes. For example, long-range regressions, which can occur especially in long sentences, cannot be predicted and explained by most serial-attention shift models like E-Z

Reader or derivative models. Not only are these models severely limited in their explanatory value for more complex sentence reading, but also are they required to be fitted to a restricted subset of the data, which is not necessarily representative of all observable behavior. Nevertheless, even if a model can theoretically predict some observation, it should also be able to reproduce qualitative patterns associated with established oculomotor, psychological, or linguistic effects. If not, fitting a model to experimental data containing such effects can cause the fitted model to exhibit implausible behavior for many otherwise likely observations in order to compensate for few otherwise unlikely observations.

1.5.2 Material

Another problem, particularly with psychological models neglecting linguistic phenomena, is that the models are optimized and evaluated on the basis of experimental fixation sequences from comparably simplistic sentence material. Many sentences in popular sentence corpora like the Potsdam Sentence Corpus (Boston et al., 2008) or the Schilling corpus (Schilling et al., 1998) exhibit very few (if any) sentences that would allow for higher-level psycholinguistic effects. This is mainly because (a) such effects are triggered by sentence constructions that are relatively rare in a natural language context, and (b) corpus sentences tend to be relatively short, precluding any “late” cognition to affect the eye movements. In other words, if an effect is predicted to occur in higher cognition that depends on lower-level processing, it necessarily takes place later during reading. At that point in time, there may not be much of the sentence left to read, i.e. the effect cannot manifest in the observed behavior. To account for such late effects, it may be necessary to use longer sentence material and consider rereading measures in addition to the disproportionately preferred first-pass reading measures.

1.5.3 Parameter Inference

Computational models in general, including computational models of reading, typically have a number of free parameters that modulate the simulated response, e.g. by changing how information is processed or the response is executed. We can expect that these parameters differ between contexts, task demands, individuals, stimuli, or even between consecutive trials. While the simulation of a response using a known set of parameters is relatively straightforward, it can be challenging to infer parameters from observations. If a model can be tuned or fitted to experimental data so that the inferred model can reliably predict the observations, the implemented theory could provide a viable explanation for patterns observed in the data. However, unsurprisingly especially when models have many free parameters, reliable inference is extremely difficult because many parameter configuration can often lead to the same observation and many observations can be the cause of a single parameter configuration.

Another common problem across cognitive models is that data are inherently hierarchical, i.e. there are multiple observations by the same participant and there are multiple observations of the same stimulus across participants, all with their very unique potential impact on the observable response. Ignoring the hierarchical nature and individual differences in data can lead to severely biased results, especially for non-linear models.

Despite these issues, as in many domains of psychology, cognitive models often lack the use of more sophisticated parameter inference methods and rely on simpler methodology (if any). Even though hand-picking parameter values has become less common in reading models, the predominant parameter optimization on the basis of deviance measures can also produce biased and unreliable results. This may be especially true for biological systems such as humans, where model parameters can be expected to be non-linear and highly correlated, a circumstance referred to as *sloppiness* in biological systems (Apgar et al., 2010; Boehm et al., 2023). In the following Section 1.6, I summarize and compare relevant methods of parameter inference, before synthesizing the research demand that motivated this dissertation project.

1.6 Methods of Parameter Inference

The dimensionality of the parameter space is a significant consideration for the modeler for at least two important reasons. Firstly, models are especially informative if they can explain a range of stereotypical behavior with few free parameters because more of the observation can be explained in terms of the implemented theory, than in terms of the experimental data. Secondly, with every free parameter, there are exponentially more possible parameter configurations, which should be tested against the observed data, especially if the parameters are continuous variables. Given that already very small parameter variations can have severe effects on predicted patterns (e.g., see Engbert et al., 2005; Schütt et al., 2017), parameter inference must be carried out in an extremely careful and sensitive manner.

1.6.1 Hand-Picked Values

For many of the models discussed above, especially for psychological models of the 20th and early 21st century, parameter values were often hand-picked by the authors (e.g., Reichle et al., 2003). Even though this can produce plausible behavior, often the rationale behind specific parameter selections is obscure and irreproducible for different contexts. Moreover, it is possible if not likely that the hand-picked selection of parameters is actually not the optimum of the parameter space.

1.6.2 Parameter Optimization

There is a number of more objective and automated parameter inference methods. The most basic form, *grid search*, basically explores combinations of predefined parameter values across the different dimensions of the parameter space, while considering all combinations. As this includes the evaluation of a priori implausible parameter configurations, this can be computationally costly, which in turn practically limits the number of testable parameter sets. There exist a number of other techniques, which use a more sophisticated way of exploring the parameter space, such as Nelder-Mead, quasi-Newton and conjugate-gradient algorithms (Bélisle, 1992; Byrd et al., 1995; Fletcher, 1964; Nelder & Mead, 1965).

1.6.3 Goodness-Of-Fit

Both grid search and other optimization algorithms require some criterion of *goodness-of-fit*, which is commonly some *deviance measure* between experimental and simulated data, $f(\mathbf{X}, \mathbf{X}' | \theta)$, or a *likelihood* (see below for details). Deviance measures often are mean squared errors between summary statistics of experimental and of simulated data, and optimization algorithms are to find their minimum in parameter space Θ , so that

$$\hat{\theta} = \arg \min_{\theta \in \Theta} f(\mathbf{X}, \mathbf{X}' | \theta) . \quad (1.1)$$

The comparisons are usually not feasible at the level of single observations and must be aggregated across many simulated or experimental trials. This can add more computational cost to the process of parameter inference, in addition to the cost induced by the algorithm itself. Moreover, the choice of the summary statistic, on which to evaluate goodness-of-fit can introduce bias into the otherwise objective parameter inference (Schütt et al., 2017). For example, if parameters of a reading model are evaluated by least-squared error optimization based of single-fixation durations alone, the resulting model may not be a reliable predictor for spatial aspects or even temporal aspects other than single-fixation duration. Even when considering multiple summary statistics, the choice of weighting the different measures is non-trivial and can introduce additional bias in the criterion.

1.6.4 Likelihood

A more objective way of evaluating the goodness-of-fit of a model is its *likelihood*. A critical advantage over deviance measures is that a model likelihood can often be applied at the level of single observations, sparing the need to aggregate across trials to calculate summary statistics, which could result in a loss of statistical resolution. Generally, the model likelihood is the probability to simulate or observe data \mathbf{X} , given the parameter configuration $\theta \in \Theta$,

$$L_M(\boldsymbol{\theta} | \mathbf{X}) = P_M(\mathbf{X} | \boldsymbol{\theta}) . \quad (1.2)$$

If the data \mathbf{X} consist of several independent trials, the probability P_M is the product of each independent trial's probability, such that

$$P_M(\mathbf{X} | \boldsymbol{\theta}) = \prod_{x \in \mathbf{X}} P_M(x | \boldsymbol{\theta}) . \quad (1.3)$$

Note that, since the product of many relatively small probabilities is computationally not tractable, usually, the log-likelihood is evaluated instead, so that

$$l_M(\boldsymbol{\theta} | \mathbf{X}) = \log P_M(\mathbf{X} | \boldsymbol{\theta}) \quad (1.4)$$

$$= \sum_{x \in \mathbf{X}} \log P_M(x | \boldsymbol{\theta}) . \quad (1.5)$$

This also entails that, as noted before, a model must always be able to attribute a non-zero likelihood to all observations, i.e. offer an account for any observable pattern. Otherwise, if only a single trial results in a zero likelihood, the model likelihood for the entire data set \mathbf{X} will also be zero. Consequently, this precludes some reading models that cannot predict the entire range of observable eye movements.

Unlike deviance measures, within the domain of parameter optimization, the likelihood L_M is to be *maximized* and the resulting parameter configuration is the *maximum-likelihood estimate* (MLE; for a tutorial, see Myung, 2003),

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg \max_{\boldsymbol{\theta} \in \Theta} L_M(\boldsymbol{\theta} | \mathbf{X}) . \quad (1.6)$$

Practically, likelihood-based inference produces less biased and more reliable results than parameter optimization involving other criteria of goodness-of-fit, such as deviance measures. They are, however, not always mathematically tractable in analytical form. In these cases, there exist different approximation techniques, including but not limited to probability-density approximation (Holmes, 2015; Palestro et al., 2018; Turner & Sederberg, 2013), pseudo-marginal likelihoods (Andrieu & Roberts, 2009), and synthetic likelihoods based on summary statistics (Wood, 2010).

Sequential likelihood. In dynamical cognitive models (see Engbert, 2021, for an introduction), which consider sequential observations over time, $\mathbf{X} = \{x_1, \dots, x_n\}$, it is possible to formulate the likelihood of an observed sequence of events as a *sequential likelihood*,

$$L_M(\boldsymbol{\theta} | \mathbf{X}) = P_M(x_1 | \boldsymbol{\theta}) \prod_{i=2}^n P_M(x_i | \boldsymbol{\theta}, x_1, \dots, x_{i-1}) , \quad (1.7)$$

where each $P_M(x_i | \theta, \dots)$ depends on the previous states or steps in the sequence. This approach can be subsumed under the general methodological approach of *data assimilation* (Engbert et al., 2022; Reich & Cotter, 2015), which refers to the integration of complex mathematical models with time-series data (e.g., see Morzfeld & Reich, 2018). Given the general advantage of likelihood-based inference to consider data at the level of observations (rather than aggregating), sequential likelihood allows not only the evaluation of the final outcome of a trial but its temporal evolution. This enables much more reliable model fitting, due to the higher resolution of the data, and more direct testing of latent model assumptions.

1.6.5 Bayesian Inference

Like parameter optimization techniques, Bayesian parameter inference attempts to find plausible configurations for parameters $\theta \in \Theta$, given the observed data \mathbf{X} . The minimum requirements in order to use Bayesian parameter inference are a model likelihood $L_M(\theta | \mathbf{X})$, as introduced above, and a prior $Q(\theta)$. While the likelihood is a strictly model-specific evaluation of the probability of a parameter configuration given some observation, the prior probability is usually conceptualized as a-priori knowledge or belief about the distribution of the parameters. Essentially, it dictates which parameter values are most likely, possibly based on knowledge about the model implementation or previous analyses. What is sometimes neglected is the fact that priors thereby also guide the algorithm as to where in parameter space Θ to consider parameter configurations.

Instead of trying to find a single parameter configuration like maximum-likelihood estimation, Bayesian parameter inference converges on a posterior probability over the parameters, given the data,

$$P(\theta | \mathbf{X}) = \frac{Q(\theta) L_M(\theta | \mathbf{X})}{P(\mathbf{X})}, \quad (1.8)$$

which incorporates the uncertainty about the fitted parameter values (Nicenboim et al., 2023).

In most sampling methods, including the influential class of Markov Chain Monte Carlo (MCMC) samplers (see Gilks et al., 1995, for an introduction), the denominator $P(\mathbf{X})$, which is the likelihood of the data, integrated over all parameters, can be disregarded because the iterative samplers usually only evaluate the relative posterior between two proposals, which effectively only considers

$$P(\theta | \mathbf{X}) \propto Q(\theta) L_M(\theta | \mathbf{X}). \quad (1.9)$$

The regularizing property of a prior distribution $Q(\theta)$ is sometimes formulated as general criticism of Bayesian inference, given that constraining the parameter space can theoretically preclude the algorithm from converging on a plausible solution. Therefore, priors must be chosen accordingly to allow any plausible parameter value, sometimes resulting in the selection of broad and weakly informative priors (Schütt et al., 2017). Likewise, the regularization

has a considerable advantage over methods like parameter optimization or grid search, which is the exclusion or attenuation of a-priori implausible parameter configurations during the exploration of the parameter space. Under the consideration of the regularizing property of priors, they can decrease the computational cost of complex and highly-dimensional models by emphasizing parameter evaluation in more plausible regions and avoiding the sometimes costly evaluation of a-priori unlikely parameter configurations.

1.7 Summary

As summarized above, in the light of continuously more complicated theoretical frameworks in the cognitive sciences, contemporary reading models are facing severe shortcomings. This includes explanatory deficiencies for purely psychological or linguistic models, inadequate corpus material (mostly for psychological models), and inefficient parameter inference. Although there are at least two modeling approaches that attempt to integrate psychological and linguistic aspects of eye movements in reading (Dotlačil, 2018; Reichle, 2021), they are either not testable, are not implemented for likelihood-based inference methods, or preclude experimental contexts, in which long-range regressions are known or expected to occur. What follows are three consecutive studies, which report on the advances of the application of these principles in the context of reading research by using the example of SWIFT (Engbert et al., 2005), culminating in the integrated reading model SEAM, which considers basic concepts of linguistic sentence processing from the cue-based memory retrieval model by Lewis and Vasishth (2005).

The following Chapter 2 introduces a new version of SWIFT, a dynamical model of fixational eye movements during reading. It presents an innovative data assimilation approach, using a combination of approximative likelihood approaches for spatial and temporal aspects of fixation sequences. Bayesian parameter inference, facilitated by an adaptive MCMC technique, demonstrates reliable estimation of model parameters for individual subjects, thereby advancing computational models of eye-movement control in general and of SWIFT in particular.

The subsequent Chapter 3 discusses the limitations posed by model complexity and high dimensionality in the context of process-oriented models of reading. A Bayesian framework is proposed to address these issues for the SWIFT model. The approach is applied to experimental data across different conditions of geometrically altered text, capturing not only differences in reading between conditions but also between individuals. Statistical analyses of model parameters between experimental conditions provide insights into the experimental effects on model behavior and thereby provide an explanation for different empirical patterns in the experimental data.

In Chapter 4, the integrated model SEAM is introduced. It combines the SWIFT model

of eye-movement control with components from a sentence processing model. Despite the added model complexity and high computational demands, the integration becomes feasible through advancements in parameter identification. The integrated model is applied to a challenging data set from a retroactive memory interference study and demonstrates the successful reproduction of eye movement patterns arising from linguistic dependency completion processes in reading. SEAM represents a pioneering achievement in the integration of process models for eye-movement control and linguistic comprehension, with implications for a comprehensive understanding of natural language comprehension in reading.

This dissertation closes with a General Discussion (Chapter 5), where I summarize and discuss the most relevant results embedded in the context of the contemporary literature, and propose directions for future research.

Chapter 2

Bayesian Parameter Estimation for the SWIFT Model of Eye-Movement Control During Reading

This chapter has been published as: Seelig, S. A., Rabe, M. M., Malem-Shinitzki, 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, Article 102313. <https://doi.org/10.1016/j.jmp.2019.102313>

Abstract

Process-oriented theories of cognition must be evaluated against time-ordered observations. Here we present a representative example for data assimilation of the SWIFT model, a dynamical model of the control of fixation positions and fixation durations during natural reading of single sentences. First, we develop and test an approximate likelihood function of the model, which is a combination of a spatial, pseudo-marginal likelihood and a temporal likelihood obtained by probability density approximation. Second, we implement a Bayesian approach to parameter inference using an adaptive Markov chain Monte Carlo procedure. Our results indicate that model parameters can be estimated reliably for individual subjects. We conclude that approximative Bayesian inference represents a considerable step forward for computational models of eye-movement control, where modeling of individual data on the basis of process-based dynamic models has not been possible so far.

2.1 Introduction

Dynamical models represent an important theoretical approach to cognitive systems, in particular, if we seek to explain time-ordered behavioral data such as sequences of movements. In dynamical models, sequential dependencies between observations are naturally explained by underlying dynamical principles that unfold over time when the model is simulated numerically (Beer, 2000; Van Gelder, 1998). Examples for the dynamical approach can be found in many fields of cognitive research, triggered by early examples from motor

control (Erlhagen & Schöner, 2002; Haken et al., 1985) or decision field theory (Busemeyer & Townsend, 1993).

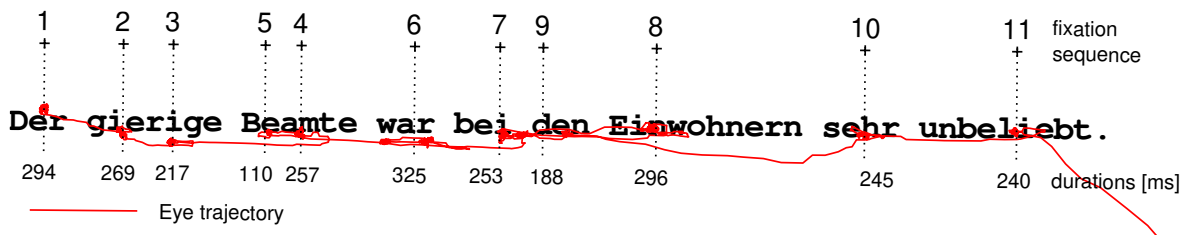
Dynamical models generate highly specific predictions on sequential data that include statistical correlations between the subsequent observations over time. As a consequence, parameter inference for dynamical models must be carried out with the fully dynamical framework of *data assimilation* (Law et al., 2015; Reich & Cotter, 2015). Here we investigate parameter inference in the SWIFT model of saccade generation during reading (Engbert et al., 2005), where the numerical computation of the model’s *likelihood function* will be the fundamental concept and main contribution of this work.

In the research area of eye-movements during reading, a number of competitor models has been proposed. These models implement alternative assumptions on the interaction of word recognition and saccade generation (see Rayner & Reichle, 2010; Reichle et al., 2003, for overviews). However, there is currently a lack of quantitative model evaluations using objective concepts. First, due to the number of different effects in experimental data, models were often compared qualitatively: Does the model reproduce an experimentally-observed effect or not? Second, in complex cognitive models, parameters were mostly hand-selected or fitted based on minimization of an arbitrary loss-function that quantifies the difference between experimental and simulated data. Third, typical models could not be fitted to data from individual subjects so far. However, explaining interindividual differences is an important aspect of model evaluation, which is precluded when fitting procedures are data hungry and require pooling of data over a large number of participants. Since model identification and model comparison are general problems in psychological and cognitive sciences, Schütt et al. (2017) recently proposed a likelihood-based, statistically well-founded Bayesian framework for parameter estimation in cognitive models. We will demonstrate the feasibility of this approach in the case of the SWIFT model for eye-movement control during reading.

In the following, the data assimilation framework will be applied to the SWIFT model of eye guidance in reading. The remaining part of this section consists of a short introduction to eye movement data and the specifics of likelihood functions for models of fixation sequences. In Section 2.2, we describe the details of the SWIFT model. A numerical approximation of the likelihood function is proposed and tested in Section 2.3. In Section 2.4, we use data from a set of readers to estimate SWIFT parameters and to model their interindividual differences. We close with a discussion of our results in Section 2.5.

2.1.1 Eye-Movement Control During Reading

Reading is based on successful word recognition, however, processing of words requires high-acuity vision that is confined to the center of the visual field (the fovea). Therefore, gaze shifts via fast eye movements (saccades) need to be generated to move words into the fovea

Figure 2.1*Sequence of Fixations During Reading*

Note. The scanpath indicates a series of fixations and saccades. Fixations are labeled by numbered dotted lines which indicate the horizontal positions. Numbers below the vertical lines are the corresponding fixation durations.

for word identification. From this general behavioral pattern, reading may be looked upon as an important example of *active vision* (Findlay & Gilchrist, 2003), which is the notion that eye movements form an essential component for almost all visual perception.

When we read texts, we perform 3 to 4 saccades per second, resulting in fixations on different words with durations between 150 and 300 ms, on average. An example is presented in Figure 2.1, where 11 fixations are placed on the words of a given sentence. Fixation durations range from 110 ms to 325 ms. In this example, some words are fixated more than once. In the case of an immediate second saccade to the same word as the currently fixated word, the event is called a *refixation* (e.g., fixations 3, a forward refixation, and 5, a backward refixation). Some words are not fixated during first-pass reading, corresponding saccades are termed *skippings* (e.g., word 6, the article “den”, was skipped in *first-pass reading*). Furthermore, it happens in roughly 5 to 10% of the fixations that a corresponding saccade returns to a previously passed region of text, which are called *regressions* (e.g., when word 6, the previously skipped article, receives fixation 9). Taken together, only about 50% of the saccades are moving the gaze forward from word n to the next word $n + 1$, which generates complicated *scanpaths* that are difficult to reproduce and predict by theoretical models of eye guidance during reading.

Eye movement research in reading has evolved into one of the fields of cognitive psychology that is strongly driven by computational models. Most of these models are based on simplified assumptions for several cognitive subsystems (e.g., oculomotor circuitry, attention and word recognition), while the core of the models is the orchestration of the subsystems to produce purposeful saccades for reading in a psychologically plausible framework. The way to this success has been paved by the E-Z Reader model (Reichle et al., 1998), a rule-based stochastic automaton model that is based on specific assumptions for the coupling of eye movements and visual attention. This model has been advanced over the years to include more and more specific assumptions (e.g., Reichle et al., 2009).

One of the major differences between existing models lies in the generation of different

types of saccades (forward saccades, skippings, refixations and regressions). While some models make explicit assumptions on saccade types or are built to have internal states representing saccade types, an alternative model considered here is motivated by the dynamical field theory of movement preparation (S.-i. Amari, 1977; Erlhagen & Schöner, 2002), which communicates the aspiration to form a general framework for human motor control. The SWIFT¹⁰ model (Engbert et al., 2002, 2005; Schad & Engbert, 2012) provides a coherent theoretical framework for reproducing all types of saccades that are observed during reading. Word processing maps to a distributed activation field that serves as a temporally evolving saccade targeting map. This model will be studied in detail with respect to parameter inference.

Given alternative theoretical models, model fitting and model comparisons will become an increasingly important topic in eye-movement research, as in cognitive science in general. So far, the minimization of ad-hoc statistical loss-functions has been used to obtain estimates for model parameters (e.g., Engbert et al., 2005; Reichle et al., 1998). For example, differences in word-frequency dependent distributions of fixation durations or skipping probabilities have been implemented as a measure of goodness-of-fit. We will replace these procedures by a Bayesian framework that exploits the likelihood function of the model.

Quantitative measures for eye movements during reading are characterized by strong interindividual differences (e.g., Risse, 2014). However, current computational models of eye-movement control could not reproduce and explain these obvious differences in human performance. It is a key message of the current work that the problem of modeling interindividual differences in reading using complex simulation models can be overcome when a likelihood-based framework of model identification, model parameter estimation, and model comparison is applied. We start with a discussion of the general concept of the likelihood function for dynamical cognitive models in the next section. The approximative computation of the likelihood function for the SWIFT model, which is the central contribution of the current work, is discussed in Section 2.3.

2.1.2 The Likelihood Function for Dynamical Cognitive Models

The key theoretical concept for the current study is the likelihood function (see Myung, 2003, for a tutorial), which is a quantitative measure of the plausibility of an observation under the assumption of a specific model M . We assume that the model depends on a set of parameters θ from parameter space Θ . In parameter inference, we are interested in the likelihood of the model parameter values θ for model M given the experimental data,

$$P_M(\theta \mid \text{data}) = P_M(\text{data} \mid \theta), \quad (2.1)$$

¹⁰Saccade Generation With Inhibition by Foveal Targets

where $P_M(\text{data}|\boldsymbol{\theta})$ is the probability of the data given model M with parameters $\boldsymbol{\theta}$.

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\text{ML}}$ is obtained by maximization of the likelihood function, i.e.,

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg \max_{\boldsymbol{\theta} \in \Theta} P_M(\boldsymbol{\theta} | \text{data}) . \quad (2.2)$$

In mathematical models of eye-movement control, a model must be evaluated against a sequence of fixations. Thus, the data is a time-ordered sequence of fixations $F = \{f_i\}$, where each fixation f_i is characterized by a position x_i on the line of text, a fixation duration T_i , and, depending on the model, also a saccade duration s_i between fixation $i - 1$ and fixation i .

In a dynamical model, fixation $f_i = (x_i, T_i, s_i)$ is generated from the sequence of previous fixations $f_1 \dots f_{i-1}$ under the control of the set of parameters $\boldsymbol{\theta}$ and, possibly, influenced by internal degrees of freedom $\boldsymbol{\xi}$, which will be discussed in Section 2.3. Since fixations are time-ordered, we can factorize the likelihood into a product of all fixations $i = 1, 2, \dots, n$, which are found in the experimental fixation sequence $F = \{f_i\}_{i=1}^n$, i.e.,

$$\begin{aligned} P_M(\boldsymbol{\theta} | F) &= P_M(\boldsymbol{\theta} | f_1, f_2, \dots, f_n) \\ &= P_M(f_1 | \boldsymbol{\theta}) \prod_{i=2}^n P_M(f_i | f_1, \dots, f_{i-1}, \boldsymbol{\theta}) , \end{aligned} \quad (2.3)$$

where $P_M(f_1|\boldsymbol{\theta})$ is the probability of the initial fixation starting at time $t = 0$. In typical experimental paradigms, however, this probability is one, since the experimental procedure determines the initial fixation position.

For complex cognitive models, the likelihood function can often be computed numerically. If numerical computation of the likelihood function is possible, we must be able to evaluate the likelihood for a large number of combinations of model parameter values $\boldsymbol{\theta}$ to find the maximum likelihood estimator, Equation (2.2), based on a given fixation sequence F .

For the implementation of numerical computations, it is advantageous to compute the log-likelihood, given as

$$\begin{aligned} l_M(\boldsymbol{\theta} | F) &= \log(P_M(\boldsymbol{\theta} | F)) \\ &= \sum_{i=1}^n \log(P_M(f_i | f_1, \dots, f_{i-1}, \boldsymbol{\theta})) , \end{aligned} \quad (2.4)$$

which prevents the addition of very small numerical values that typically occur for some of the additive terms $P_M(f_i|f_1, \dots, f_{i-1}, \boldsymbol{\theta})$ for the fixations f_i .

If we can compute the log-likelihood $l_M(\boldsymbol{\theta}|F)$ for model M efficiently using numerical simulation, then it will be possible to apply Bayesian parameter inference (see Gelman et al., 2013; Marin & Robert, 2007, for overviews). In Bayesian inference, we seek to compute the posterior distribution $P(\boldsymbol{\theta}|F)$ over the parameter vector $\boldsymbol{\theta}$ after the observation of the fixation

sequence F . In addition to the likelihood that represents constraints from the experimental data, we specify a prior probability $Q(\theta)$ that indicates our a-priori knowledge on the model parameters. The posterior distribution is given by

$$P(\theta | F) \propto Q(\theta) P_M(\theta | F), \quad (2.5)$$

where the constant of proportionality, which is the normalization constant of the posterior, can be omitted, if Markov Chain Monte Carlo (MCMC) methods are used (Gilks et al., 1995; Robert & Casella, 2013).

So far, we discussed the structure of the likelihood function for a single experimentally observed fixation sequence F . In a typical experiment, however, we obtain a set of fixation sequences F_s from a participant who read a corpus of S sentences ($s = 1, 2, 3, \dots, S$), i.e., the data set $\{F_s\}$ is composed of S fixation sequences. Since fixation sequences are statistically independent observations of the reading process, the numerical computation of the likelihood can be carried out independently for each fixation sequence F_s . This statistical independence can be exploited to accelerate computations via parallel evaluations of a large number of fixation sequences, which we will discuss in Section 2.4.

In summary, the likelihood function for dynamical models of sequential data factorizes as explained in Equation (2.3), which turns out to be basis for incremental numerical computation. If we implement the computation in an efficient way numerically, then Bayesian parameter inference is available using MCMC methods. Before we discuss and apply the MCMC framework, we introduce the SWIFT model in the next section. In Section 2.3, we present the numerical computation of the likelihood function. The MCMC simulation for Bayesian inference will be discussed in Section 2.4.

2.2 The SWIFT Model of Saccade Generation During Reading

Since word recognition is the key process driving eye movements during reading, a natural assumption is that the time-course of ongoing word processing is closely linked to target selection for saccades. In the SWIFT model, each word is represented by a separate activation variable (lexical activation) that is tracking the word's current progress in word recognition. The resulting set of lexical activations determines the probability for saccade target selection (so-called spatial or *where* pathway). Whenever a saccade is prepared, the set of lexical activations provides a flexible mechanism for target selection. As time evolves, the relative activations change, so that a continuous-time representation of the next saccade target exists.

Fixation times are adjusted to the fixated (foveal) word by an inhibitory mechanism (the temporal or *when* pathway). According to an influential proposal (Findlay & Walker, 1999) the spatial and temporal pathways of saccade generation are partially independent. The

SWIFT model is compatible with this view, in the sense that control of fixation duration and saccade target selection are basically independent, however, interactions exist due to the coupling of both pathways via the set of lexical activations.

2.2.1 Saccade Target Selection and Temporal Evolution of Activations

Each word m in a sentence of N_w words is represented by a time-dependent activation $a_m(t)$. The activation is initially increasing during lexical access (word recognition), and later decreasing during post-lexical processing. The set of activations $\{a_j(t)\}$, ($j = 1, 2, 3, \dots, N_w$) must be built up by parallel processing of words, which is the key assumption that distinguishes SWIFT from other models (e.g., Engbert & Kliegl, 2011; Reichle et al., 2003). If a saccade target has to be selected at time t , then the probability $\pi_m(t)$ for target selection of word m is given by the relative activation, i.e.,

$$\pi(m, t) = \frac{(a_m(t))^\gamma}{\sum_{j=1}^{N_w} (a_j(t))^\gamma}, \quad (2.6)$$

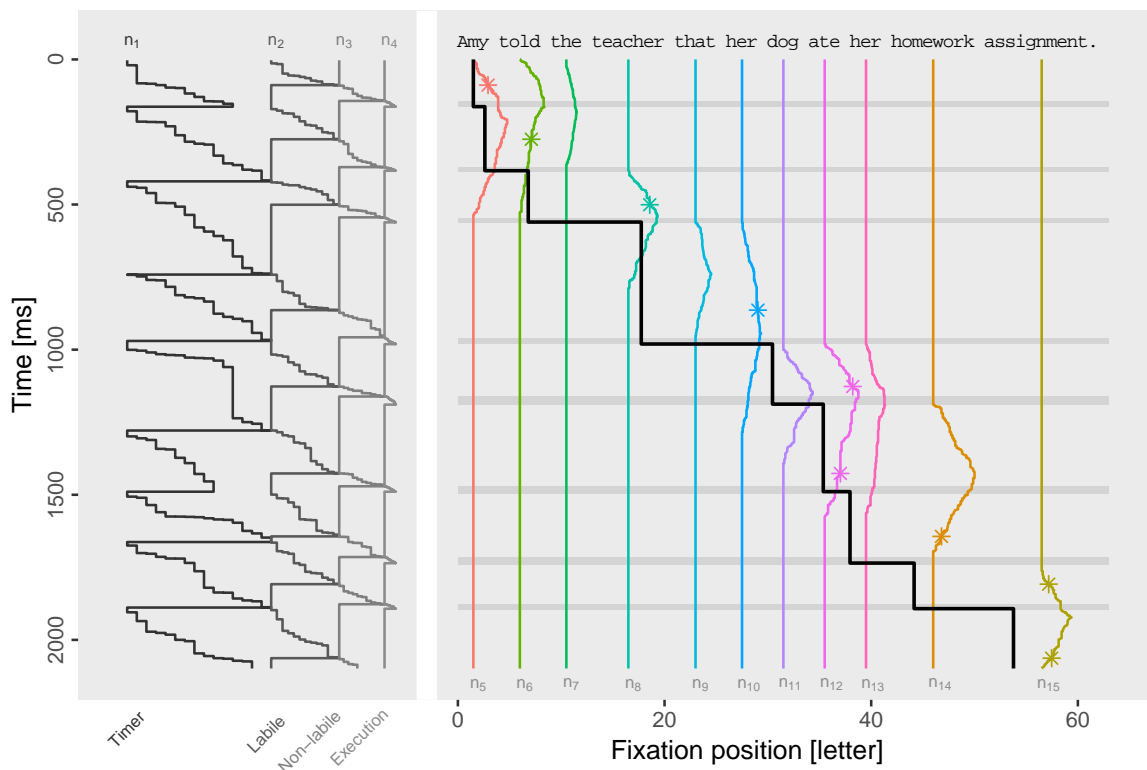
which is normalized as $\sum_{m=1}^{N_w} \pi_m(t) = 1$ for all $t > 0$. The parameter γ introduces a weighting of the set of lexical activations, so that switching between different selection schemes is controlled by a variation of γ :

$$\pi_m(t) \rightarrow \begin{cases} \text{winner-takes-all} & : \gamma \rightarrow \infty \\ \text{Luce's choice rule} & : \gamma = 1 \\ \text{random selection} & : \gamma \rightarrow 0 \end{cases} . \quad (2.7)$$

An example for a simulated scanpath and the full time-series of lexical activation is illustrated in Figure 2.2. As one can see from figure, all internal sub-processes of the model are implemented by discrete random walks. In the leftmost column, the saccade timer increases as a one-step process from $n_1 = 0$ up to a maximum number N_t with transition rate w_1 . The stepping rate was chosen as N_t/t_{sac} , so that the mean time to reach state N_t is the mean time inter-saccadic time t_{sac} of the model.

When the saccade timer terminates at state N_t , a new saccade timer run is initiated at $n_1 = 0$ and, at the same time, a labile saccade program start with $n_2 = 0$ until its threshold N_l is reached. If this labile program terminates, a saccade target is chosen (see asterisks in Figure 2.2). After the non-labile stage, which is described by state variable n_3 , the corresponding saccade (state variable n_4) is executed.

In addition to the saccadic processes, lexical activations are also described by discrete random walks (note, however, the increasing and decreasing parts in the case of lexical activations). Thus, all sub-processes saccade timing, labile and non-labile saccade pro-

Figure 2.2*Simulation Example for the SWIFT Model*

Note. The activation field (colored lines) determines the target selection probability $\pi_m(t)$ that evolves dynamically over time (running downwards). The resulting scanpath (fixation sequence) is indicated by the black line. Several random walks (grey, left) generate saccade timer intervals and labile and non-labile saccade latencies. The transition between labile and non-labile stage is the point in time for saccade target selection (asterisk). The saccade timer sends commands to the saccade programming cascade, but also receives inhibition during foveal load (visible shortly after 1000 ms in the example) and is reset for refixations (e.g., second fixation).

gramming, saccade execution, and change of lexical activations are represented as one-step stochastic processes between discrete states.

The state of the model at time t is given by the vector $n = (n_1, n_2, \dots, n_{4+N_w})$, where the components n_j represent the states of the subprocesses with transition rates w_j . Components 1 to 4 are saccade-related processes and additional stochastic variables n_5 to n_{4+N_w} are keeping track of the (post-)lexical processing of words. We assume a discrete-state, continuous-time stochastic process with Markov property, so that a one-step transition table describes all possible transitions between internal states (Table 2.1). In each of the possible transitions from state $n = (n_1, n_2, \dots)$ to $n' = (n'_1, n'_2, \dots)$ only one of the components n_i is changes by one unit, e.g., if the saccade timer generates a transition, then the model's internal change steps from $n = (n_1, n_2, n_3, \dots)$ to $n' = (n_1 + 1, n_2, n_3, \dots)$.

Table 2.1

Stochastic Transitions Between Adjoined States From $n = (n_1, n_2, \dots) \mapsto n' = (n'_1, n'_2, \dots)$

| Process | Transition to ... | Transition rate $W_{n'n}$ |
|--------------------|----------------------------|---|
| Saccade timer | $n'_1 = n_1 + 1$ | $w_1 = N_t/t_{\text{sac}} \cdot (1 + h a_k(t)/\alpha)^{-1}$ |
| Labile program | $n'_2 = n_2 + 1$ | $w_2 = N_l/\tau_l$ |
| Non-labile program | $n'_3 = n_3 + 1$ | $w_3 = N_n/\tau_n$ |
| Saccade execution | $n'_4 = n_4 + 1$ | $w_4 = N_x/\tau_x$ |
| Word processing | $n'_{4+j} = n_{4+j} \pm 1$ | $w_{4+j} = N_a/\alpha \cdot \Lambda_j(t)$ (for word j) |

A numerical algorithm for the simulation of a trajectory of the SWIFT model can be derived easily from our assumptions. The temporal evolution of the probability over the model's internal states is given by the master equation¹¹,

$$\frac{\partial}{\partial t} p(n, t | n'') = \sum_{n'} [W_{n'n'} p(n', t | n'') - W_{n'n} p(n, t | n'')] , \quad (2.8)$$

which is specified by the transition probabilities $W_{n'n}$ for transitions between state vectors $n \mapsto n'$ shown in Table 2.1 with initial condition $p(n'', 0)$, the probability of state n'' at time $t = 0$. When simulating a single trajectory, the system is in a specific state n with certainty and the transition probabilities determine both the waiting time distribution for the next transition and the relative stepping probability to the adjoined states given in Table 2.1, which will be explained below.

2.2.2 Temporal Control of Saccades and Foveal Inhibition

Gaze duration, defined as the sum of the durations of all immediately consecutive fixations on a word, is probably the best measure of required processing time for this word during natural reading (Rayner, 1998). Gaze durations and word recognition times depend linearly on the logarithm of the word's frequency (printed word frequency can be estimated from the word's occurrences in large text corpora). Since word recognition is the basis for text comprehension, an adaptive mechanism for the modulation of fixation duration by word frequency is essential for all models of eye-movement control.

In general, the required fixation duration for successful word recognition can be attained by two opposing mechanisms: The current fixation can be prolonged by inhibiting the next saccade or, alternatively, the word can be refixated to increase gaze duration. Experimentally, there is only a weak influence of word frequency on the mean first-fixation duration (Kliegl

¹¹The master equation can be interpreted as a conservation equation for probability (Gardiner, 1985; Van Kampen, 1992), where the temporal change of probability in state n on the left side of the equation equals the *gain* in probability for state n that is generated by transitions from neighboring states $n' \mapsto n$ and the *loss* in probability generated by transitions from n to neighboring states $n \mapsto n'$.

et al., 2004). In contrast, we find a strong effect of word frequency on the probability for refixation. Therefore, there is a preferred strategy for extending the processing time (gaze duration) via generation of a refixation. However, saccade-inhibiting processes can be assumed to contribute a weaker effect (compared to refixation) to the increase in gaze duration by prolonging the ongoing fixation (Engbert et al., 2002, 2005).

Motivated by these observations, the second central assumption in the SWIFT model is *random timing* of fixation duration with additional *foveal inhibition* (Engbert et al., 2002) that delays the start of the next saccade program to extend the current fixation. We assume that foveal inhibition modulates the transition rate w_1 for transitions between elementary steps of a random-walk that implements the saccade timer (leftmost column in Fig. 2.2), i.e.,

$$w_1 = \frac{N_t}{t_{\text{sac}}} \cdot \left(1 + \frac{h}{\alpha} a_k(t)\right)^{-1}, \quad (2.9)$$

where N_t is the number of states of the timer's random walk and t_{sac} is the mean value of the timer; the activation $a_k(t)$ of the fixated word k (i.e., the word in the fovea) at time t is the key variable that modulates the transition rate of the timer. Using numerical simulations of the model, it can be shown that for $h > 0$, foveal inhibition can produce a modulation of the fixation duration that is in good agreement with experimental data (Engbert et al., 2002, 2005).

2.2.3 Character-Based Visual Processing

Word recognition starts with visual processing of letters, which is done in parallel for all the letters of a given word. We define the spatial region where word activations can be influenced in the model as the *processing span*. Within this region, parallel processing is limited by the fact that processing rate depends on the letter's *eccentricity* (i.e., the distance of the letter position from the position of the current fixation). Mathematically, we define an inverted parabolic processing span from the fovea to position $-\delta_L$ on the left and to position $+\delta_R$ on the right of fixation, i.e.,

$$\lambda(\epsilon) = \lambda_0 \cdot \begin{cases} 0, & \text{for } \epsilon < -\delta_L \\ 1 - \epsilon^2/\delta_L^2, & \text{for } -\delta_L \leq \epsilon < 0 \\ 1 - \epsilon^2/\delta_R^2, & \text{for } 0 \leq \epsilon \leq \delta_R \\ 0, & \text{for } \delta_R < \epsilon \end{cases}, \quad (2.10)$$

where λ_0 is a constant given as

$$\lambda_0 = \frac{3}{2} \cdot \frac{1}{\delta_L + \delta_R}, \quad (2.11)$$

which is necessary to normalize the total processing rate, i.e., $\int_{-\infty}^{+\infty} \lambda(\epsilon) d\epsilon = 1$.

Experimentally, a strong asymmetry of the *perceptual span* with an extension of 4 to 5 letters to the left of the fixation position and up to 15 letters to the right was found (Rayner et al., 1980). Therefore, parameters δ_L and δ_R should be estimated separately from experimental data. In the following, we estimate $\delta_0 \equiv \delta_L = \delta_R$ for simplicity.

2.2.4 Word-Based Processing Rate

Because of the assumption of a processing span, Equation (2.10), processing rates for letters depend on spatial eccentricities. Letter j of word i is processed with rate $\lambda(\epsilon_{ij})$, if it is located at a spatial position with eccentricity $\epsilon_{ij}(t)$ relative to gaze position at time t . This letter-based processing rate must be related to the effective word-based processing rate $\Lambda_i(t)$ of word i at time t .

Because of parallel processing of the letters of a given word, each letter contributes to word recognition. In the case of long words, some letters will have large eccentricities, so that their processing rate will be small (or zero) according to Equation (2.10). To capture these opposing effects in a parametric model, we make the assumption that the word-based processing rate has the form

$$\Lambda_i(t) = M_i^{-\eta} \sum_{j=1}^{L_i} \lambda(\epsilon_{ij}(t)) , \quad (2.12)$$

where M_i is the word length (i.e., number of letters) of word i and η is the word length exponent, with $0 < \eta < 1$. For $\eta = 0$, long words will have a processing advantage. For $\eta = 1$, word processing rate is the arithmetic mean of the letter-based processing rates (mean over all letters of a given word); therefore, we will observe a disadvantage for long words in the case $\eta = 1$. We expect a numerical value for η about 0.5.

With the assumptions on spatial aspects of letter- and word-based processing rates, the temporal aspects of word processing need to be specified. As discussed for the motivation of the SWIFT model, a time-dependent activation field will provide probabilistic control of saccadic eye movements. Word-based activations $a_i(t)$ for the words of a given sentence are increasing during the initial stage of processing called *lexical processing*. After reaching the maximum of activation D_i for word i , the activation starts to decrease (*post-lexical processing*). The maximum of activation is interpreted as processing difficulty, which is a logarithmic function of word frequency Ω_i , i.e.,

$$D_i = \alpha \left(1 - \beta \frac{\log \Omega_i}{\log \Omega^{\max}} \right) , \quad (2.13)$$

where Ω^{\max} is the highest word frequency in a given language and parameter β determines the strength of the word frequency effect.

For word processing, we assume that current activation for each word $i = 1, 2, 3, \dots, N_w$ is related to the discrete state n_{4+i} of word processing (Table 2.1), given by

$$a_i(t) = D_i \frac{n_{4+i}}{N_a}, \quad (2.14)$$

where D_i is the word's processing difficulty, Equation (2.13).

Global decay of activation. Maintaining words in working memory during reading cannot be done without loss. Since word activations $\{a_n(t)\}$ represent the state-of-processing, we introduce a global decay of activation. If the processing rate of a word is smaller than the constant ω , then we assume a decay of activation with rate ω .

Processing during saccades. During saccadic eye movements, lexical processing is paused because of *saccadic suppression* (Matin, 1974). In the SWIFT model, lexical processing is paused during saccades in the lexical processing stage (increasing activation), while post-lexical processing (decreasing activation) continues during saccades.

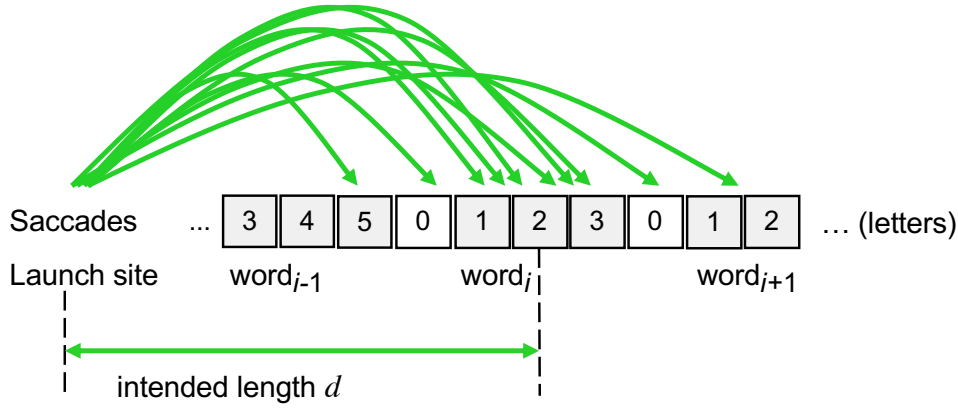
2.2.5 Oculomotor Assumptions

Our assumption of two-stage saccade programming are motivated by the experimental findings of the double-step paradigm (Becker & Jürgens, 1979). A saccade program starts with a labile stage; during this stage, the saccadic gaze center is forced to prepare the next saccade (Findlay & Walker, 1999), however, a new decision to start a labile saccade program during an ongoing labile stage leads to cancelation and replacement of the earlier saccade program. After the transition to the non-labile stage, the saccade can no longer be canceled or modified.

Oculomotor errors make an important contribution to eye-movement control during reading. In 1988, based on the analysis of initial fixation positions within words, McConkie and coworkers suggested that a considerable fraction of saccades landed on different words than the intended target words (McConkie et al., 1988). Using an iterative oculomotor modeling approach, Engbert and Nuthmann (2008) showed that about 10% to 20% of the saccades during natural text reading are mislocated on an unintended word.

McConkie et al. (1988) showed that saccadic errors can be decomposed into a random (approximately Gaussian) error component and a systematic shift (called saccadic range error). The critical variable that determines the size of both random and systematic error components turned out to be the intended saccade length (distance d from the launch site of the saccade to the center of the target word). Therefore, saccades targeting a word center at $x = 0$ will be normally distributed with

$$x \sim \mathcal{N} \left(\epsilon_{\text{sre}}, \sigma_{\text{sre}}^2 \right), \quad (2.15)$$

Figure 2.3*Oculomotor Error Model*

Note. Saccades start at a launch site and aim at the word center of the selected target word i . Oculomotor errors are normally distributed, which can lead to misplaced fixations on word $i - 1$ (undershoot error) or word $i + 1$ (overshoot error). Both the standard deviation σ_{sre} and the mean shift ϵ_{sre} from the intended word's center depend on the intended saccade length d .

where both parameters depend linearly on the intended saccade length d , i.e.,

$$\epsilon_{\text{sre}} = r_1 - r_2 d \quad (2.16)$$

$$\sigma_{\text{sre}} = s_1 + s_2 d, \quad (2.17)$$

where d is the physical distance between the launch site of the saccade and the word center of the target word, measured in units of character spaces. The oculomotor parameters r_1 , r_2 , s_1 , and s_2 will vary depending on the type of saccade (e.g., refixation or skipping), which is discussed in earlier papers (Engbert et al., 2005; Krügel & Engbert, 2010). We would like to remark that McConkie et al.'s descriptive model of saccadic errors could be replaced by a process-oriented Bayesian model (Engbert & Krügel, 2010; Krügel & Engbert, 2014) in perspective.

2.2.6 Modulation of the Duration of the Labile Stage

An important problem is the observation of a reduced average fixation duration for refixations. As a solution, we assume that the duration of the labile stage of saccade programming is reduced by factor R ($0 < R \leq 1$), if the fixation is a refixation.

Closely related is the phenomenon of mislocated fixations (Engbert & Nuthmann, 2008). If the realized fixation position (the saccadic landing position) strongly deviates from the word center, so that the landing position will fall onto the neighboring word, then a mislocated fixation will occur. In this case, the duration of the next saccade program will be

reduced by factor M ($0 < M \leq 1$). Such a mechanism is a possible explanation of the inverted optimal viewing-position effect (Nuthmann et al., 2005; Vitu et al., 2001) of fixation durations that indicates reduced average fixation duration at word edges compared to the word center. In the SWIFT version used here, the probability of misplaced fixation is given as $p_{\text{mis}} = 0.9 \cdot (2\delta/M)^4$, where δ is the fixation error (distance from word center) and M is the length of the fixated word.

2.2.7 Numerical Simulation and Model Parameters

For numerical simulations of single trajectories of the SWIFT model, the *minimal process method* by Gillespie (1976), an exact and efficient numerical algorithm, can be derived from the master equation, Equation (2.8). If the model is in state n at time $t_0 = 0$ with certainty, all other states will have zero probability, i.e., $p(n', t|n)$ for $n' \neq n$. Therefore, the master equation, Equation (2.8), reduces to

$$\frac{\partial}{\partial t} p(n, t | n) = - \sum_{n'} W_{n'n} p(n, t | n) = -W_n p(n, t | n) , \quad (2.18)$$

where $W_n = \sum_{n'} W_{n'n}$ is the total transition probability from state n . From Equation (2.18), we obtain an exponentially distributed waiting time for the next transition from state n to an adjoined state $n' \neq n$. Following Gillespie (1976), a two-step algorithm can be derived: In step 1, an exponentially-distributed random number is generated; in step 2, a transition (Table 2.1) is chosen according to relative transition probabilities, $W_{n'n}/W_n$ with $n' \neq n$. This algorithm is numerically efficient, since it restricts computations to the transitions when simulating the system's trajectory.

For the simulations in this paper we used a restricted version of the SWIFT model to reduce the number of free parameters to 11 (see Table 2.2, cf. Engbert et al., 2005). Moreover, we fixed seven of these parameters to estimate four free parameters in the simulation examples. Future simulation studies will be carried out with more free parameters (see Section 2.5). The number of possible random-walk states varies between subprocesses; based on earlier simulations (Schad & Engbert, 2012), we used the following numbers: $N_t = 15$ (saccade timer), $N_l = 12$ (labile saccade stage), $N_n = 10$ (non-labile saccade stage), $N_x = 20$ (saccade execution), and $N_a = 30$ (word activations).

2.3 Likelihood Function for the SWIFT Model

For the parameter estimation procedure discussed in the introduction, we aim at a framework that computes the likelihood of a series of experimentally observed fixations incrementally, Equation (2.3). For fixation f_i , we need to compute the likelihood function

Table 2.2*Model Parameters of the SWIFT Model*

| Parameter | Symbol | Typical Value | Reference |
|-------------------------------|---------------------------|---------------|-----------------|
| Lexical difficulty: Intercept | α | 50 | Equation (2.13) |
| Lexical difficulty: Slope | β | 0.75 | Equation (2.13) |
| Processing span | $\delta_0 = \delta_{L,R}$ | 8 | Equation (2.10) |
| Word-length exponent | η | 0.5 | Equation (2.12) |
| Saccade timer | t_{sac} | 250 ms | Table 2.1 |
| Foveal inhibition | h | 0.6 | Equation (2.9) |
| Labile saccade program | τ_l | 120 ms | Table 2.1 |
| Non-labile program | τ_n | 80 ms | Table 2.1 |
| Saccade execution | τ_x | 20 ms | Table 2.1 |
| Refixation factor | R | 0.9 | Section 2.2.6 |
| Mislocated fixation | M | 1.5 | Section 2.2.6 |

Note. Numerical values are chosen in agreement with earlier publications (see text).

$P_M(f_i | f_1, \dots, f_{i-1}, \boldsymbol{\theta}, \boldsymbol{\xi})$ given the previous fixations f_1, f_2, \dots, f_{i-1} , the model parameters $\boldsymbol{\theta}$, and the internal states $\boldsymbol{\xi}$ of model M , which we not addressed in Equation (2.3). In SWIFT each fixation event $f_i = (x_i, T_i, s_i)$ is defined by a fixation position x_i given by the fixated word v_i and the fixated letter l_i within the word, the fixation duration T_i , and the saccade duration s_i . The likelihood for fixation f_i is composed of a spatial contribution and a temporal contribution. At time t , fixation i starts on letter l_i of word v_i , which is predicted by the SWIFT model with a probability determined by word activations and oculomotor assumptions. After fixation i started, the model can make another prediction for the fixation duration T_i of fixation i . Next, the likelihood for fixation i can be decomposed into the spatial and temporal contributions, i.e.,

$$P_M(v_i, l_i, T_i | F_{i-1}, \boldsymbol{\theta}, \boldsymbol{\xi}) = P_{\text{temp}}(T_i | v_i, l_i, F_{i-1}, \boldsymbol{\theta}, \boldsymbol{\xi}) \cdot P_{\text{spat}}(v_i, l_i | F_{i-1}, \boldsymbol{\theta}, \boldsymbol{\xi}), \quad (2.19)$$

where we introduced $F_{i-1} \equiv \{f_1, f_2, \dots, f_{i-1}\}$ to simplify the notation.

For the *spatial likelihood* P_{spat} , the dynamically evolving word activations in SWIFT determine the time-dependent probability for selecting a particular word as the next target word. Additionally, the target-selection probability is modified by oculomotor noise. Due to dynamical dependencies, we compute the likelihood of an experimentally realized fixation position based on the previous fixations. However, the internal states $\boldsymbol{\xi}$ are given by the stochastic dynamics and are, therefore, unknown. In principle, we could integrate over many possible realizations of the internal states $\boldsymbol{\xi}$, which is, however, time-consuming for the numerical computations. Therefore, we compute P_{spat} for one realization of the internal states $\boldsymbol{\xi}$, which results in fluctuating numerical values for P_{spat} . Thus, instead of integrating out the internal degrees of freedom $\boldsymbol{\xi}$, we used a pseudo-marginal likelihood (Andrieu & Roberts,

2009) and eliminated the dependence on ξ for the spatial likelihood in Equation (2.19).

For the *temporal likelihood* P_{temp} , SWIFT computations start with a realized fixation position on letter l_i of word v_i , however, with internal states ξ . Given this fixation position, the distribution of fixation durations can be predicted by the model. The generated estimate of the likelihood of the experimentally realized fixation duration is approximated by averaging over many realizations of the internal states ξ (e.g., the internal states of the various saccade programming stages). As a result, both P_{temp} and P_{spat} are random variables, which will be discussed in detail in the next two sections.

2.3.1 Spatial Likelihood

In SWIFT, saccadic gaze shifts are generated in two steps: First, a target word is determined in a probabilistic selection process based on relative word activations. Second, after a short delay generated by saccade programming, the saccade is executed with oculomotor errors influenced by the saccadic landing position distribution. These oculomotor errors induce stochastic variability in the within-word fixation position and can also induce mislocated fixations (Engbert & Nuthmann, 2008; Nuthmann et al., 2005), where the realized fixation position is placed on a different word than the selected target.

The combination of activation-based saccadic selection and oculomotor errors generates a non-zero probability for all fixation positions (Fig. 2.3). The target selection probability $\pi(m, t - \tau_n - \tau_x)$ (see. Equation 2.6) is the probability of selecting word m as a saccade target for a fixation starting at time t . It is important to note that target selection occurs at the time of transition from the labile to the non-labile saccade program, so that the probability $\pi(\cdot)$ for selecting the next target word has to be evaluated with an average time delay $\tau_n + \tau_x$. According to our oculomotor assumptions, the saccadic error generates a probability $q(v, l | m, x_{\text{gaze}})$ of fixating word v at letter l given that word m is the selected target word and x_{gaze} is the previous gaze position (or saccade launch site). Thus, the spatial likelihood of an observed saccade starting at time t_i towards letter position l of word v is therefore given by

$$P_{\text{spat}}(v, l | F_{i-1}, \theta) = \sum_{m=1}^{N_w} \pi(m, t_i - \tau_n - \tau_x) q(v, l | m, x_{\text{gaze}}), \quad (2.20)$$

where we dropped the conditional arguments to simplify the notation. Moreover, the time-dependency is now written explicitly, since t_i for the computation of the spatial likelihood of fixation i is given by the sum of fixation durations and saccade durations of the previous fixations in the sequence, $t_i = \sum_{l=1}^{i-1} T_l + s_l$.

The oculomotor system generates systematic and random errors that introduce deviations between the target word's center and the realized fixation position. In SWIFT, we adopt McConkie et al.'s (1988) range-error framework by assuming a Gaussian distribution that is

shifted with respect to the target word's center. Thus, the probability of landing at letter l of word v , given a target word m , is given by

$$q(v, l | m, x_{\text{gaze}}) = \frac{1}{\sqrt{2\pi}\sigma_{\text{sre}}} \exp\left(-\frac{[(v_m + \epsilon_{\text{sre}}) - x_{v,l}]^2}{2\sigma_{\text{sre}}^2}\right) \cdot \Delta x, \quad (2.21)$$

where v_m is the spatial position of the target word's center, $x_{v,l}$ is the spatial position of the fixated letter l of word v , and $\Delta x = 1$ is the unit width of a letter. The oculomotor parameters $\epsilon_{\text{sre}}(d)$ and $\sigma_{\text{sre}}(d)$ of the range-error model specify systematic shift (saccadic range error) and standard deviation of the random error (oculomotor noise), respectively, Equations (2.16, 2.17); the intended saccade length $d = \|v_m - x_{\text{gaze}}\|$ is given as the distance between the target word's center v_m and the fixation position before the saccade x_{gaze} .

2.3.2 Temporal Likelihood

Because of two-stage saccade programming and due to the fact that fixations are bounded by two saccades in time, SWIFT's fixation durations are given as linear combinations of realizations of random variables. For the saccade timer and saccade programming stages, resulting durations are gamma-distributed random variables, which are generated by continuous-time discrete-state random walks according to the master equation, Equation (2.8).

The saccade timer controls the initiation of the saccade programming cascade with consecutive labile and non-labile stages and a saccade execution stage. The time interval between the end point of the previous and the beginning of the next saccade execution is defined as the experimentally observed fixation duration. However, the saccade timer is continuously inhibited by word activations. As a consequence, the mean waiting times (the inverse of the transition probabilities) of the elementary steps of the saccade timer's random walk will be time-dependent. Additionally, the mean durations of the labile stages of saccade programming depend on the type of fixation (i.e., whether it is a refixation, a mislocated fixation, or neither of these). Finally, if the saccade timer produces a short interval, then saccade cancellation will be likely, which results in a higher mean value of the predicted fixation duration.

Since each fixation duration is bounded by two saccades (i.e., the i th fixation duration lies between $(i-1)$ th saccade offset and i th saccade onset), each observed fixation duration T_i is compared to the simulated realization \tilde{T}_i that is given as the sum of the following terms (see Fig. 2.4a),

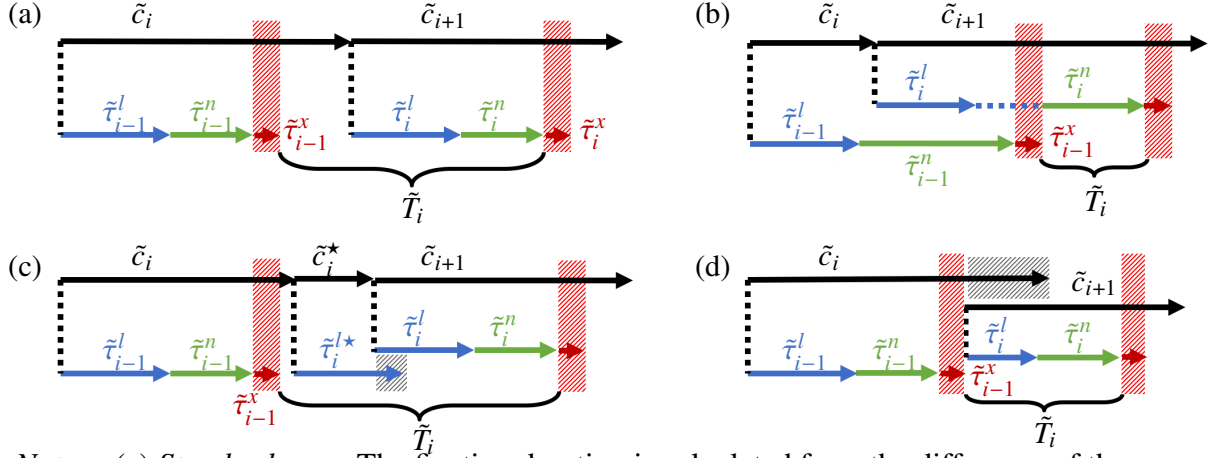
$$\tilde{T}_i = \tilde{c}_i + \tilde{\tau}_i^l + \tilde{\tau}_i^n - \tilde{\tau}_{i-1}^l - \tilde{\tau}_{i-1}^n - \tilde{\tau}_{i-1}^x, \quad (2.22)$$

where \tilde{c}_i is the realized saccade timer duration, $\tilde{\tau}_i^l$ and $\tilde{\tau}_i^n$ are realized durations of the labile and non-labile saccade programming stages respectively, and $\tilde{\tau}_i^x$ is the realized saccade duration.

Our strategy for the computation of the temporal likelihood of the i th fixation duration

Figure 2.4

Schematic Illustrations of the Generation of Fixation Durations for Different Types of Fixations in SWIFT



Note. (a) *Standard case*: The fixation duration is calculated from the difference of the sum of the saccade timer c_i , the labile and non-labile saccade latencies $\tilde{\tau}_i^l$ and $\tilde{\tau}_i^n$, respectively, and the sum of saccade latencies $\tilde{\tau}_{i-1}^l$, $\tilde{\tau}_{i-1}^n$ and $\tilde{\tau}_{i-1}^x$. (b) *Labile pausing*: If a saccade program reached the non-labile stage it cannot be aborted anymore. A newly started labile programming stage will transition to its non-labile stage only after the current saccade program is terminated at saccade offset. (c) *Saccade cancellation*: If the saccade timer finishes earlier than the concurrent labile saccade program, the ongoing labile saccade program is canceled—consequently, both the labile program and the saccade timer are restarted. The realized duration of the premature saccade timer \tilde{c}_i^* is added to the new realization \tilde{c}_i . (d) *Refixation and Mislocated Fixation*: If the current fixation is either a refixation or considered to be a mislocated fixation, the saccade timer realization \tilde{c}_i is reset immediately at fixation onset and a new labile saccade program is initiated. The fixation duration is then given as the sum of the current labile and non-labile durations $\tilde{\tau}_i^l$ and $\tilde{\tau}_i^n$ respectively.

T_i is to simulate many realizations of \tilde{T}_i from Equation (2.22) to numerically approximate the theoretical distribution of fixation durations with kernel density estimation¹². In the context of Bayesian analysis, this approach is termed probability density approximation (PDA) (Holmes, 2015; Palestro et al., 2018; Turner & Sederberg, 2013), which falls into the broad class of likelihood-free procedures in approximate Bayesian computation (ABC; see Sisson & Fan, 2011, for a review).

Since all of the terms in Equation (2.22) are random realizations of stochastic variables, the order of terminations of the subprocesses shown in Fig. 2.4(a) can be violated. In the following, we discuss all possible cases:

1. *Labile pausing* happens if the labile saccade program terminates during an ongoing non-labile saccade program. Since we assume that there cannot be more than one

¹²While it is possible to derive an iterative algorithm for the distribution of linear combinations of gamma-distributed random numbers (S. V. Amari & Misra, 1997; Coelho, 1998), it turned out that these solutions are numerically unstable.

non-labile saccade program active at a time, the current labile program is paused immediately before termination, thus its duration is extended until the current non-labile program and saccade execution finish (Fig. 2.4b). Formally, this situation is encountered if $\tilde{c}_i + \tilde{\tau}_i^l < \tilde{\tau}_{i-1}^l + \tilde{\tau}_{i-1}^n + \tilde{\tau}_{i-1}^x$. In this case, the interval $\tilde{\tau}_i^l$ is increased and the calculation of \tilde{T}_i is simplified to the duration of the non-labile saccade program, i.e.,

$$\tilde{T}_i = \tilde{\tau}_i^n . \quad (2.23)$$

Since the duration of the labile program is extended, however, there will be increased probability for the saccade timer to terminate during the ongoing labile program, while will cause saccade cancelation.

2. *Saccade cancelation* occurs if the main saccade timer realization \tilde{c}_{i+1} terminates during an ongoing labile saccade programming stage $\tilde{\tau}_i^l$, i.e., $\tilde{c}_i^* < \tilde{\tau}_i^{l*}$, which is illustrated in Figure 2.4c. In this case the labile saccade program is canceled and replaced with the new labile saccade program initiated by restarting of the saccade timer. As a result, the duration of the timer \tilde{c}_i in Equation (2.22) is replaced by the sum $\tilde{c}_i + \tilde{c}_i^*$. Therefore, the corresponding distribution T_i for saccade cancelation is given by

$$\tilde{T}_i = \tilde{c}_i + \tilde{c}_i^* + \tilde{\tau}_i^l + \tilde{\tau}_i^n - \tilde{\tau}_{i-1}^l - \tilde{\tau}_{i-1}^n - \tilde{\tau}_{i-1}^x , \quad \text{if } \tilde{c}_i^* < \tilde{\tau}_i^{l*} . \quad (2.24)$$

In principle, saccade cancelation can happen repeatedly within the same fixation, depending on the choice of parameters.

3. *Refixations and mislocated fixations* represent another special case, where a new saccade program is triggered immediately after the fixation onset (Fig. 2.4d). In both cases the saccade timer realization \tilde{c}_i is reset and a new labile saccade program is initiated. The mean duration of the new labile stage is modified by coefficients $f^r = 1/R$ and $f^m = 1/M$ for refixations and mislocated fixation, resp. (see 2.2.5). As a result, the observed fixation duration is given as

$$\tilde{T}_i = f^{r,m} \tilde{\tau}_i^l + \tilde{\tau}_i^n . \quad (2.25)$$

The SWIFT model includes inhibition of fixation durations by word activation; in its simplest form, the activation of the fixated (foveal) word inhibits the fixation duration by decreasing the transition rates of the saccade timer (Equation 2.9). Because of the complicated time-course of the activation field (i.e., sudden changes of activation evolution due to saccades), stochastic simulations are necessary to estimate the distribution of \tilde{T}_i .

To compute the likelihood $L_{\text{temp}}(T_i)$ of an observed fixation duration T_i we first simulate the activation evolution for words in the perceptual span from time $t = 0$ until the point in time

that corresponds to the end of fixation i . We start simulating the stochastic contributions by initially going backwards from the time of fixation onset by sampling the saccade latencies $\tilde{\tau}_{i-1}^x$, $\tilde{\tau}_{i-1}^n$, and $\tilde{\tau}_{i-1}^l$ to determine the onset of the saccade timer c_i . The previously sampled activations provide information for the simulation of the saccade timer with inhibition by foveal word activations, similar to the generative process. If $\tilde{c}_i < \tilde{\tau}_{i-1}^l$, both realizations are discarded and sampled again with the same procedure (we are not interested in saccade cancelation events which do not affect the fixation duration under consideration). The offset of \tilde{c}_i demarks the onset of \tilde{c}_{i+1} and, following the rules of the previously discussed order violations, we can easily simulate the timer cascade until fixation offset and hence obtain a sample from the distribution of fixation durations as provided by the SWIFT framework with respect to the history of the fixation sequence.

Once $N = 300$ fixation durations are sampled, the distribution of T_i^{exp} is approximated via KDE. Increasing the number of samples increases the accuracy of the approximation but is costly in terms of computation time. For the density estimation we use the Epanechnikov kernel (Epanechnikov, 1969) with a bandwidth setting according to Scott's rule (2015). The Epanechnikov kernel is computationally efficient as it only integrates samples within its limited interval given by the bandwidth. However this can result in situations where no data point is covered by the kernel. To prevent estimates with zero probability, the bandwidth of the kernel was adjusted to the 1.1-fold of the distance between T_i^{exp} and the nearest sample of \tilde{T}_i , so that at least one sample will lie within the kernel.

2.3.3 Evaluation of the Log-Likelihood Using Single-Parameter Variations

A simple test of the likelihood function and its inherent stochastic contributions can be done by repeatedly evaluating the likelihood of a simulated dataset for which the parameters are known and keeping all parameters but one at their respective true values (i.e., the values used in generating the data). Systematically varying the parameter under consideration reveals its impact on the likelihood. Since the likelihood function is composed of two terms from spatial and temporal contributions (Equation 2.19), separating both components can also prove insightful with regard to the strength and direction of the parameter's influence.

To investigate the properties of the likelihood function for a relevant subset of parameters, we simulated 1624 fixations on 114 sentences (Figure 2.5) from the sentence corpus of Risse and Seelig (2019). The examined parameters are given in Table 2.3, with the remaining parameters set according to Table 2.2. The likelihood was then evaluated for 1000 different, evenly spaced values within the given interval (Table 2.3) separately for each parameter. Since all other parameters were fixed at their true values, any systematic change in the resulting log-likelihood can only be attributed to the parameter under consideration.

Table 2.3*Parameters of the SWIFT Model Considered in Bayesian Estimation*

| Parameter | Symbol | Range | True value |
|----------------------|------------------|--------------|------------|
| Saccadic timer | t_{sac} | 150...350 ms | 260 ms |
| Refixation factor | R | 0.2...1.8 | 0.9 |
| Processing span | δ_0 | 4...15 | 8.5 |
| Word length exponent | η | 0...1 | 0.4 |

Note. True values apply to the synthetic data generated for verification of the likelihood function.

Figure 2.5a indicates that the saccade timer t_{sac} influences the temporal likelihood, while there is no influence on the spatial likelihood. A similar behavior is observed for the refixation factor R (Figure 2.5b). In both cases, there is a clear maximum in the likelihood profile at the true parameter values, $t_{\text{sac}} = 260$ ms and $R = 0.9$, resp. A different dependence can be seen for the processing span δ_0 , which clearly influences the spatial likelihood (maximum at the true value $\delta_0 = 8.5$), but exerts only a minimal influence on the temporal likelihood (Figure 2.5c). For the word-length exponent η , there is an influence on both spatial and temporal likelihoods (Fig. 2.5d), with a maximum for both likelihood profiles at the true parameter value $\eta = 0.4$.

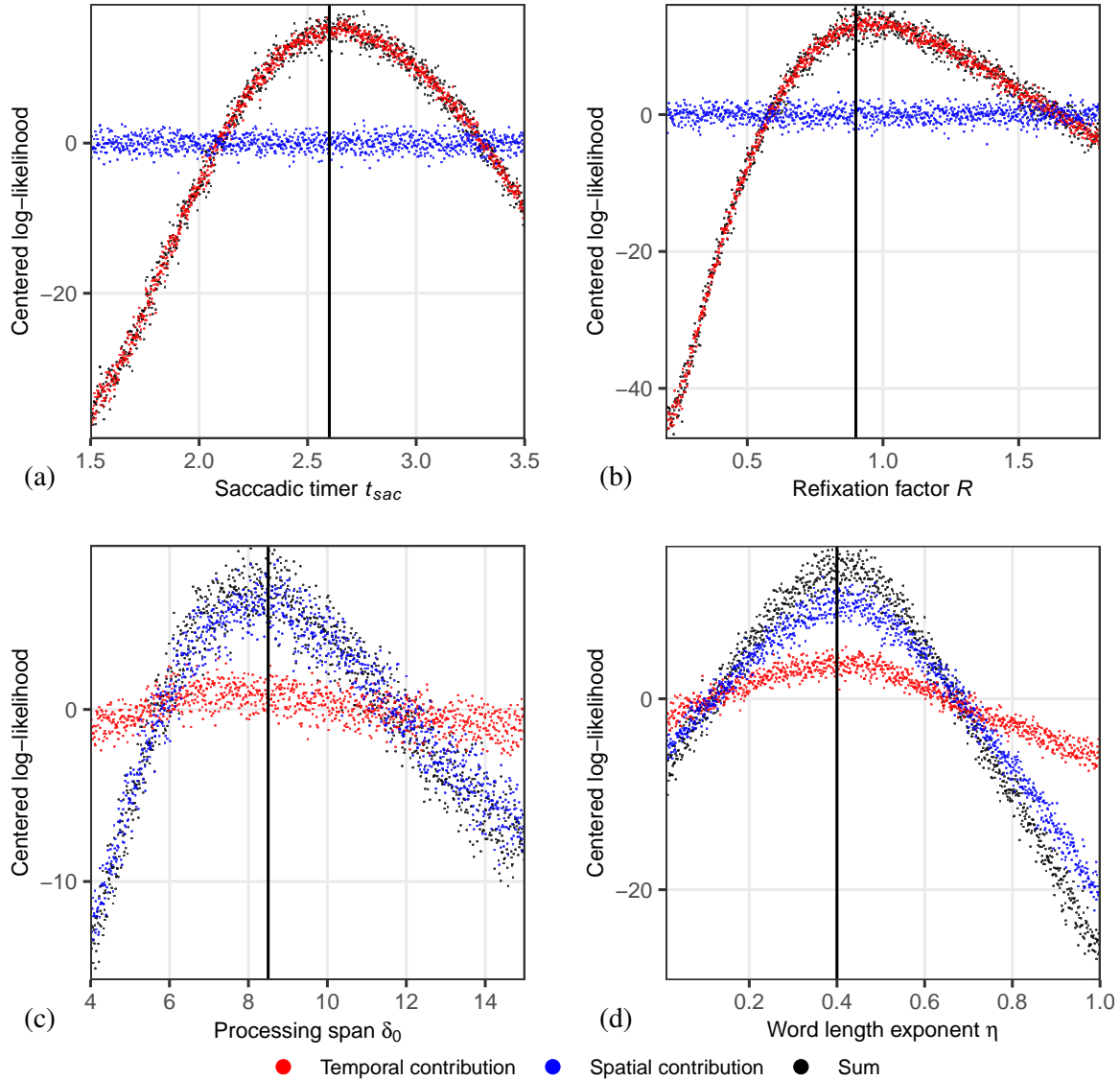
Thus, our numerical implementation of the likelihood function indicates clear maxima at the true parameter values for simulated data, while stochastic fluctuations due to the approximative account for internal degrees of freedom ξ are small. In the next section, we will apply an adaptive MCMC framework for Bayesian parameter estimation using simulated and real (experimental) data.

2.4 Likelihood-Based Parameter Inference Using MCMC

With the implementation of the numerical computation of the likelihood function for the SWIFT model from the previous section, we developed the critical step for adopting the Bayesian framework for parameter inference. We will discuss the Markov Chain Monte Carlo approach used for inference, discuss the efficient implementation on a digital computer, present results for parameter recovery from simulated data with known parameters, and, finally, estimate parameters for experimental data.

2.4.1 Markov Chain Monte Carlo Simulation for the SWIFT Model

As described in Section 2.1.2, the computability of the likelihood $L_M(\theta|F)$, Equation (2.3), for a given set of parameters θ and a given fixation sequence F is critical for maximum-likelihood and Bayesian inference. For the numerical procedures of Markov

Figure 2.5*Temporal, Spatial and Total Log-Likelihood Profiles*

Note. Temporal contributions in red and spatial contributions in blue to the total log-likelihoods in black. Likelihood was evaluated for a simulated dataset of 1624 fixations on 114 sentences from the Risse and Seelig (2019) corpus. Single parameters were varied within an interval around the respective true parameter value used in creating the data. The log-likelihoods were centered around their respective mean value.

Chain Monte Carlo type, we use a variant of the Metropolis Hastings (MH) algorithm (Hastings, 1970). In the random-walk MH algorithm, a random walk in the parameter space is generated, where the probability of the random-walk steps depends on the ratio of the likelihoods associated with the random walk's current and proposed new positions.

Starting from an arbitrary initial point X_0 in parameter space, every move is determined by two steps:

1. A proposal Y_n is generated by a random-walk step from position X_{n-1} ,

$$Y_n = X_{n-1} + SU_n, \quad (2.26)$$

where $U_n \sim \mathcal{N}(0, \sigma)$. Both the shape matrix S and the width σ of the proposal distribution must be chosen beforehand and kept constant during a run of the algorithm.

2. The proposal is then accepted with the probability

$$\alpha_n := \alpha(X_{n-1}, Y_n) := \min \left\{ 1, \frac{\pi(Y_n)}{\pi(X_{n-1})} \right\}, \quad (2.27)$$

in which case $X_n = Y_n$, i.e. the walker moves to the proposed position. If the proposal is rejected, then the random walk remains at the current position $X_n = X_{n-1}$.

By recursively following these rules the chain of accepted samples of the algorithm asymptotically converges to the true distribution of π . However, the speed of convergence greatly depends on an optimal choice of both the shape matrix S and the width parameter σ of the proposal distribution. Poor choices lead to abundant rejections (i.e. the chain is stationary most of the time if S is chosen badly or σ is too large) or strong autocorrelations of the samples (i.e., movements are very small if σ is chosen too small, even if S is optimal). Both parameters are however not known in advance and cannot be obtained due to analytical intractability of SWIFT model's likelihood function.

Therefore, we used the *Robust Adaptive Metropolis* (RAM) algorithm by Vihola (2012) which progressively captures the parameters' covariance structure shape and at the same time attains a predefined acceptance rate (see G. O. Roberts et al., 1997). The speed of the adaptation can also be specified parametrically. Although the RAM algorithm is a good strategy for parameter estimation, it is still computationally expensive, as exploration is naturally slow, if subsequent samples are dependent. Furthermore, it is necessary to use several independent chains with randomly dispersed initial values, each requiring a burn-in phase necessary for the sampler to progress to the vicinity of the stationary distribution.

An additional modification of the MCMC algorithm is necessary because of the stochastic pseudo-likelihood function of the SWIFT model. If, by chance, an exceptionally high log-likelihood value is obtained for a proposal, the acceptance rate for the subsequent proposal will be very low, which might stall the chain (Holmes, 2015). Therefore we re-evaluate $\pi(X_{n-1})$ for every iteration of the algorithm, which, however, doubles the computation time of the sampling.

To increase computational efficiency, we introduced parallel computation at two levels. First, while the likelihood of a fixation is dependent on all preceding fixations in the respective fixation sequence, likelihoods of whole fixation sequences can be computed independently from each other and added up later. This procedure enables computing the

log-likelihood for independent fixation sequences in F in parallel using a multi-core compute cluster. Second, different chains are independent of each other and can therefore be calculated in parallel as well.

2.4.2 Parameter Recovery Using Simulated Data

Before we demonstrate the application of the MCMC framework for the SWIFT model to experimental data, we investigate its performance for simulated data with known parameters. While we tested the likelihood function using single-parameter variation around the true value in Section 2.3.3, we now estimate all four selected parameters (Table 2.3) simultaneously using the MCMC procedure for the same dataset. We specified truncated normal distributions centered at parameter ranges (see Table 2.3). The standard deviation was set to one half of the estimation range in order to obtain an uninformative prior. We ran 5 independent chains with $N = 4,000$ iterations each and the default adaptation parameter value of $\gamma = 2/3$. The resulting marginal posterior distributions are given in Figure 2.6. The results suggest that the likelihood-based MCMC framework is very promising for parameter estimation based on data from single participants.

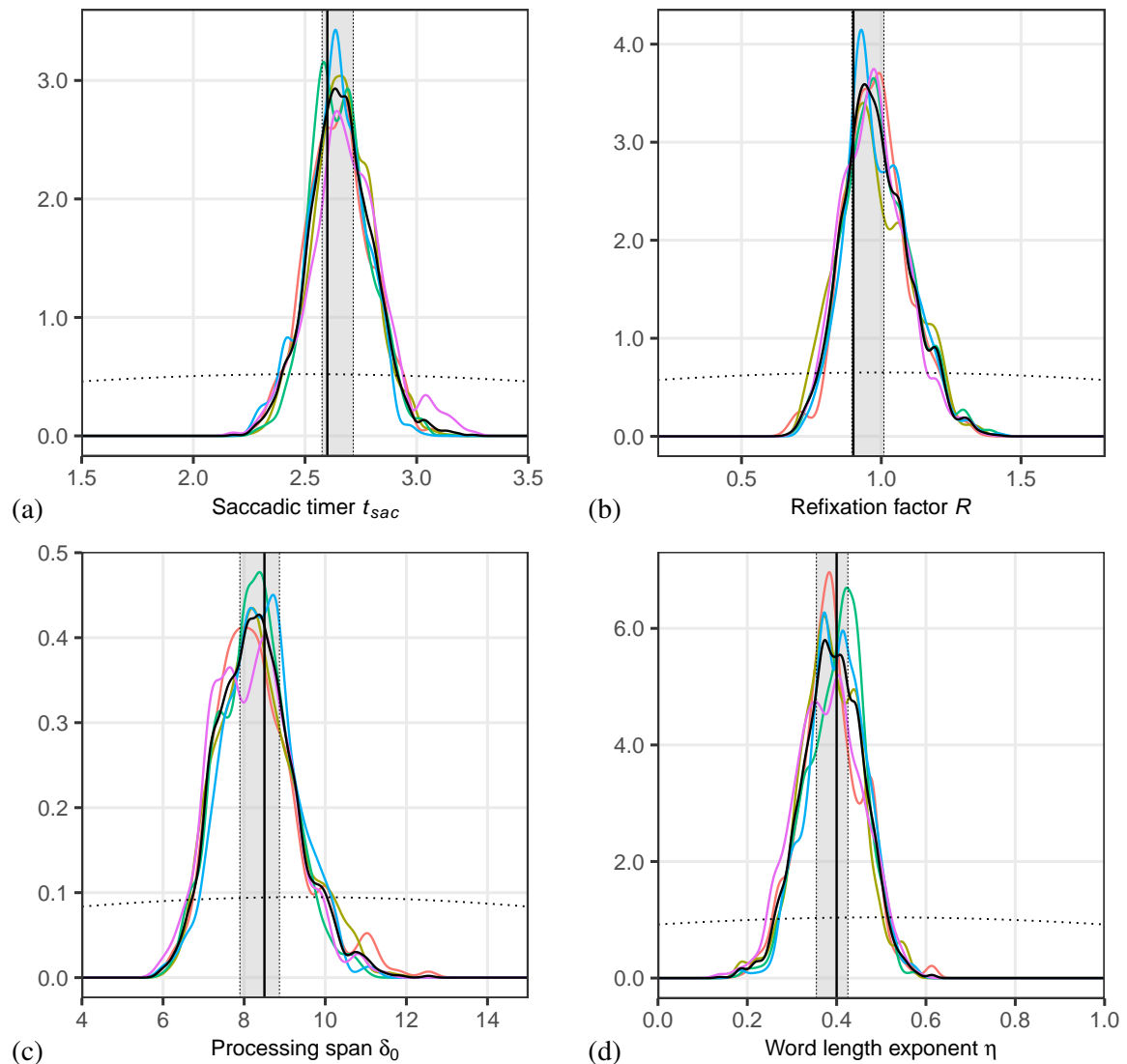
2.4.3 Estimation of Parameters Based on Experimental Data

In the next step, we estimated the same parameters for data from an eye tracking experiment. We used the control condition from a larger experimental study on parafoveal processing using the boundary paradigm (see Risse & Seelig, 2019, for a detailed description of the boundary paradigm). We ran 10 chains per participant, each with 4,000 iterations. We used the last 2,000 samples (50%) after the burn-in to estimate the posterior density. The resulting marginal posterior densities for a single participant are plotted in Figure 2.8. While there is an increased variance in the posterior densities for the estimation using experimental data compared to the simulated data (Fig. 2.6, we observe clear convergence of the independent chains to a common posterior estimate. Since there is qualitative agreement for the results on simulated and experimental data, the method seems promising to investigate interindividual differences via parameter estimation, which is discussed in the next section.

2.4.4 Interindividual Differences and Model Parameters

In this section we study interindividual differences in model parameters across 34 subjects that served as participants in the experiment by Risse and Seelig (2019). Figure 2.8 shows the posterior densities for all subjects, demonstrating considerable interindividual differences over the model parameters t_{sac} , R , and δ_0 , whereas estimates of η fall close to zero.

Figure 2.6
Exemplary Posterior Distributions

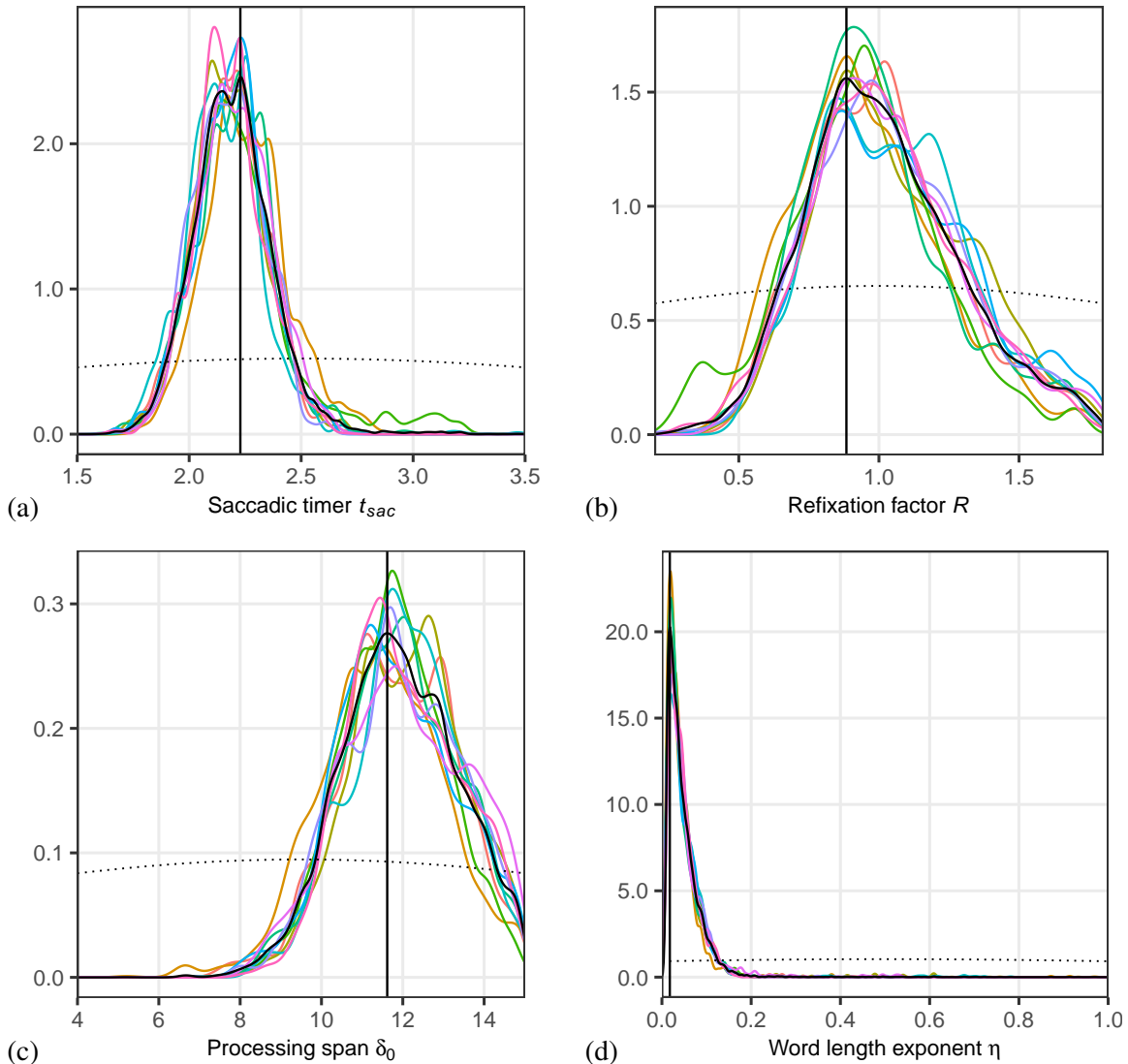


Note. Distributions are from five individual chains (different colors) for four parameters based on simulated data. The black vertical lines indicate the true parameter values. Grey areas indicate the 40% HPDI of all chains. The scale of the parameter range reflects the width of the prior (black, dotted).

A critical question is how much of the differences in reading behavior could be explained by the estimated differences in model parameters. Therefore, we used the maximum a posteriori (MAP) estimator (i.e. the mode) of the pooled chains for each subject as input parameters for the generative model and created a simulated data set that corresponds to the experimental data.

Fixation durations. For both the experimental and the artificial data, we calculated participant-wise averages in different measures of fixation durations. Specifically we com-

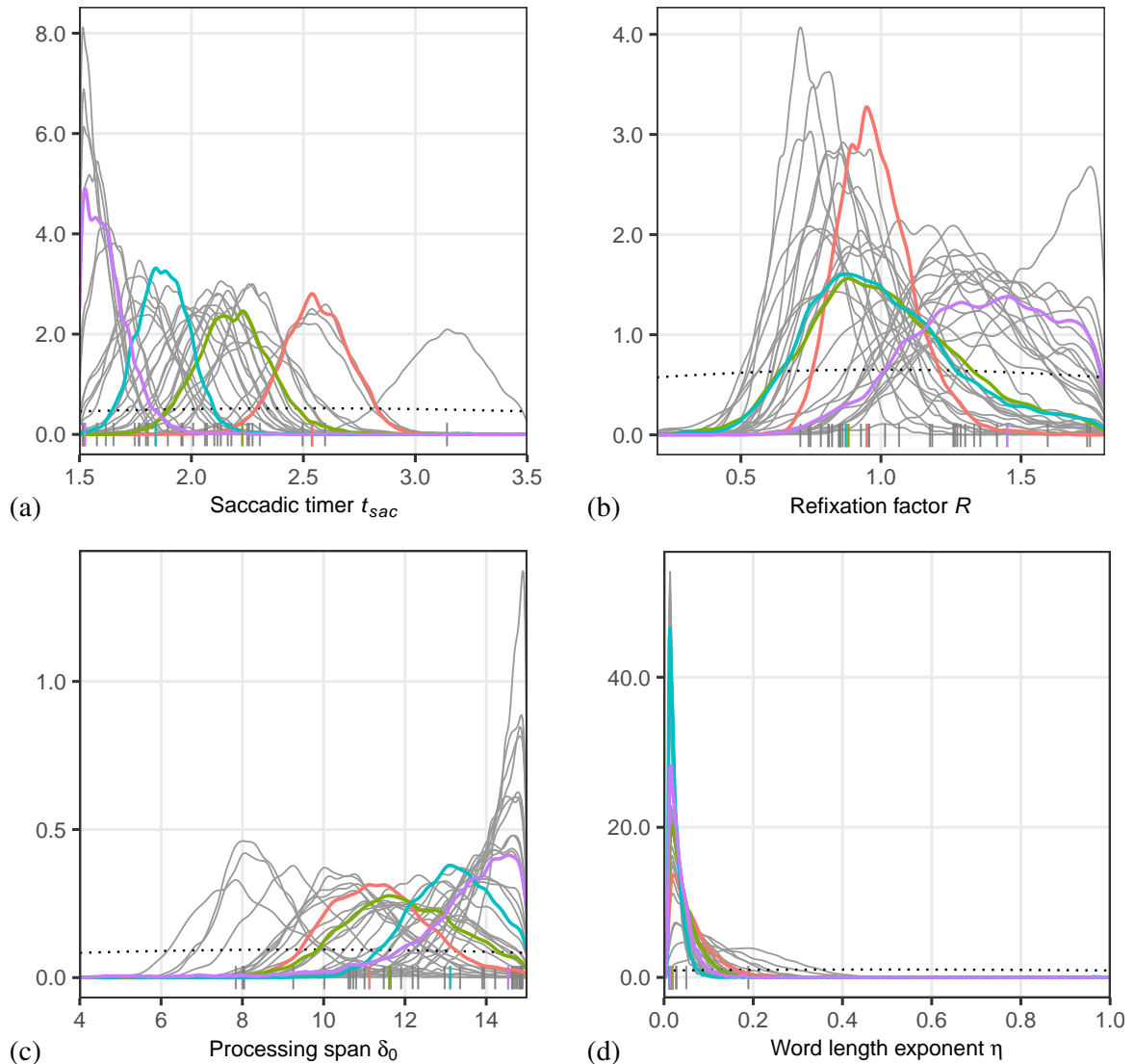
Figure 2.7
Example Posterior Densities for Single Participants



Note. Posterior densities for 10 independent chains (colored) for experimental data from a single participant. The MAP estimator for the pooled chains (black) of each respective parameter is indicated by the black vertical line. The prior is indicated by the black dotted line.

pared durations of single fixations (*SFD*; when the word was fixated only once in first-pass), first fixations (*FFD*; when the word was fixated once or more in first-pass), refixations (*RFD*; the second fixation on words, which were fixated more than once consecutively in first-pass), gaze durations (*GD*; the total time spent on a word in first-pass) and total viewing time (*TVT*; the total time spent on a word regardless of first, second or more passes). The results (Fig. 2.10a) indicate a remarkably good fit between the experimental data and model predictions for individual participants for *RFD* and *GD*. Mean *FFD* and *SFD* generated by the

Figure 2.8
Posterior Distributions of 34 Participants



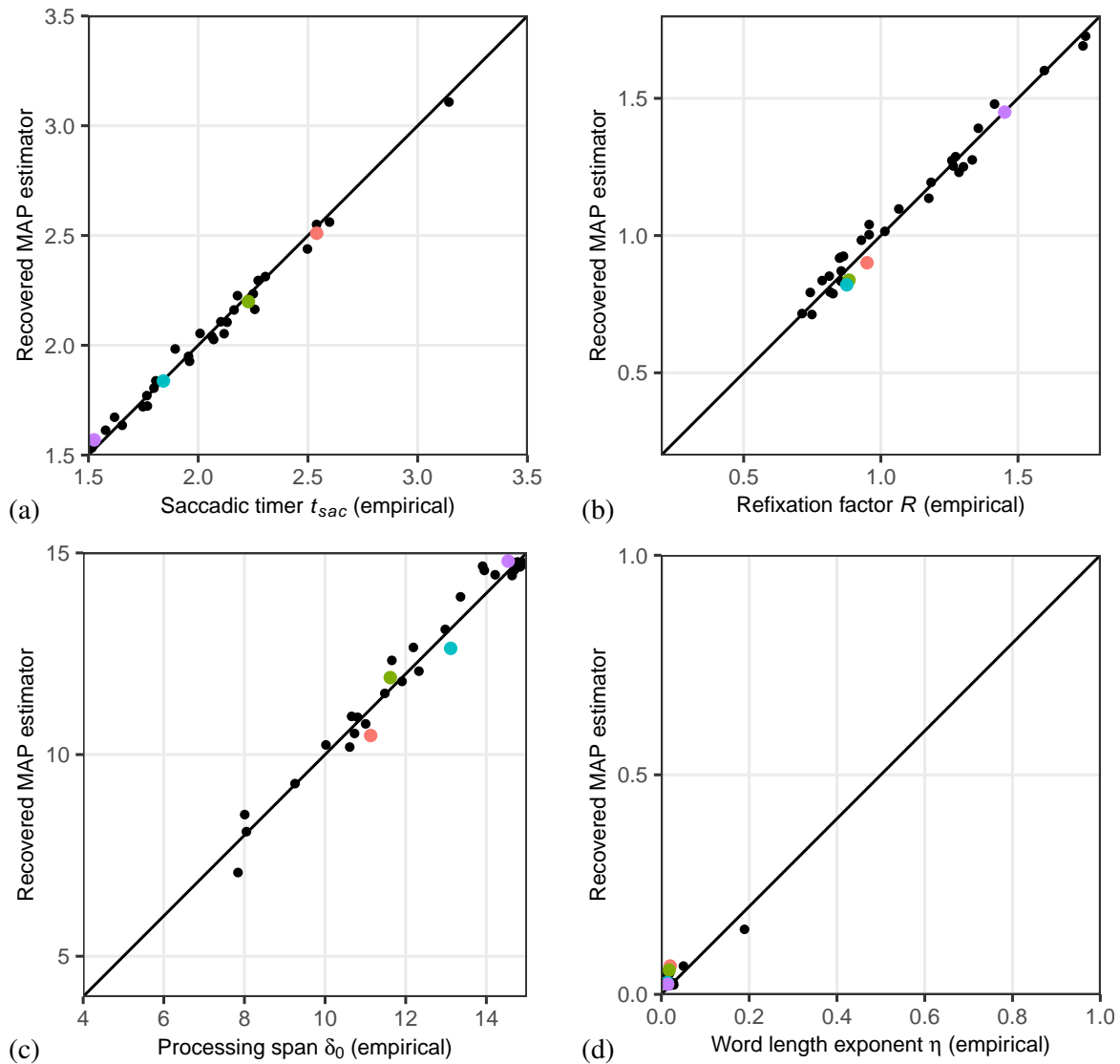
Note. Each density is calculated from the pooled data of 10 chains after the burn-in interval. Black ticks at the bottom indicate the MAP estimators for the individual chains. The prior distributions are indicated by the dotted, black line. Curves with the same color correspond to 4 highlighted participants.

model tend to be slightly underestimated for participants with longer initial fixations. Mean TVT, however, is higher in the model predictions than in the experiment. It is important to note that the TVT measure captures more complex gaze behavior, since it also incorporates additional fixation time due to regressions.

Fixation probabilities. Similar to the analysis of fixation durations, we calculated word-based probabilities for single fixations (SF), refixations (RF), regressions (RG), and word skipping (SK) (Fig. 2.10b). While in the experiment words are more likely to receive single

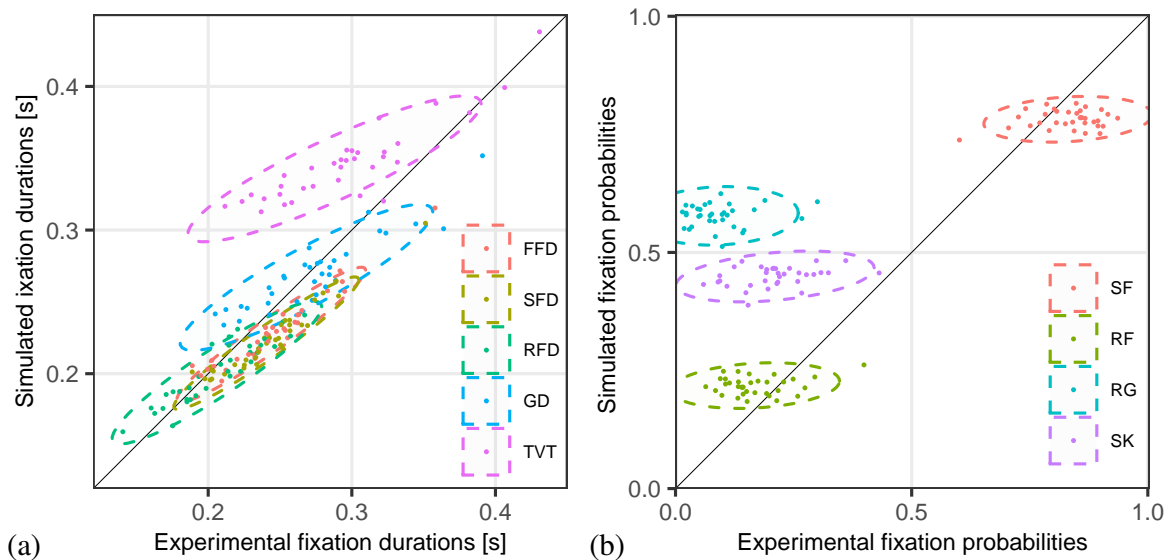
Figure 2.9

Relationship Between True Parameters and Estimated Parameter Values of Generated Data



Note. Relationship between true parameters (horizontal axis) and estimated parameter values of generated data (vertical axis). Parameters used are the MAP estimators for the experimental data. The colored points correspond to the same participants as in Figure 2.8.

fixations as compared to the simulated data, they consequently have a lower probability of receiving refixations. Additionally, the model predicts higher skipping probabilities and also higher probabilities of serving as regression target. It should be noted that the mismatch between experimental and simulated regression probabilities and experimental and simulated TVT (discussed above) is closely related. In general, part of the regressions might be looked upon as a more complicated psycholinguistic measure related to various aspects of post-lexical processing (Rayner, 1998) that cannot be captured in the SWIFT model, while another

Figure 2.10*Correlation of Empirical and Simulated Fixation Durations and Probabilities*

Note. (a) Means of different measures of fixation duration for experimental and corresponding simulated data. Each point represents one participant. Simulated data were created using the mean estimated parameters for each respective participant. The colored ellipses represent the 95% confidence boundaries. (b) Means of word based fixation probabilities. Again each point represents one participant.

portion of the regressions might be of oculomotor origin and can be found even in scanning tasks (Nuthmann et al., 2005).

In summary, our results indicate that estimated parameters can explain some of the interindividual differences in fixation durations and fixation probabilities. Thus, the likelihood-based MCMC approach to parameter inference could be applied successfully to estimate model parameters from individual behavioral data.

2.5 Discussion

Current approaches to parameter inference and model comparison (e.g., Reichle et al., 2003) for dynamical cognitive models are insufficient in at least three ways: First, dynamical models need to be tested against time-ordered observations. Second, a likelihood-based procedure is necessary for statistical inference. Third, parameter estimates are needed for individual subjects to explain interindividual differences based on specific model assumptions or components. We set out to solve these three issues in current modeling in computational cognitive science using the SWIFT model of eye-movement control during reading (Engbert et al., 2005) as a case study.

The approach discussed here is fundamentally based on the likelihood function of the

model. Therefore, we proposed and investigated the numerical likelihood computation of the SWIFT model. This approach is based on the observation that incremental prediction of fixation positions and fixation durations by the generative model can be exploited to determine the likelihood of the next fixation.

Since the likelihood can be decomposed into a spatial (i.e., fixation position) and a temporal part (i.e., fixation duration), we tried to find separate solutions to both problems. In the spatial part of the likelihood function, internal degrees of freedom (stochastic internal states) could not be integrated out due to numerical efficiency considerations; therefore, we computed a (stochastic) pseudo-likelihood (see Andrieu & Roberts, 2009). In the temporal part, the theoretical likelihood function was unavailable. Therefore, we constructed an approximate likelihood function using a sufficient number of predicted fixation durations from the SWIFT model and KDE for the approximation of the likelihood. In sum, we combined a pseudo-marginal spatial likelihood and an approximated pseudo-likelihood (see Holmes, 2015, for nomenclature) function to obtain the likelihood function of the model (Sisson & Fan, 2011).

Before we applied our framework to real data, we demonstrated that, in a simplified model version with 4 free parameters, we could reconstruct the true parameter values from simulated data. We used a Bayesian approach using MCMC sampling from the posterior distribution based on an adaptive sampling algorithm (Vihola, 2012). The size of the simulated data-set was comparable to a typical experimental data set that is recorded from an individual participant during a one-hour session of eye-tracking experimentation. Next, the same procedure was applied to experimental data. Motivated by the results from simulated data, we estimated model parameters independently for 34 subjects.

Finally, our results indicate that it is possible to relate interindividual differences in reading behavior (characterized by 5 different measures of fixation durations and 4 different measures of fixation probabilities) to differences in the estimated model parameters. Given the typical state-of-the-art models of eye-movement control in reading, this is a major step for generating hypotheses on the observed interindividual differences in a task as complex as reading.

Throughout the current work, we focused on the numerical implementation of the likelihood function for the SWIFT model. Since likelihood-based Bayesian inference turned out to be a viable and sound alternative to ad-hoc parameter estimation procedures, we expect that our approach can be further advanced for both theory building and modeling of interindividual differences. For example, for higher dimensional parameter spaces Differential Evolution MCMC algorithms (see, e.g., ter Braak, 2006; ter Braak & Vrugt, 2008) might be more adequate. Additionally, we expect that a hierarchical Bayesian design will help to increase the stability of the posterior estimates for individual subjects—even if we apply our methods to data sets smaller than used in the current work.

Chapter 3

A Bayesian Approach to Dynamical Modeling of Eye-Movement Control in Reading of Normal, Mirrored, and Scrambled Texts

This chapter has been published as: Rabe, M. M., Chandra, J., Krügel, A., Seelig, S. A., Vasishth, S., & Engbert, R. (2021). A Bayesian approach to dynamical modeling of eye-movement control in reading of normal, mirrored, and scrambled texts. *Psychological Review*, 128(5), 803–823. <https://doi.org/10.1037/rev0000268>

Abstract

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 et al., *Psychological Review*, 112, 2005, pp. 777–813). 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.

3.1 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 et al., 2006).

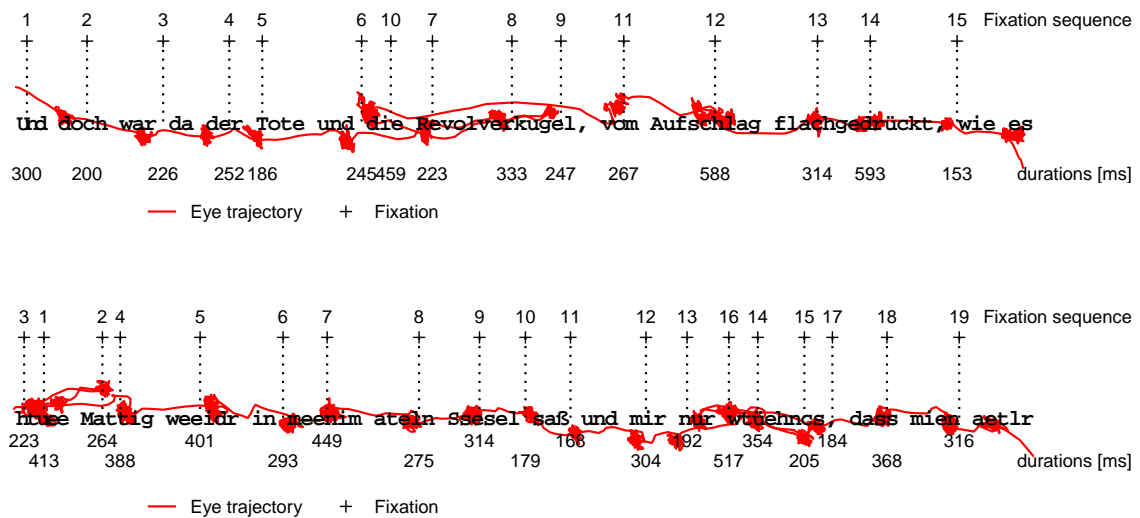
During reading of normal texts, the reader generates 3 to 4 saccades per second; word processing occurs during fixations on different words with average durations in a range between 150 and 300 ms (Rayner, 1998). The number of fixations in a given text is of the same order as the number of words, however, 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 3.1. 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 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 3.1).

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 et al., 2021). 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 (a) run more extensive

Figure 3.1*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.

simulations using more free parameters, (b) use a more powerful MPMC algorithm in the Bayesian framework, (c) reproduce a more representative range of reading behaviors using the full covariance structure of the fitted posterior distributions, (d) evaluate an experiment with 5 different reading conditions, and (e) 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 et al., 2021), our approach includes (a) rigorous statistical inference, (b) an evaluation of goodness-of-fit for specific effects, and (c) 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.¹³

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.

¹³Analyses are available online at <https://doi.org/10.17605/osf.io/t9sbf>.

Table 3.1
Reading Conditions Used As Modeling Targets

| Code | Description | Order | Example |
|------|------------------|-------|--|
| N | Normal | LR | Jede Sprache der Welt besitzt eine Grammatik |
| mL | Mirrored letters | LR | Ƶebɛ ʒqrɔscɛbɛr Wɛlt ɔɛɪtʒt ɛɪnɛ ɔrɔmmɔtɪk |
| sL | Scrambled | LR | Jdee Scrahpe der Wlet bsizett enie Gmartimak |
| iW | Reversed letters | RL | edeJ ehcarpS red tleWtztiseb enie kitammarG |
| mW | Mirrored word | RL | ɛbɛ ʒqrɔscɛbɛr Wɛlt ɔɛɪtʒt ɛɪnɛ ɔrɔmmɔtɪk |

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

3.1.1 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 et al., 2021). 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,

$$X_n = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \\ x_{n1} & & x_{nm} \end{pmatrix}, \quad (3.1)$$

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 $P_M(\theta | X)$ for a model M specified by a set of parameters θ after observing data X , we

first need to determine the likelihood $L_M(\boldsymbol{\theta} | X)$ of the data X given some parameter set $\boldsymbol{\theta}$ and the prior probability distribution $Q(\boldsymbol{\theta})$ over parameters $\boldsymbol{\theta}$, so that

$$P_M(\boldsymbol{\theta} | X) \propto L_M(\boldsymbol{\theta} | X) \cdot Q(\boldsymbol{\theta}) . \quad (3.2)$$

While the definition of $L_M(\boldsymbol{\theta} | X)$ 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 (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 (a) to characterize different theoretical entities (i.e., subjects or items) and (b) 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 $L_M(X_n | \boldsymbol{\theta})$ for time-ordered dataset X_n given some model configuration $\boldsymbol{\theta}$. The likelihood function is the product of the likelihood of all instances x_i of X_n , each conditional on model parameters $\boldsymbol{\theta}$ and all previous instances $X_{i-1} = (x_1, \dots, x_{i-1})$, i.e.,

$$L_M(X_n | \boldsymbol{\theta}, \boldsymbol{\xi}) = L_M(x_1 | \boldsymbol{\theta}, \boldsymbol{\xi}) \prod_{i=2}^n L_M(x_i | \boldsymbol{\theta}, \boldsymbol{\xi}, X_{i-1}) . \quad (3.3)$$

The additional variable $\boldsymbol{\xi}$ 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 $L_M(X_n | \boldsymbol{\theta}, \boldsymbol{\xi})$ (Andrieu & Roberts, 2009) with internal degrees of freedom $\boldsymbol{\xi}$.

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 et al. (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.

3.1.2 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 et al. (2021) 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,¹⁴
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.

¹⁴As 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.

3.1.3 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 (S. 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 et al., 2012; Reilly & Radach, 2006; Snell et al., 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 skipplings, 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 via the modified Gamma-distributed saccade lengths (Appendix C).

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.

3.2 The SWIFT Model of Eye-Movement Control

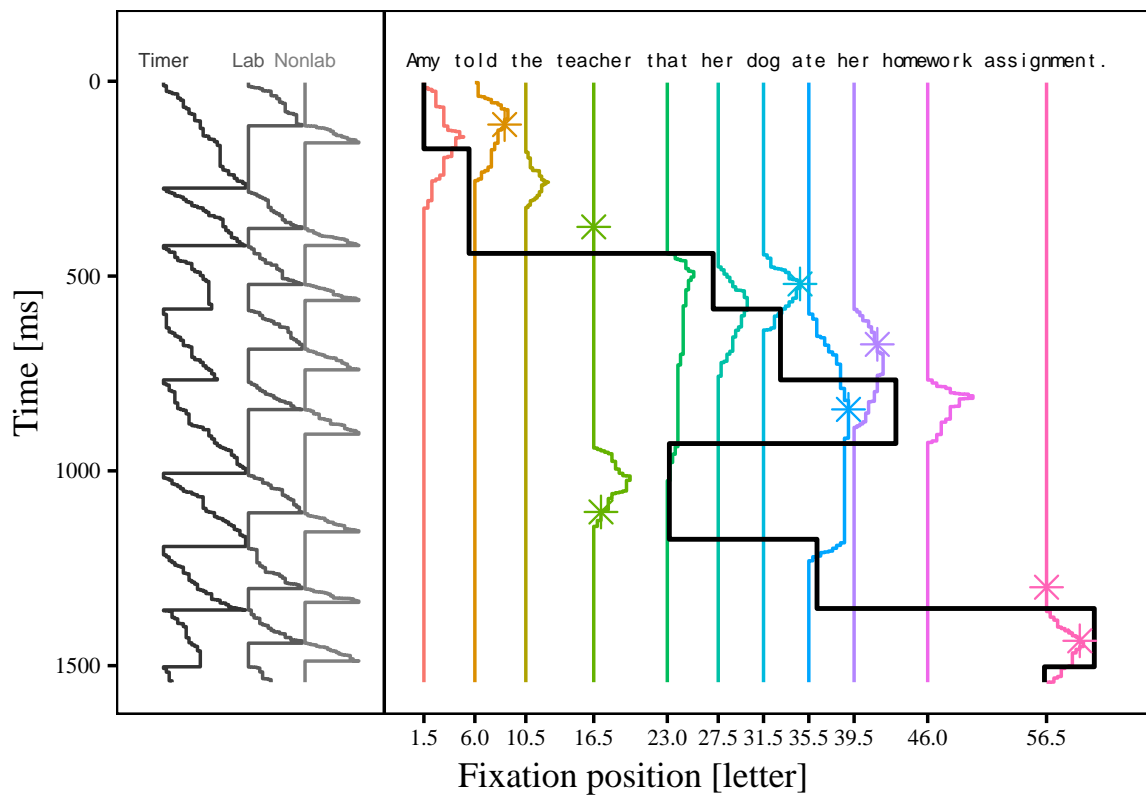
The SWIFT model (Engbert et al., 2005; Seelig et al., 2020) is a dynamical cognitive model of eye-movement control during reading. The model can describe, explain, and predict temporal and spatial aspects that are commonly observed in eye trajectories recorded during natural reading. It is among several competitor models that aim at predicting and explaining similar eye movement statistics (e.g., see Reichle et al., 1998; Reilly & Radach, 2006; Snell et al., 2018). In Figure 3.2, 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 et al., 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 3.2) 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.

Saccade execution is modulated by oculomotor errors (Engbert & Krügel, 2010; Krügel & Engbert, 2014). In fact, McConkie et al. (1988) proposed the *saccadic range error* (SRE) model, stating that the landing position is driven by systematic and random contributions,

Figure 3.2*An Eye Trajectory As Simulated in SWIFT*

Note. The thick black line is the simulated eye 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.

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 overshoot 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 B provides more mathematical details of key aspects of SWIFT, while Appendix C extends on the oculomotor assumptions.

3.3 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 3.2) 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 et al., 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 ,

$$F_n = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} k_1 & l_1 & T_1 & s_1 \\ \vdots & \vdots & \vdots & \vdots \\ k_n & l_n & T_n & s_n \end{pmatrix} \quad (3.4)$$

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.,

$$L_M(k_i, l_i, T_i | F_{i-1}, \boldsymbol{\theta}, \boldsymbol{\xi}) = P_{\text{temp}}(T_i | k_i, l_i, F_{i-1}, \boldsymbol{\theta}, \boldsymbol{\xi}) \cdot P_{\text{spat}}(k_i, l_i | F_{i-1}, \boldsymbol{\theta}, \boldsymbol{\xi}), \quad (3.5)$$

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

The internal degrees of freedom $\boldsymbol{\xi}$ 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 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.

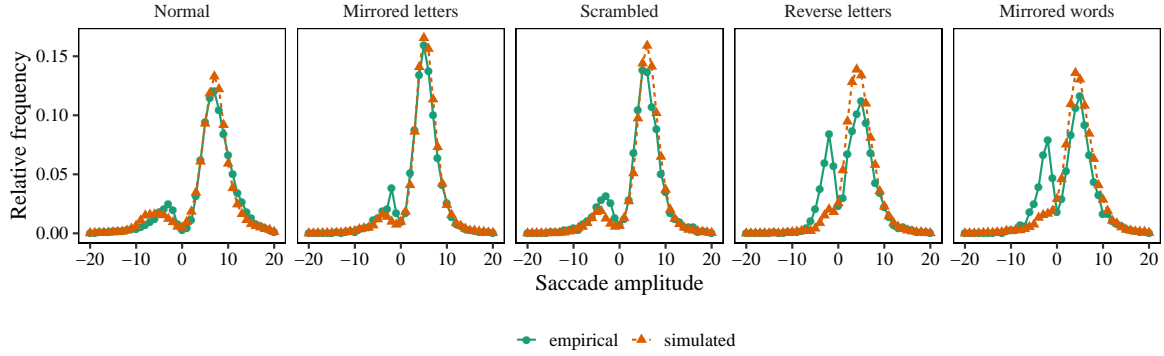
The spatial likelihood P_{spat} is further decomposed into the probability q to land on word k_i and letter l_i within word k_i after having selected word m with selection probability¹⁵ π following the initiation of a saccade at time T_i (see Equation 3.6). For an observed fixation i , it is unknown 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 $\pi(m|\cdot)$ and oculomotor error probability $q(k_i, l_i | m, \cdot)$ by summation over all words m of the sentence ($m = 1, 2, \dots, N_W$), i.e.,

$$P_{\text{spat}}(k_i, l_i | T_i, F_{i-1}, \boldsymbol{\theta}, \boldsymbol{\xi}) = \sum_{m=1}^{N_W} \pi(m | T_i, F_{i-1}, \boldsymbol{\theta}, \boldsymbol{\xi}) \cdot q(k_i, l_i | m, F_{i-1}, \boldsymbol{\theta}) \quad (3.6)$$

¹⁵The selection probabilities π are normalized so that $\sum_{m=1}^{N_W} \pi(m) = 1$.

Figure 3.3

Empirical and Previously Simulated Gaussian Saccade Amplitudes Aggregated Across All Subjects in Each Experimental Condition



Note. For saccade amplitudes generated with improved oculomotor assumptions, see Figure 3.10.

According to the SRE model of saccade amplitudes (McConkie et al., 1988), the systematic component ϵ_{sre} , Equation C.5, mainly shifts the mean landing position and the random component σ_{sre} , Equation C.6, modulates the variance of the distribution of landing positions (see Appendix C for mathematical details). While the selection probability $\pi(\cdot)$ in SWIFT is driven by a time-dependent word activation field, the observable landing position, or its probability $q(\cdot)$, depends on oculomotor process assumptions and only indirectly on the implementation of the word activation field.

The oculomotor assumptions, explicitly given by the probability $q(\cdot)$, Equation (3.6), 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 (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 3.3). 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 C).

Finally, for the temporal probability density P_{temp} in Equation (3.6), 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, 2013).

3.4 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 3.2) 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.

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 et al., 2005; Vitu et al., 2001) or lag and successor effects as indicators for spatially distributed processing (Kliegl et al., 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 ap-

Table 3.2
Fitted SWIFT Model Parameters

| Parameter | Description |
|------------------------------|---|
| α | Baseline word difficulty |
| β | Word frequency modulation |
| ω | Global decay during postlexical processing |
| δ | Non-dynamical processing span in letter spaces |
| η | Word length modulation |
| γ | Target selection exponent |
| M | Relative duration of the labile saccade stage for misplaced fixations |
| t_{sac} | Relative duration of global saccade program |
| omn_1 | Intercept term for random oculomotor noise ^a |
| omn_2 | Slope term for random oculomotor noise ^a |
| R | Relative duration of the labile saccade stage for well-placed refixations |
| sre_1 | Intercept term for saccadic range error for forward fixations and skipping ^b |
| $\text{sre}_1^{(\text{RF})}$ | Intercept term for saccadic range error for refixations ^b |
| $\text{sre}_2^{(\text{FS})}$ | Slope term for saccadic range error for forward fixations ^b |
| $\text{sre}_2^{(\text{RF})}$ | Slope term for saccadic range error for refixations ^b |
| $\text{sre}_2^{(\text{SK})}$ | Slope term for saccadic range error for skipplings ^b |
| $\tau_{\text{n/l}}$ | Mean durations of the labile and non-labile saccade programs ^c |

Note. ^aParameters 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. ^b 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 $\text{sre}_1 = \text{sre}_2 = 0$ to disable saccadic range error. ^cParameters 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 $\frac{1}{2}\tau_l = \tau_n = \tau_{\text{n/l}}$.

prox. 10,000 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 Figure 3.6). 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.

3.5 Experiment

In order to demonstrate our approach and validate the model, we chose an experimental study¹⁶ 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 3.1 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.

3.5.1 Data Preprocessing

From the initially recorded data, all trials including blinks were discarded. We used the velocity-based algorithm by Engbert et al. (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

¹⁶The experimental data are available at <https://osf.io/bmvr/x/>.

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).

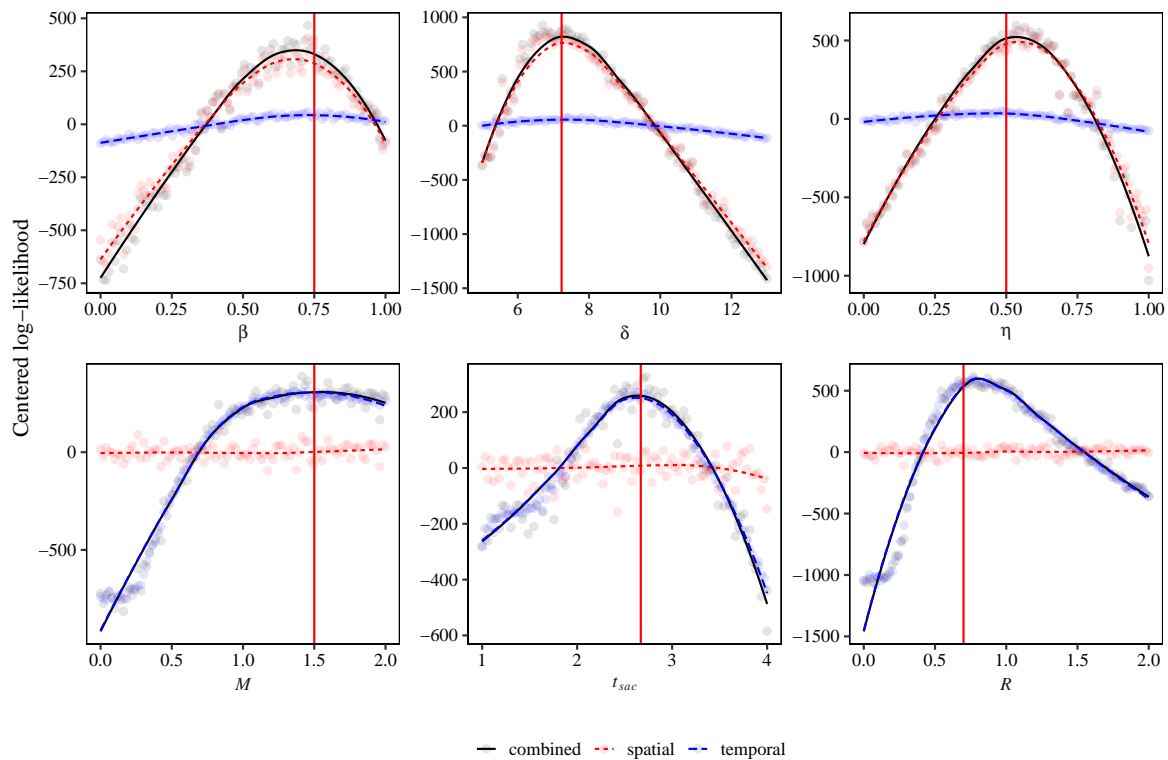
For each subject in each condition, remaining datasets were split into a fitting (training) and validation (test) dataset. Trials within each dataset were shuffled, keeping the sequence of fixations within a trial intact, and the split criterion incrementally shifted trial by trial until 70% of all fixations 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.

3.6 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 statistical analysis of model parameter estimates across participants and experimental conditions.

3.6.1 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, Equation 3.5) as well as the combined likelihood (L_M), which is exclusively used for further fitting purposes. As shown below, likelihood profiles (a) peak at the true value, (b) have identical maxima for simulated data, and (c) 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 3.4).

Figure 3.4*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.

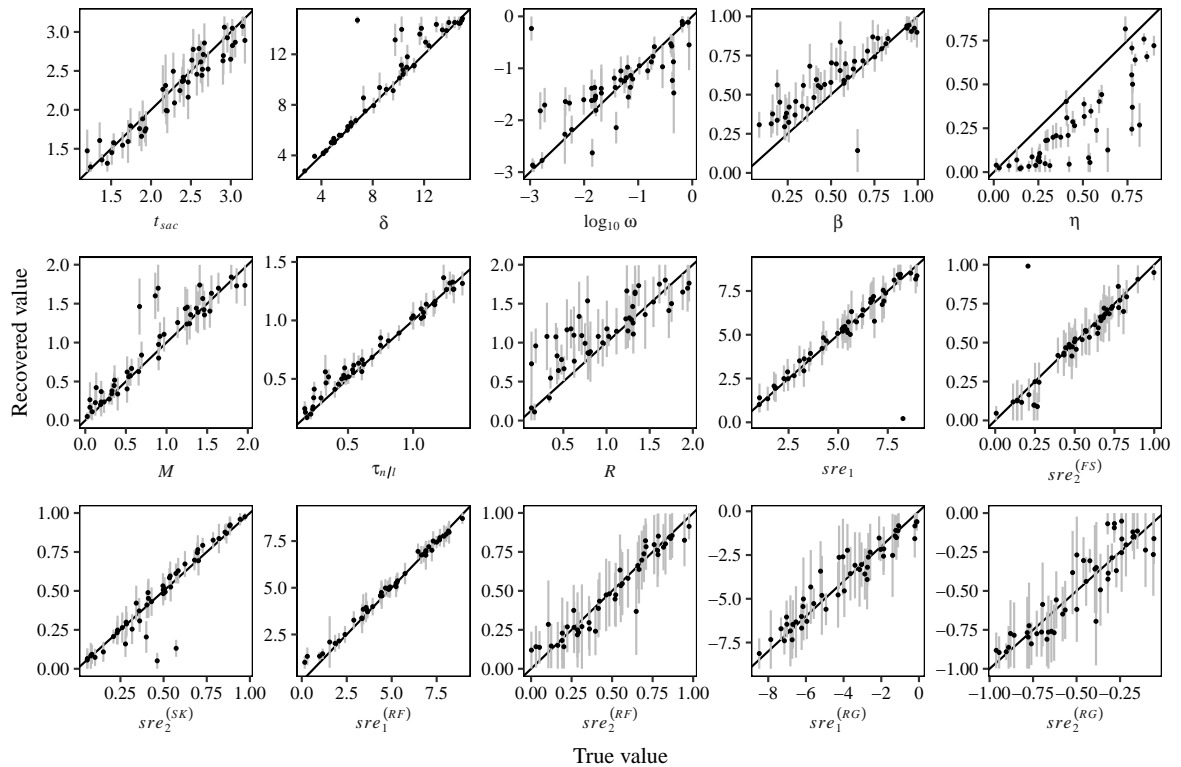
3.6.2 Parameter Recovery

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 3.5, 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

Figure 3.5

Scatterplot of True and Recovered Parameters With 60% HPDIs



Note. Vertical grey lines are 60% HPDIs across simulated data sets. The diagonal line indicates identity, i.e., credible intervals touching the diagonal include the true value.

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.

3.6.3 Experimental Data: Summary Statistics

In Table 3.3, 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

Table 3.3
Empirical Means and Standard Errors in Summary Statistics Aggregated Across Subjects

| Metric | 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.

indicated none of these as statistically different from the accuracy observed in the normal reading condition.

3.6.4 Parameter Estimates

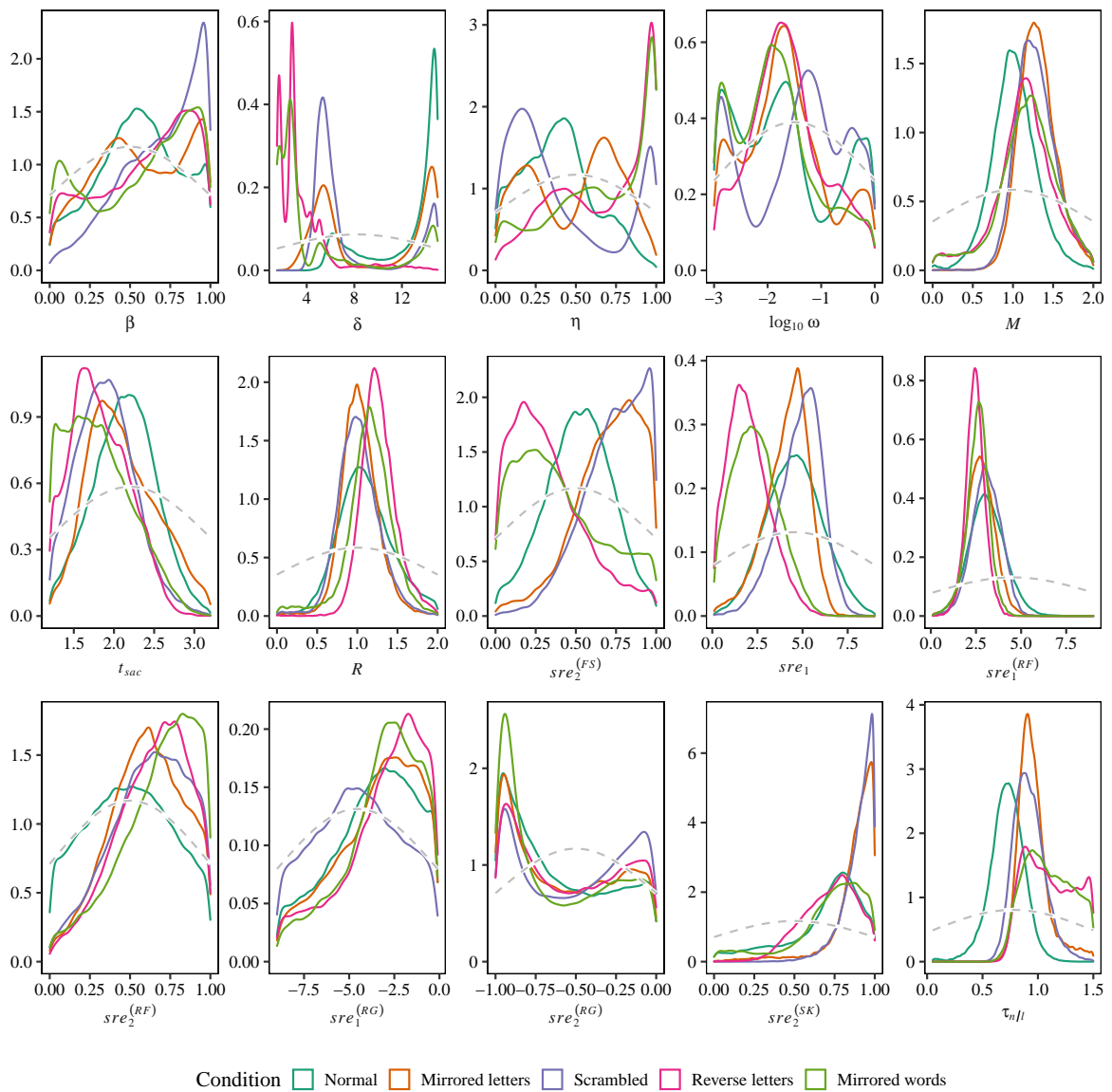
For every participant of the experiment, posteriors were generated using MCMC sampling in both the normal and a manipulated reading conditions. In Figure 3.6, all samples were aggregated across subjects for the five experimental conditions. The corresponding distributions indicate how the posteriors deviate, on average, between experimental conditions. It appears that some model parameters (e.g., $\text{sre}_1^{(\text{RF})}$) 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 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.

3.6.5 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

Figure 3.6
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 3.4

Change in MSE Across Subject-Level Summary Statistics Between Posterior Sampling and Point Estimates

| Metric | 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% |

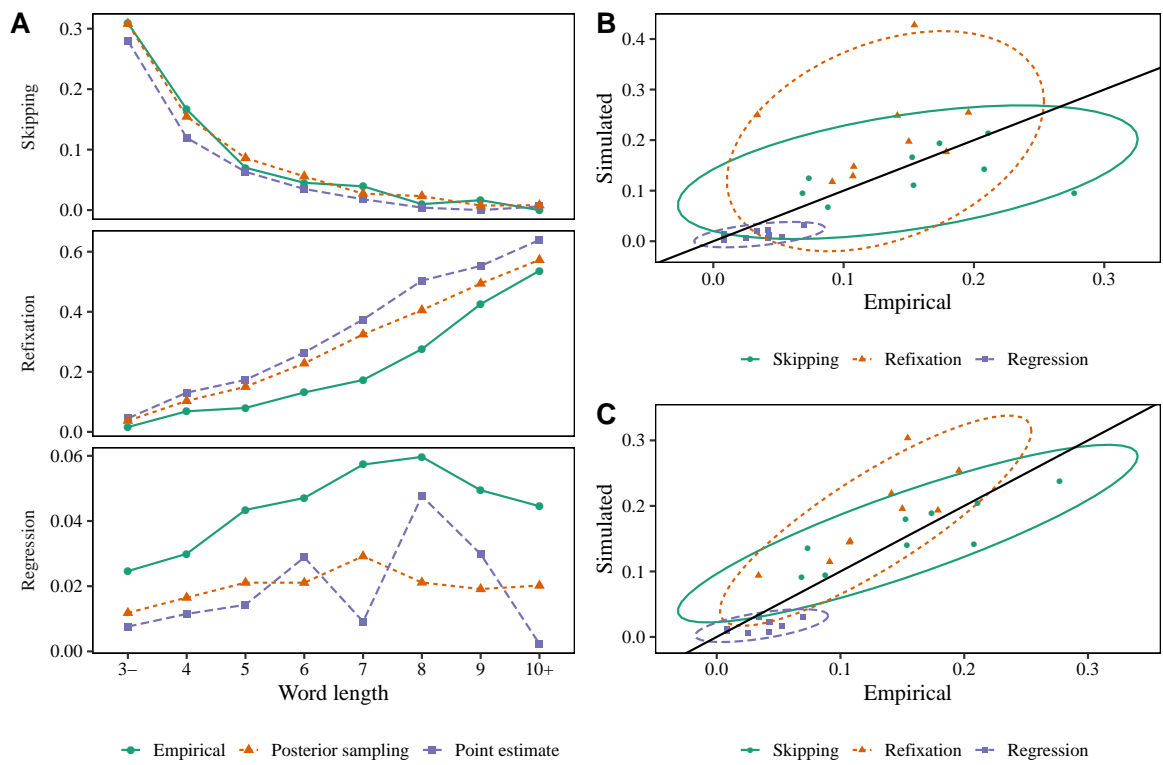
Note. Negative percentages are reductions of the mean squared error (MSE) when using posterior sampling relative to the MSE when using point estimates. Positive percentages are increases.

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 3.4 and Figure 3.7, 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.

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 3.8), including word length effects on those at a

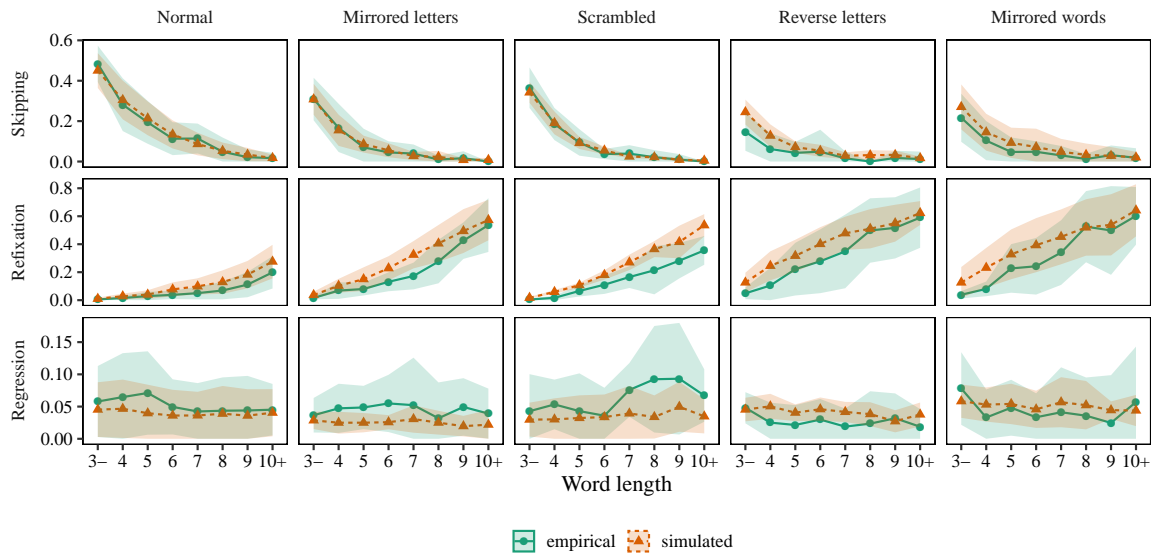
Figure 3.7
Comparison of Simulated Summary Statistics When Sampling From the Posterior Vs. using Point Estimates



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).

Figure 3.8

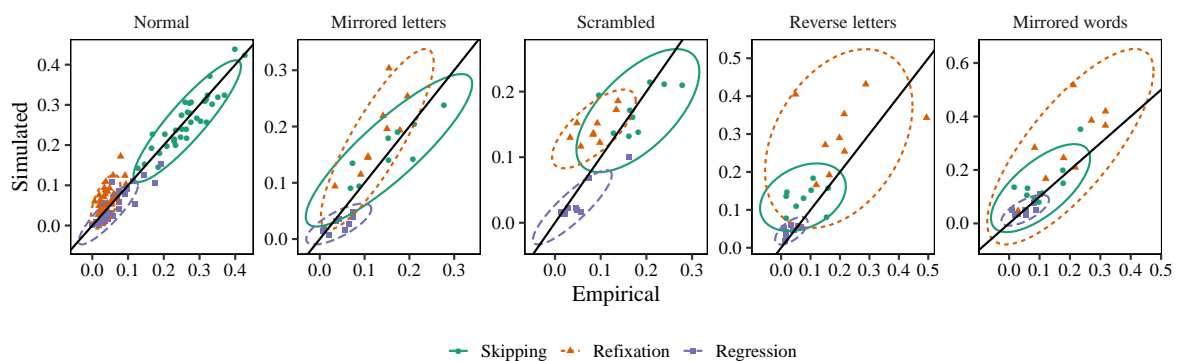
Empirical and Simulated Spatial Summary Statistics (Fixation Probabilities) for Different Experimental Conditions, Aggregated Across Subjects, As a Function of Word Length



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

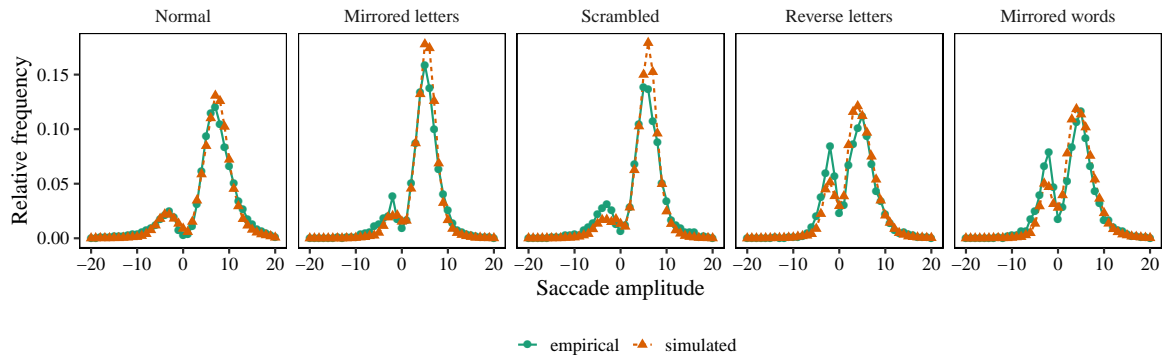
Figure 3.9

Correlation Between Empirical (Horizontal Axis) and Simulated (Vertical Axis) Spatial Summary Statistics (Fixation Probabilities)



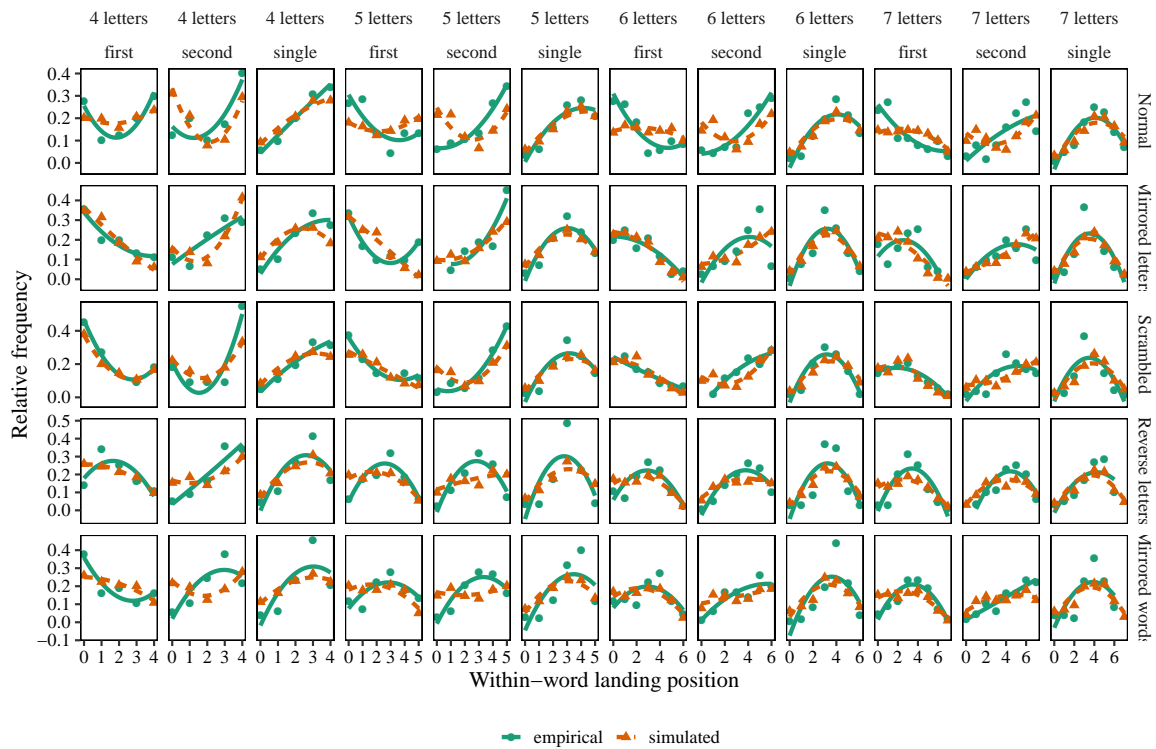
Note. Each subject is represented by one dot in each color in the respective experimental condition (panel).

Figure 3.10
Empirical and Simulated Saccade Amplitudes Aggregated Across All Subjects in Each Experimental Condition



Note. Saccades were generated using Gamma-distributed saccade amplitudes.

Figure 3.11
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.

Table 3.5
Correlations Across Subjects for Empirical Vs. simulated Summary Statistics

| Metric | 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_S -values in parentheses (bold font for $p_S < 0.01$).

global level. There is, however, still some divergence with regard to regression probabilities, namely that the model 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 3.9 and Table 3.5, most conditions, the averages across subjects (ellipsis midpoints) correlate closely across statistics and the subject-level variance (covariance within each ellipsis) is captured very reliably.

Moreover, results for saccade amplitudes are clearly important (see Figure 3.10), supporting the notion of Gamma-distributed saccade length distributions. In contrast to the previous Gaussian saccade amplitudes (see Figure 3.3), 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 C.1 in Appendix C also shows a comparison to previously Gaussian distributed saccade amplitudes, which fit the data less satisfactorily. As depicted in Figure 3.11, 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 3.12). When compared at

a by-subject level (see Figure 3.13 and Table 3.5), 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).

3.6.6 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 3.14) 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 et al., 2020; Schad 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, H_{0_4} 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

Figure 3.12

Empirical and Simulated Temporal Summary Statistics (Fixation Durations) for Different Experimental Conditions, Aggregated Across Subjects, As a Function of Word Length

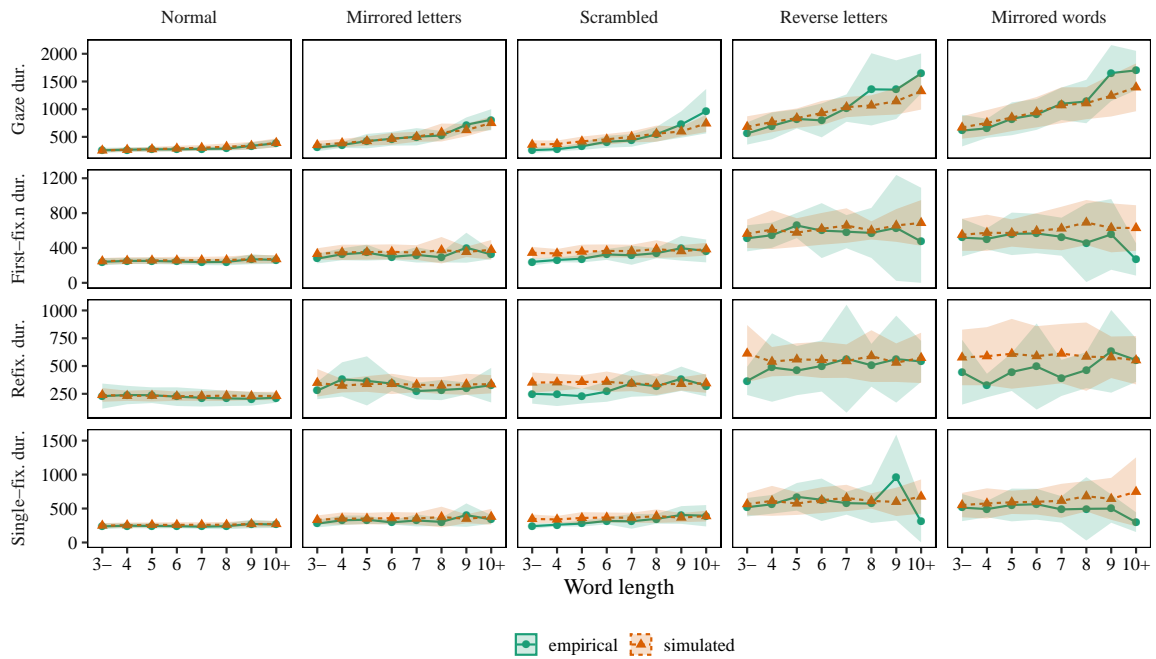
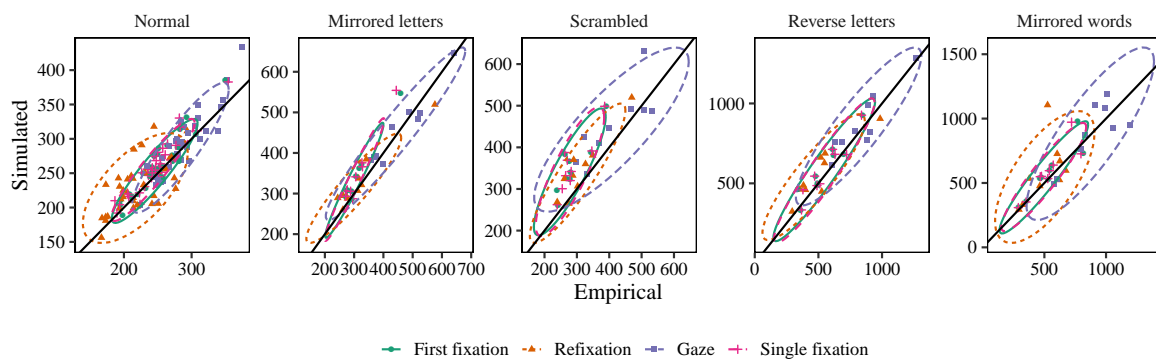
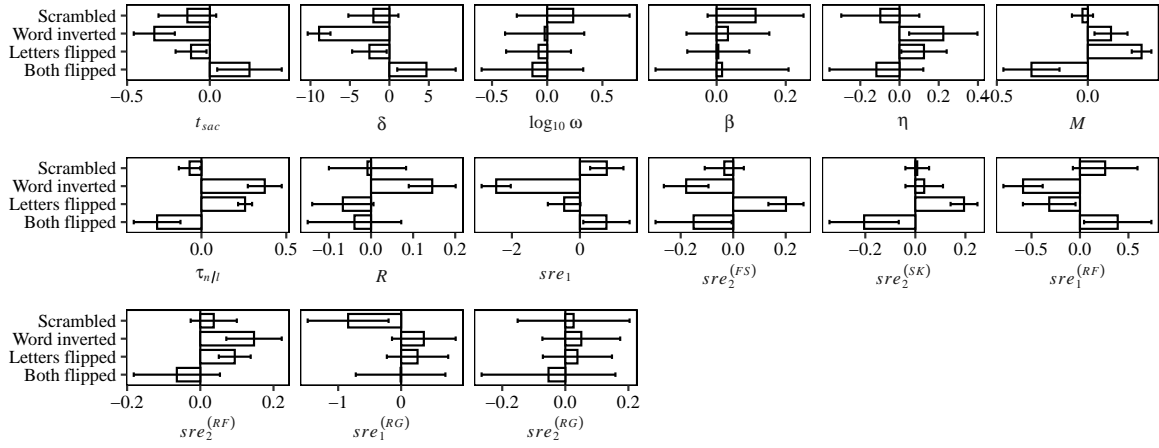


Figure 3.13

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).

Figure 3.14*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.

($sre_1^{(RF)}$, $b = -0.59$, $p_S < 0.003$), 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, skipings, or refixations, as suggested by greater SRE slopes and thus reduced oculomotor control, $\text{sre}_2^{(\text{FS})}$ ($b = 0.20$, $p_S < 0.002$), $\text{sre}_2^{(\text{SK})}$ ($b = 0.19$, $p_S < 0.001$), and $\text{sre}_2^{(\text{RF})}$ ($b = 0.09$, $p_S < 0.017$), 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 $\tau_{n/1}$ ($b = -0.26$, $p_S < 0.020$) 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.

3.7 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 et al., 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 pre-defined 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 et al., 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 et al., 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 et al., 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.

Chapter 4

SEAM: An Integrated Activation-Coupled Model of Sentence Processing and Eye Movements in Reading

This chapter has been submitted for publication to the Journal of Memory and Language and posted as a preprint: Rabe, M. M., Paape, D., Mertzen, D., Vasishth, S., & Engbert, R. (2023). *SEAM: An integrated activation-coupled model of sentence processing and eye movements in reading*. <https://doi.org/10.48550/arXiv.2303.05221>

Abstract

Models of eye-movement control during reading, developed largely within psychology, usually focus on visual, attentional, lexical, and motor processes but neglect post-lexical language processing; by contrast, models of sentence comprehension processes, developed largely within psycholinguistics, generally focus only on post-lexical language processes. We present a model that combines these two research threads, by integrating eye-movement control and sentence processing. Developing such an integrated model is extremely challenging and computationally demanding, but such an integration is an important step toward complete mathematical models of natural language comprehension in reading. We combine the SWIFT model of eye-movement control (Seelig et al., *Journal of Mathematical Psychology*, 95, 2020, Article 102313) with key components of the Lewis and Vasishth sentence processing model (Lewis and Vasishth, *Cognitive Science*, 29, 2005, pp. 375–419). This integration becomes possible, for the first time, due in part to recent advances in successful parameter identification in dynamical models, which allows us to investigate profile log-likelihoods for individual model parameters. We present a fully implemented proof-of-concept model demonstrating how such an integrated model can be achieved; our approach includes Bayesian model inference with Markov Chain Monte Carlo (MCMC) sampling as a key computational tool. The integrated model, SEAM, can successfully reproduce eye movement patterns that arise due to similarity-based interference in reading. To our knowledge, this is the first-ever integration of a complete process model of eye-movement control with

linguistic dependency completion processes in sentence comprehension. In future work, this proof of concept model will need to be evaluated using a comprehensive set of benchmark data.

4.1 Introduction

What is the relationship between sentence processing and eye movements during reading? As an answer to this question, Just and Carpenter (1980, pp. 330–331) famously coined the eye-mind assumption, which states that “the eye remains fixated on a word as long as the word is being processed”, and that “there is no appreciable lag between what is being fixated and what is being processed”. But what does it mean for a word to be “processed”? Just and Carpenter’s model of reading has three stages: Encoding of the word form and lexical access, identification of relationships between the words in a sentence (such as agent-action-object), and integration with information from previous sentences. Once these three stages are finished, the eyes proceed to the next word.¹⁷ Just and Carpenter’s processing model is highly serial, which matches most readers’ subjective experience that sentences are processed in an incremental, left-to-right fashion (Snell & Grainger, 2019). However, while readers do tend to make fixations incrementally in the reading direction, fixation sequences are not always in serial order: Instead of systematically shifting the gaze from one word to the next – something that only happens in about 50% of fixations – readers also skip words, refixate the same word, or regress to previous words (Kliegl et al., 2004; Rayner, 1998).

This more complicated picture of reading aligns with the fact that the structure of many sentences in natural language does not correspond to simple agent-action-object sequences. Consider a sentence like (1), taken from Mertzen et al. (2023):

- (1) It turned out that the attorney whose secretary had forgotten that the visitor was important frequently complained about the salary at the firm.

In this sentence, there are several dependencies between non-adjacent words, most strikingly the long-distance dependency between the noun *attorney* and the verb *complained*. It is difficult to argue that the processing of the word *attorney* is finished once the preamble *It turned out that the attorney . . .* has been read: It is clear that a verb must arrive at some point of which *attorney* is the subject. Complete integration of *attorney* can thus only be achieved when *complained* is read after ten intervening words have been processed. It is therefore clear that the eyes will have to move forward even if the current word has not been completely integrated into the sentence structure.

¹⁷There is a fourth stage in the model, called wrap-up, which only occurs at the end of a sentence, and whose purpose is to finish any processing that could not be completed at a previous point during reading (but see Warren et al., 2009, for a critical discussion).

A well-established assumption in sentence processing is that a noun like *attorney* is held in working memory until the dependency is completed, and needs to be retrieved when the verb is reached (Gibson, 1998, 2000; Lewis et al., 2006). A strong interpretation of the eye-mind assumption would predict that, given that the processing of *attorney* is finalized at *complained*, readers should refixate *attorney* once lexical access of *complained* is complete. However, this is not what usually happens: While readers do make more regressions in more complex sentences that involve memory retrievals (e.g., Gordon et al., 2006; Jäger et al., 2015; Lee et al., 2007; Mertzen et al., 2023), regressive eye movements nevertheless occur only in a minority of trials. Furthermore, even in difficult sentences that may require multiple passes to parse correctly, readers do not necessarily regress to the most syntactically informative words in the sentence (e.g., Christianson et al., 2017; Engelmann et al., 2013; von der Malsburg & Vasishth, 2011; von der Malsburg & Vasishth, 2013). Thus, while there is undoubtedly a connection between sentence processing and eye movements (Clifton et al., 2007; Frazier & Rayner, 1982; Rayner, 1998), it is much less direct than posited by the strong version of the eye-mind assumption, as Reichle et al. (2009) have pointed out. On the other hand, there *is* evidence that readers can and do move their eyes into the vicinity of critical words (Inhoff & Weger, 2005; Meseguer et al., 2002; Mitchell et al., 2008; Schotter et al., 2014; Weger & Inhoff, 2007), which suggests the need for a model with *some* linguistically-mediated guidance of regressive eye movements.

Psycholinguistic studies of sentence processing typically rely on aggregated reading measures such as total fixation times, and models of language processing during reading, such as the classic Just and Carpenter (1980) model, usually ignore the complexity of eye-movement control. However, highly detailed models of eye-movement control do exist. An important line of work in cognitive psychology seeks to explain reading processes at the level of individual fixations and saccades by unpacking the underlying dynamics of the latent subprocesses involved. Several influential mathematical models of eye-movement control exist; a prominent example is the E-Z Reader model (Reichle et al., 2003). These models have historically focused on the effects of word-level properties such as word length, frequency, and predictability, and do not take into account higher-level processes such as linguistic dependency completion. However, there have been several attempts at integrating models of sentence processing difficulty with eye-movement control, including E-Z Reader (Reichle et al., 2009), the model of Engelmann et al. (2013), and Über-Reader (Reichle, 2021; Veldre et al., 2020). These models focus on different aspects of sentence processing, and have been evaluated against corpus data, such as the Schilling corpus (Schilling et al., 1998). Two models that investigate the interaction between eye-movement control and sentence comprehension using data from planned experiments are reported in Vasishth and Engelmann (2022) and Dotlačil (2021); both these investigations use a highly simplified version of E-Z Reader, that is, the Eye Movements and Movement of Attention (EMMA) model embedded

within the ACT-R architecture (Salvucci, 2001). The simplified EMMA model has important limitations; for example, as discussed in Engelmann et al. (2013), the model only allows regressive eye movements to the preceding word.

All of these existing models do capture a range of selected empirical phenomena and furnish important insights into the interaction between eye-movement control and sentence parsing processes. However, to our knowledge, no model exists that uses a fully specified mainstream model of eye-movement control that is integrated with a model of dependency completion in language comprehension; furthermore, as far as we are aware, such a detailed process model has never been evaluated using data from a planned psycholinguistic experiment.

A major difficulty in developing a more complex integrated model is that a considerable number of model parameters will need to be estimated using empirical data. For models of such complexity, conventional methods like grid search will lead to intractability. In order to implement such a complex model, Bayesian parameter estimation using the model's likelihood function (or an approximation) provides a rigorous approach to statistical inference (Rabe et al., 2021; Schütt et al., 2017). Two major advantages of the Bayesian approach are that parameters can be regularized or constrained a priori, which makes computation more efficient compared to the traditional grid search method, and that the uncertainty of the parameter estimates can be taken into account when evaluating model fit. Regularization makes parameter estimation more tractable, and incorporating the uncertainty of parameter estimates gives a more realistic picture of model fit (Nicenboim et al., 2023). Although Bayesian model fitting has been implemented for a basic reading model (Dotlačil, 2018), this line of work currently still neglects many low-level physiological and higher-level cognitive aspects of reading.

In this context, the major recent advance in Bayesian parameter inference for modeling process-based models has been proposed by Rabe et al. (2021) and Seelig et al. (2020) (for an overview, see Engbert et al., 2022). This line of work relies on the dynamical model of eye movement control developed by Engbert et al. (2005), and demonstrates how the Bayesian approach can be deployed in highly complex process models. Compared to other models of eye-movement control in reading such as E-Z Reader (Reichle et al., 2003), SWIFT has several advantages that make it a potentially better candidate for the purpose of integrating higher-level processing: It (1) is available for Bayesian parameter inference due to the likelihood implementation (Rabe et al., 2021; Seelig et al., 2020), (2) has a time-dependent word-activation field that can serve as the basis for memory encodings, and (3) has mechanisms that allow for long-range regressions, which are of particular interest when investigating dependencies that span several words. SWIFT and E-Z Reader also differ with regard to theoretical assumptions such as serial vs. parallel processing of words, but these are not our primary focus. Based on the methodological advances by Rabe et al. (2021) and Seelig et al.

(2020), we are able to find an objective answer to the question: Can the complex lower-level cognitive and physiological principles of eye movements be integrated with a computational model of higher-level linguistic processing, taking into account the cost of long-distance dependency completion?

Below, we present the Sentence-Processing and Eye-Movement Activation-Coupled Model (SEAM), a novel integrated model of sentence processing and eye movement control in reading. By combining the Saccade-Generation With Inhibition by Foveal Targets (SWIFT) model with the cue-based memory retrieval model proposed by Lewis and Vasishth (2005), we can integrate spatially-distributed processing in eye movement control with rule-based dependency completion in a Bayesian model-fitting framework. We carry out model simulation using a principled Bayesian workflow (Schad et al., 2021) to demonstrate the activation-based coupling between SWIFT and the Lewis and Vasishth (2005) model. As a result, our model yields reliable Bayesian parameter estimates by generating simulated data with known parameters, and then recovering these parameters using the Bayesian parameter estimation approach.

We also fit SEAM to recently-published empirical data from an eye-tracking experiment investigating similarity-based interference (Mertzen et al., 2023), providing model-driven explanations for the observed eye movement patterns. Given that SEAM simulates time-ordered fixation sequences, the model makes predictions for all spatial and temporal summary statistics that are relevant in the reading research literature (e.g., fixation probabilities, landing positions/saccade amplitudes, and fixation durations/reading times). This capability of the SEAM architecture makes it an important candidate model for theory development in psycholinguistics.

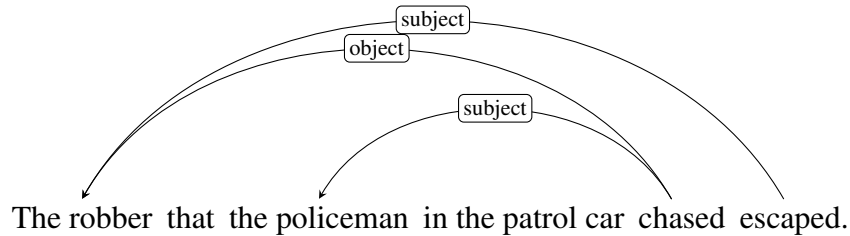
We will first introduce the Lewis and Vasishth (2005) model of sentence processing, then introduce the basic workings of SWIFT, and finally proceed to our integrated model SEAM.

4.1.1 The Activation-Based Model of Sentence Processing (Lewis & Vasishth, 2005)

During sentence reading, the human sentence processor has to incrementally integrate individual words into a syntactic structure, based on which sentence meaning can be derived. Lewis and Vasishth (2005) proposed a model of sentence processing (hereafter, we refer to this model as LV05) that is based on the cognitive architecture ACT-R (Anderson & Lebiere, 1998; Anderson, 2005). In the LV05 model, incoming words are incrementally integrated into syntactic constituents that are stored in memory as *chunks*. Memory chunks in LV05 carry information in the form of features, which can be used to access them in memory later on. Chunks also have fluctuating activation values that are determined by recency and by cue match during retrieval events. For instance, in a sentence like (2), as the sentence is read

word-by-word, the noun phrases *the robber* and *the policeman* are stored as memory chunks as soon as they are read. The verbs *chased* and *escaped* then each trigger retrievals of their respective arguments from memory.

(2)



Taking the retrieval at the verb *escaped* as an example, the dependency needs to be completed by searching working memory for a suitable memory chunk to serve as a syntactic subject. The search process is cue-based, that is, the verb specifies a set of linguistic features such as \pm noun or \pm animate to identify the correct dependent, and existing memory chunks are reactivated based on their feature specifications. The best-matching candidate is usually retrieved, but because memory activation is noisy, misretrievals occasionally occur. In addition, processing is slowed when multiple memory chunks, such as *the robber* and *the policeman* in (2), match the retrieval cues and compete for activation, which is called the fan effect (e.g., Anderson, 1990).

In LV05, the latency of a given retrieval is governed by a set of equations taken from the ACT-R architecture (Anderson et al., 2004), which determine each chunk's activation at a given point in time. Suppose that a noun phrase, say *the robber* in (2), has been stored in memory as memory chunk k . When a retrieval is triggered while processing word n (*escaped*) later on, chunk k 's activation value at word n is calculated as

$$A_{k,n}(t) = S_k(t) + P_k(t) + B_k(t) , \quad (4.1)$$

where S_k is the memory association strength, P_k is the mismatch penalty, and B_k is the chunk-specific base-level activation. The fan effects $\phi_{kl}(t)$ of competing retrieval candidates of all l features of memory chunk k decrease the chunk's activation strength, which also depends on the S_{\max} (*maximum activation strength*) parameter, i.e.,

$$S_k(t) = \sum_l [S_{\max} - \log \phi_{kl}(t)] . \quad (4.2)$$

The fan effect variable $\phi_{kl}(t)$ is defined as the number of memory chunks with feature l at time t , including memory chunk k itself so that $\phi_{kl}(t) \geq 1$.

The mismatch penalty decreases activation for all retrieval cues l that do not match the cor-

responding feature of memory chunk k , i.e.,

$$P_k(t) = \sum_l \Delta_{kl}, \quad (4.3)$$

where

$$\Delta_{kl} := \begin{cases} 0 & \text{if } \text{cue}_l = \text{feature}_{kl} \\ -p & \text{otherwise} \end{cases} \quad (4.4)$$

and $p \geq 0$ is a free parameter specifying the mismatch penalty incurred by each unmatched feature.

Chunks become active when words are encoded or when retrievals are performed, and then start to decay. The resulting base-level activation at time t is given by

$$B_k(t) = \sum_i \exp(-d \cdot \lfloor t - t_{ik} \rfloor) \quad (4.5)$$

where d is a decay parameter and t_{ik} is the i -th memory access (encoding or retrieval) of memory chunk k .

Note that in our implementation, in contrast to the original LV05 model, S_k , ϕ_{kl} , and P_k are functions of time. This is because the memory schedule, that is, the set of words encoded in memory chunks, changes dynamically each time a word is encoded in memory. As encodings can happen at any time t , the memory schedule, and therefore the predicted fan effects and penalties, may change even while a retrieval is ongoing. This assumption is necessary to allow for dependency resolution in the case that a retrieval trigger is processed before a potential target has been stored in memory.

Activation values are subject to stochastic noise controlled by the *ans* (*activation noise*) parameter, so that

$$A'_{k,n}(t) \sim \text{Logistic}(A_{k,n}(t), \text{ans}) . \quad (4.6)$$

The memory chunk k_n^* with the highest memory activation $A'_{k,n}$ is matched for the retrieval n , and the retrieval latency is computed as

$$t_{k,n} = F \cdot \exp[-A'_{k,n}(t)] , \quad (4.7)$$

where F is the *latency factor*, a free linear scaling parameter.

Equation (4.7) can be used to make quantitative predictions for reading times, and the LV05 model has been used to model a variety of phenomena in the sentence-processing literature (for a review, see Engelmann et al., 2019; Vasishth et al., 2019). However, the LV05 model can only be straightforwardly applied to paradigms in which sentences are read strictly incrementally, such as self-paced reading: The model can create chunks, track their

activations, and integrate them with each other via retrievals, but it does not account for eye fixations, and cannot capture cases in which the order of fixations mismatches the serial word order due to skipings and regressions. To fully capture “natural” sentence reading, the LV05 model thus needs to be interactively integrated with a model that accounts for spatial and temporal aspects of eye movements.

The dynamical SWIFT model (Engbert et al., 2002, 2005) is a good candidate for integration with the LV05 model. Its main advantages are that it

- has recently been implemented for Bayesian parameter inference (Rabe et al., 2021; Seelig et al., 2020),
- predicts and explains all empirically observable saccades in sentence reading, and
- allows for (but does not enforce) parallel processing of words.

Even though SWIFT itself does not follow an ACT-R based architecture like EMMA (Engelmann et al., 2013; Salvucci, 2001; Vasishth et al., 2019), an integration with ACT-R-based models such as LV05 is possible via activation-based coupling, as we will detail below after a brief introduction of SWIFT.

4.1.2 The SWIFT Model of Eye-Movement Control (Engbert et al., 2005)

SWIFT is a model of eye-movement control in reading implemented in a dynamical cognitive modeling framework (Beer, 2000; Engbert, 2021). At its core, its internal timing processes and word activations govern the temporal control and target selection for saccadic eye movements. Words with high activation values are more likely to be selected as saccade targets. SWIFT assumes that all words that fall within a *processing span* around the current fixation location are processed in parallel (Engbert et al., 2002).¹⁸ The processing rate $\Lambda_j(t)$ of any given word j at time t depends on a number of factors such as gaze eccentricity, that is, the distance between word j and the currently fixated word, such that words that are further away from the visual focus are processed more slowly.

In SWIFT, each word in the sentence passes through a *lexical* and *post-lexical* processing stage. During lexical processing, word recognition and identification take place. As word recognition is ongoing, the discrete activation associated with the processed word j , $n_j(t)$, rises up to a maximum threshold, N_j . The threshold is modulated by the word’s corpus frequency, as frequent words generally require less processing than less frequent words, and word predictability. Note, however, that we did not include predictability effects in our

¹⁸Other examples of parallel processing models include *Glenmore* (Reilly & Radach, 2006) and *OBI-Reader* (Snell et al., 2018). These models contrast with sequential attention shift models such as *E-Z Reader* (Reichle et al., 1998).

model implementation. SWIFT also largely ignores low-level sensory perception and letter-level processing, which can have effects on the further (post-lexical) processing of a word and the sentence as a whole. In future work, processes such as bigram identification (Snell et al., 2018) and surprisal (Huang et al., 2023) are worth considering as extensions to SWIFT (or derivative models) to account for more aspects of lexical processing.

Once the word is identified, post-lexical processing begins and word activation decreases again. Post-lexical processing, however, is not explicitly modeled in SWIFT. Although SWIFT keeps track of the processing stage of words in the sentence, it has no higher-level representation of its constituents or of the entire word sequence. Adjacent words may have an influence on processing difficulty, but there is no mechanism to account for difficulty due to dependency completion processes at the sentence level.

While the relative word activations at the time of programming a saccade determine the relative probability of each word to be selected as the upcoming target, the timing of saccades is relatively independent (Findlay & Walker, 1999) and involves a cascade of several processes. The cascade starts with a global timer, which triggers the *labile* and subsequent *non-labile* saccade stages, a distinction motivated by oculomotor performance in the double-step paradigm (Becker & Jürgens, 1979). During the labile stage, saccades can be canceled and a new target can be selected. During the non-labile stage, cancellation is no longer possible. The execution of the saccade itself is a noisy process subject to systematic (range) and random error (McConkie et al., 1988), where the systematic error component can be explained by a Bayesian-optimal estimation of the saccade target position (Engbert & Krügel, 2010). for saccade amplitudes based on significantly better model fits in previous work (Rabe et al., 2021).

Target selection in SWIFT is inherently stochastic, as it depends on the dynamic, relative word activations at any given point in time. Words with high activation values are more likely to be selected as targets than words with lower activation. The probability $\pi_j(t)$ to select word j at time t as the next saccade target is given as

$$\pi_j(t) = \frac{[a_j(t)]^\gamma}{\sum_{k=1}^{N_W} [a_k(t)]^\gamma} \quad (4.8)$$

where N_W is the number of words in the sentence and

$$a_j(t) = \frac{n_j(t)}{N_a} \quad (4.9)$$

is the normalized activation of word j at time t , which is the processing state of the word, normalized by parameter N_a , the highest possible threshold of a word in a given corpus.

The relation between the activation $a_j(t)$ of a word and its selection probability $\pi_j(t)$ also entails that words requiring little processing (i.e., “easy-to-process” words)

pass through lexical and post-lexical processing faster than less frequent (i.e., “difficult-to-process”) words. The former words are therefore in a state of higher activation for a shorter time period, consequently less likely to be fixated, and thus often skipped. The free parameter γ modulates the relationship between word activations and selection probabilities. For $\gamma \rightarrow 0$, words are selected randomly with equal probability, regardless of their actual activation values (if greater than zero). If $\gamma \rightarrow 1$, there is a perfect linear relationship between activations and selection probabilities (Luce’s choice rule). Higher values $\gamma \rightarrow \infty$ enforce a winner-takes-all principle so that the word with the highest activation always “wins.”

The evolution of word activations in the original version of SWIFT (Engbert et al., 2002, 2005) was governed by ordinary differential equations (ODEs). In the more recent versions by Rabe et al. (2021) and Seelig et al. (2020), the dynamics of SWIFT changed toward a model with discrete internal states that evolve stochastically over continuous time. Word activations and saccade timers are random walks that increase/decrease over time with different transition rates for different timers and individual word activations. The state of the model at time t is given by a vector $n = (n_1, n_2, \dots, n_{4+N_W})$, where the components n_j represent the states of the subprocesses. Components 1 to N_W are keeping track of the (post-)lexical processing of words, while components $N_W + 1$ to $N_W + 4$ are saccade-related and additional stochastic variables (Table 4.1). In each of the possible transitions from state $n = (n_1, n_2, \dots)$ to $n' = (n'_1, n'_2, \dots)$ only one of the sub-processes n_i is changed by one unit. The discrete stochastic variables $\{n_j\}$ at time t map to the activation variables $\{a_j(t)\}$.

For the numerical simulation of the model, an algorithm can be derived from the master equation (see Seelig et al., 2020, for details),

$$\frac{\partial}{\partial t} p(n, t | n'') = \sum_{n'} [W_{nn'} p(n', t | n'') - W_{n'n} p(n, t | n'')] , \quad (4.10)$$

which describes the temporal evolution of the model’s internal states (Gardiner, 1985; Van Kampen, 1992). It is specified by the transition rates $W_{n'n}$, which in turn govern the transitions between state vectors $n \mapsto n'$.

Table 4.1

Stochastic Transitions Between Internal States From $n = (n_1, n_2, \dots) \mapsto n' = (n'_1, n'_2, \dots)$

| Process | Transition to ... | Transition rate $W_{n'n}$ |
|--------------------|------------------------------|---|
| Word processing | $n'_j = n_j \pm 1$ | see Equation (4.11) for w_{j, \dots, N_W} |
| Saccade timer | $n'_{N_W+1} = n_{N_W+1} + 1$ | $w_{N_W+1} = N_t/t_{\text{sac}} \cdot (1 + h a_k(t)/\alpha)^{-1}$ |
| Labile program | $n'_{N_W+2} = n_{N_W+2} + 1$ | $w_{N_W+2} = N_l/\tau_l$ |
| Non-labile program | $n'_{N_W+3} = n_{N_W+3} + 1$ | $w_{N_W+3} = N_n/\tau_n$ |
| Saccade execution | $n'_{N_W+4} = n_{N_W+4} + 1$ | $w_{N_W+4} = N_x/\tau_x$ |

Implementation of more detailed assumptions on the post-lexical stage can be achieved by changing the transitions rates $\{w_j(t)\}$ that control the stochastic transitions for the internal states $\{n_j(t)\}$ and thus for activations $\{a_j(t)\}$. Transition rates are a measure of the expected number of transitions in a given time unit (milliseconds in SWIFT) and are the inverse of the expected time between two consecutive transitions. Transition rates, in combination with thresholds N_j , are therefore directly related to processing speed. While the rates for the saccade timers are either constant or determined according to an invariant rule (see Table 4.1), the determination of transition rates for word processing components varies between processing stages, i.e.,

$$w_j(t) = \begin{cases} \alpha \cdot \Lambda_j(t) & \text{in lexical stage} \\ \max[\alpha \cdot \Lambda_j(t) \cdot \text{proc}, \omega] & \text{in post-lexical stage} \\ 0 & \text{otherwise (complete)} \end{cases}, \quad (4.11)$$

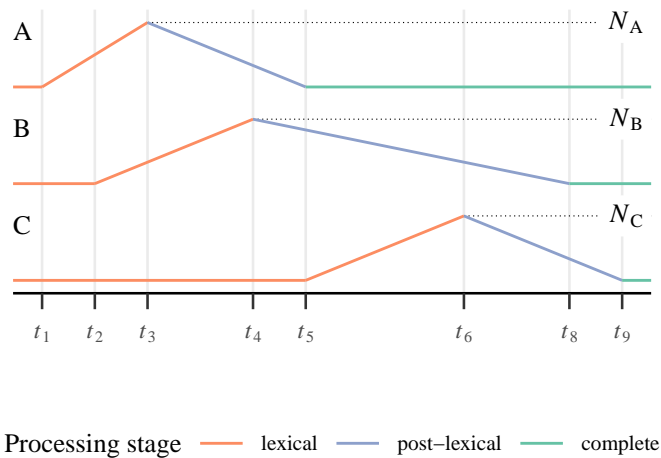
where α is the baseline processing difficulty, Λ is the processing rate, proc is the relative processing speed for post-lexical processing, and ω is a minimum decay parameter.¹⁹ In the integrated SEAM model, word activations in SWIFT are coupled with memory activations in LV05 in a Bayesian modeling framework by adapting the formula in Equation (4.11).

The fact that the SWIFT implements detailed mechanisms on word processing and saccade preparation is reflected by the number of parameters. Fitting the eye-movement model to experimental data started with hand-picking plausible parameter values, grid search (Reichle et al., 1998), genetic algorithms (Engbert et al., 2002), while optimizing the fit between empirical and simulated summary statistics. Based on the development of a likelihood approximation (Seelig et al., 2020), a fully Bayesian framework is now available for parameter inference (Rabe et al., 2021). The likelihood framework permits objective parameter fitting independent of a set of selected summary statistics, since fixation sequences are involved for likelihood computation. Using large-scale numerical simulations, it has been shown that SWIFT can reliably reproduce fixation durations, fixation probabilities and saccade amplitudes at the level of global and by-participant summary statistics, without using those summary statistics for the purpose of parameter fitting.

4.1.3 SEAM: Activation-Based Coupling of SWIFT and LV05

In baseline SWIFT, processing a word always starts out in the lexical processing stage. Once the word activation $n_j(t)$ has reached its threshold N_j at time t , it begins post-lexical processing, and activation starts to decrease. When the activation has returned to zero, the

¹⁹The transition rate for post-lexical word j cannot be lower than ω , which ensures a decaying word activation even if there is no or little processing at a given time t , e.g. when the word is not within the processing span.

Figure 4.1*Word Activation in SWIFT*

Note. Theoretical activation history of three words (A, B, and C). Colors of line segments correspond to the processing stage active at that given time. Activation maxima are N_A , N_B , and N_C , respectively. Activations are displayed as continuous but are actually implemented as discrete counters.

word is completely processed.

Figure 4.1 abstractly shows the activation histories of three hypothetical words. The figure assumes that the eyes move sequentially from word (a), to (b), to (c), leading to a somewhat sequential onset of their first processing (t_1 , t_2 , and t_5). The first stage of processing is the lexical stage. During this stage, activations rise until they reach their respective maxima (N_A , N_B , and N_C), which depend on printed word frequency. Given that saccade targeting depends on activation, the words in question are most likely to be selected as a saccade target if the upcoming saccade is programmed at times t_3 , t_4 , and t_6 . This happens as well when the words enter the post-lexical processing stage. During post-lexical processing, activations decrease again, making it in turn less likely for the respective word to be selected as a target. Once the activation returns to zero (t_5 , t_8 , and t_9), the word is assumed to have completed processing.

A feature common to the SWIFT and LV05 is that both models use activation values to guide processing. SWIFT uses word activations to select words as saccade targets, while LV05 uses memory activations to select memory chunks as retrieval targets. Our integrated model SEAM keeps these activations separate, but implements an interaction, so that memory activations in LV05 modulate word activations in the SWIFT model. Therefore, rather than assuming that the sentence processor has direct control of the eye-movement targeting system, we propose an indirect, stochastic influence on saccade targeting via memory activations. This is in good agreement with eye-tracking studies carried out with larger-than-usual

sample sizes that show that the effects of sentence processing cost due to memory interference on fixation and other measures have relatively small magnitudes (e.g., Jäger et al., 2020); larger effect sizes are generally driven by lower-level factors such as frequency and word length (Boston et al., 2008).

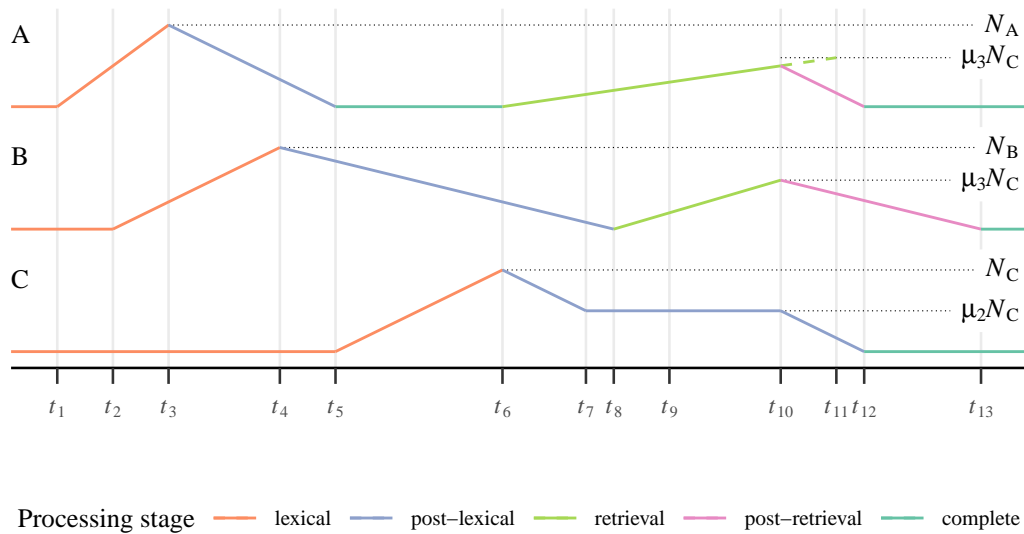
In SEAM, activations in the LV05 component reflect the construction of a sentence representation, which affect word activations and thereby stochastically influences target selection in the eye-movement component. As in SWIFT, the activation gradient of a word in SEAM is mainly determined by the transition rate, which varies between processing stages. Compared to SWIFT, the sequence of processing stages in SEAM is extended by stages that reflect the cost of memory retrieval, which can account for additional processing difficulty. Possible interactions of memory retrieval and the word activations during dependency resolution include: (a) the retrieval process delays post-lexical processing of the the currently fixated region that caused the retrieval (that is, the retrieval trigger); and (b) retrieval candidates are reactivated so that they attract regressions from the retrieval trigger.

In Figure 4.2, activation histories of the same three words from the SWIFT example in Figure 4.1 are shown. Like the baseline SWIFT model, words in SEAM go through a lexical and post-lexical processing stage before they are considered *completely processed*. However, SEAM additionally accounts for the resolution of a linguistic dependency during post-lexical processing of word C. Once the words are lexically accessed (t_3 , t_4 , and t_6), they are encoded as chunks in SEAM's memory module, along with their features, as in the LV05 model. Words A and B are assumed to not trigger a dependency completion process; this is the case for most nouns. However, when word C, which could be a verb, is processed and the associated chunk is stored in memory, a subject-verb dependency must be resolved. A retrieval is thus triggered. The assumption that nouns do not trigger a dependency completion process is obviously an oversimplification; but this simplification is reasonable for the data being modeled in this paper, as in the experiment design of Mertzen et al. (2023), the theoretically interesting dependency completion occurs at the verb.

During retrieval, all words that are fully processed before the processing of word C completes are counted as retrieval candidates. Candidate words enter into a retrieval stage in which activation increases until the retrieval process finishes.²⁰ The activation increase differs by the degree to which the retrieval candidate features match the retrieval cues, implementing a core assumption of the LV05 model.

The effect of the memory activations on word activations is mainly modulated by the new parameters μ_2 and μ_3 . The retrieval stage ends when one candidate reaches a threshold

²⁰A word can also become a candidate after the retrieval process has started. Word A, for example, is already a candidate at the time the post-lexical processing of word C starts at time t_6 , given that it was already completely processed at time t_5 . Therefore, the retrieval stage of word A starts immediately with the start of the post-lexical stage of word C. This mechanism is necessary for the rare but possible case that a retrieval is triggered before any candidate word has been encoded as a chunk in memory.

Figure 4.2*Word Activation in SEAM*

Note. Theoretical activation history of three words (A, B, and C). Colors of line segments correspond to the processing stage active at that given time. Activation maxima are N_A , N_B , and N_C , respectively, for the transition from lexical to post-lexical processing, $\mu_2 N_C$. Activations are displayed as continuous but are actually implemented as discrete counters.

value, which is a fraction μ_3 of the maximum activation of the retrieval trigger N_C . Because post-lexical processing in SEAM is only finished after all dependencies have been resolved, the post-lexical activation of the retrieval trigger is guaranteed not to fall below a fraction μ_2 of its maximum activation during retrieval. This is why the post-lexical activation of word C does not change between t_7 and t_{10} . In this example, despite entering the retrieval phase at a later time, word B reaches the retrieval threshold at time t_{10} before word A, thereby concluding the retrieval process. Consequently, the post-lexical processing of word C continues and all retrieval candidates, that is, word A and word B, enter a post-retrieval stage, which is equivalent to an additional post-lexical processing stage. This also entails that the retrieval phase of word A is aborted, which would otherwise have reached threshold at time t_{11} .

The transition rates of the baseline SWIFT model for word j , Equation (4.11), are re-

placed by

$$w'_j(t) = \begin{cases} \alpha \cdot \Lambda_j(t) & \text{in lexical stage} \\ \max[\alpha \cdot \Lambda_j(t) \cdot \text{proc}, \omega] & \text{as retrieval trigger } (j = m \wedge n_j(t) > \mu_2 N_j) \\ 0 & \text{as retrieval trigger } (j = m \wedge n_j(t) \leq \mu_2 N_j) \\ \max[\alpha \cdot \Lambda_j(t) \cdot \text{proc}, \omega] & \text{in post-lexical stage} \\ \mu_3 N_m F^{-1} \exp[A'_{j,m}(t)] & \text{as retrieval candidate } (j \neq m) \\ \max[\alpha \cdot \Lambda_j(t) \cdot \text{proc}, \omega] & \text{in post-retrieval stage} \\ 0 & \text{otherwise (complete)} \end{cases} \quad (4.12)$$

where m is the current retrieval trigger that needs to form a dependency. The transition rate for the retrieval candidate j , triggered by dependency resolution for word m , is chosen to ensure that the total duration of reaching threshold (i.e., the time for j to be matched as a dependent of m), matches the retrieval latency predicted by LV05. Therefore, it is computed as the threshold value $\mu_3 N_m$ divided by the expected total duration of j in that stage, $F \exp[-A'_{j,m}(t)]$.

Altogether, SEAM extends the baseline SWIFT model parameters (Rabe et al., 2021; Seelig et al., 2020) with seven additional model parameters. The parameters d (decay), S_{\max} (maximum memory activation strength), F (retrieval latency scaling factor) and p (mismatch penalty), which modulate $w'(t)$ through $A'_{j,m}(t)$, are directly based off their LV05 implementations (Lewis & Vasishth, 2005). Moreover, the link between word activations in LV05 and processing rate in SWIFT is complemented by the three new model parameters μ_1 , μ_2 , and μ_3 , as detailed above. Some parameters of the LV05 model, in particular for goal activation and noise (G , and ans), are ignored in the present implementation. Variation in the goal activation parameter is usually used to model individual-level capacity differences (e.g., Daily et al., 2001; Mätzig et al., 2018; Vasishth & Engelmann, 2022), which is not of interest in the present work. The goal activation is fixed at 1.0, which gives equal weight to all retrieval cues. The noise parameter ans is replaced by the built-in stochasticity of SWIFT. Moreover, the parameters S_{\max} and F are not independent in terms of the resulting retrieval latency and transition rate, which is why we will only estimate F as a free parameter and keep S_{\max} at a fixed default value of 1.5. In the present study, we also exclude μ_1 , the fixed time needed to execute a production rule, by setting it to 0, because we assume this time to overlap with some of the oculomotor processes already present in the model. Since S_{\max} is fixed, we also decided to fix mismatch penalty p at its default value, as the relation of the two parameters is critical. Thus, the only parameters that were fit to the Mertzen et al. (2023) data were F , d , μ_2 , and μ_3 . For a complete list of model parameters and default values in SEAM, see Appendix D.

For our implementation of SEAM, we opted for a simplified version of the LV05 model (Engelmann, 2015) and the latest version of SWIFT (Rabe et al., 2021).²¹ SEAM connects the baseline eye-movement control architecture of SWIFT with the interactive working memory module of LV05 via activation-based coupling: reading words in SWIFT leads to the creation of memory chunks and can trigger retrievals in LV05, whereas chunk activations computed by LV05 modulate word activations in SWIFT.

4.2 Data Availability

All experimental and simulated data, analysis code, and computational models (SEAM and SWIFT) reported in this paper are available at the Open Science Framework (<https://doi.org/10.17605/osf.io/ad5nx>) and at GitHub (<https://github.com/mmrabe/SEAM-2023-Paper>).

4.3 Experimental Study (Mertzen et al., 2023)

To test the predictions of the integrated model, we use data from a memory interference experiment conducted with 61 English native speakers (Mertzen et al., 2023). This experiment was originally planned with 120 participants, but due to the pandemic, data collection had to be aborted. Our inability to reach the target number of participants has consequences for model evaluation, as discussed later.

The Mertzen et al. (2023) experiment employed a fully crossed distractor subjecthood (2) × animacy (2) design that closely mirrored an experiment reported in Van Dyke (2007). Examples of the four conditions are shown below in example (3).

- (3) a. It turned out that the **attorney**_{+anim}^{+subj} whose secretary had forgotten about the important meeting_{-anim}^{-subj} frequently **complained**_{anim}^{subj} about the salary at the firm.
- b. It turned out that the **attorney**_{+anim}^{+subj} whose secretary had forgotten about the important visitor_{+anim}^{-subj} frequently **complained**_{anim}^{subj} about the salary at the firm.
- c. It turned out that the **attorney**_{+anim}^{+subj} whose secretary had forgotten that the meeting_{-anim}^{+subj} was important frequently **complained**_{anim}^{subj} about the salary at the firm.
- d. It turned out that the **attorney**_{+anim}^{+subj} whose secretary had forgotten that the visitor_{+anim}^{+subj} was important frequently **complained**_{anim}^{subj} about the salary at the firm.

²¹The principal reason for using the simplified version of the LV05 model is tractability. Using the full ACT-R architecture, which is Lisp-based, would require much more complex engineering decisions, and would make the model inaccessible to researchers unfamiliar with Lisp but who are interested in exploring its behavior with novel data.

In the example above, processing the verb *complained* is expected to trigger a retrieval for an animate subject noun phrase. In all sentences, *attorney* is the grammatically correct subject of *complained*, and should thus be retrieved. However, the distractor noun phrase (*meeting* or *visitor*) may interfere with the retrieval of *attorney*. The distractor is *visitor* in the +animate or *meeting* in the –animate condition, and it is either a subject (+subject) or an object (–subject) of the embedded clause.

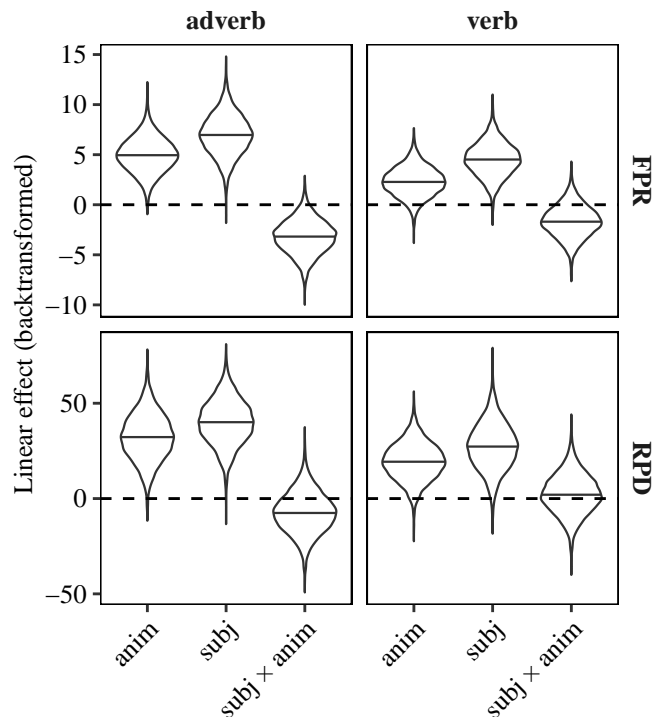
According to cue-based retrieval theory, both subjecthood and animacy of the distractor should lead to additional difficulty for resolving the critical dependency. This is due to the fan effect (e.g., Anderson, 1990), which is also known as similarity-based interference (Jäger et al., 2017): When the feature specification of a distractor overlaps with that of the retrieval target, it diverts some of the retrieval activation from the target to itself. The activation of both the target and distractor are reduced, leading to longer retrieval time; what ends up being retrieved in a particular simulation run (target or distractor) depends on which chunk happens to have higher activation (this can vary in simulation runs due to stochastic noise in the activation). It is therefore possible that the distractor is sometimes erroneously retrieved. As indices of increased processing difficulty, we expect additive effects of animacy and subjecthood of the distractor on regression path duration and outgoing regression probabilities on the critical verb (*complained*). The primary region of interest where the effect of the subjecthood and animacy manipulation should manifest is the verb; however, because similarity-based interference effects have been shown to occur in the region just before the verb (Lago et al., 2021; Van Dyke, 2007), Mertzen et al. (2023) also investigated the effect at the adverb (*frequently*) that preceded the critical verb. For this reason, in our investigations we also report model fits for this pre-critical region.

In summary, similarity-based interference accounts predict that conditions (3b,d) should be more difficult to process than conditions (3a,c) due to the animacy of *visitor*, and conditions (3c,d) should be more difficult to process than conditions (3a,b) due to the distractor being in subject position.

As indices of increased processing difficulty, additive effects of distractor animacy and distractor subjecthood were expected in reading times and outgoing regression probabilities. An interaction of distractor subjecthood and animacy was not predicted but is reported in Mertzen et al. (2023) for completeness; in the Mertzen et al. (2023) analysis, there was no evidence for an interaction.

In this summary of the Mertzen et al. (2023) results, we report only regression path duration and first-pass regressions out (FPR) from the pre-critical adverb and the critical verb; for full details of all experimental results, please see the original paper.

The effects of animacy and subjecthood (coded as sum contrasts) were analyzed using Bayesian mixed-effects models. Subject and item were specified as random effects in the models, with a full variance-covariance matrix for subject and item random effects. The

Figure 4.3*Experimental Effects of Mertzen et al. (2023)*

Note. Plotted violins are the estimated posterior distributions of experimental effects of subjecthood (*subj*) and animacy (*anim*) on regression path duration (RPD) and first-pass outgoing regression probability (FPR) from Bayesian mixed-effects regressions, as analyzed and reported by Mertzen et al. (2023). Posteriors are backtransformed linear effects in ms (for RPD) or % (for FPR).

models were implemented with *brms* (Bürkner, 2017, 2018, 2021), an interface to *Stan* (Carpenter et al., 2017). Priors were mildly informative Gaussian distributions for the linear model coefficients (intercept and slopes) and mildly informative regularizing Lewandowski-Kurowicka-Joe (LKJ) priors (Lewandowski et al., 2009) for random effects correlation matrices; setting the LKJ prior's parameter ν to 2 downweights extreme correlations like ± 1 . For a detailed tutorial on linear mixed models in the Bayesian setting, see chapter 5 of Nicenboim et al. (2023), or Sorensen et al. (2016).

The results in Mertzen et al. (2023) showed reading time patterns consistent with effects of subjecthood (syntactic interference) and effects of animacy (semantic interference). Figure 4.3 shows that on the pre-critical adverb, the effect of subjecthood shows longer regression path duration (RPD) and more first-pass regressions out for conditions that have a +subject distractor (95% credible intervals (CrIs): RPD [17,63] ms, FPR [3,11]%). Similarly, the effect of animacy shows longer regression-path duration and an increase in first-pass regressions out for conditions with animate distractors compared to conditions with

inanimate distractors (95% CrIs: RPD [8,57] ms, FPR [2,8]%). The subjecthood \times animacy interaction in regression-path duration is centered on zero; for first-pass regressions, the interaction has a negative sign ($[-7,0]$ %).

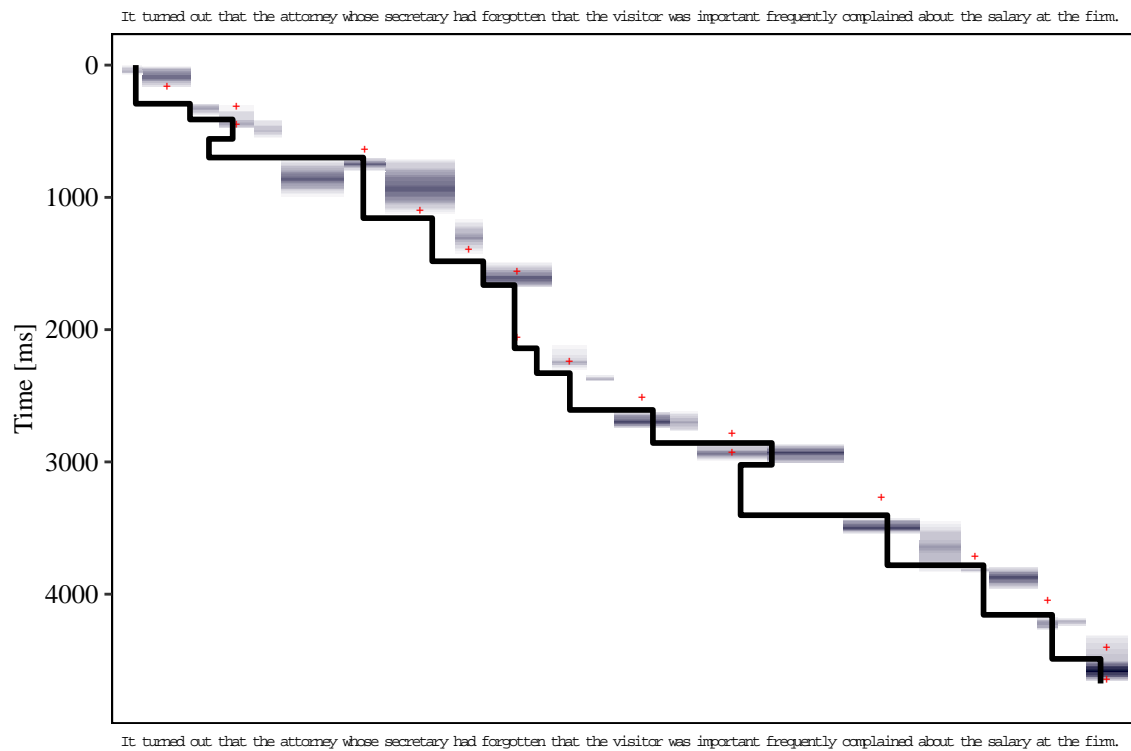
On the critical verb, the effects of subjecthood and animacy show a similar pattern of longer regression path duration and an increase in first-pass regressions out (Subjecthood 95% CrIs: RPD [3,52] ms, FPR [1,8]%; Animacy 95% CrIs: RPD [0,39] ms, FPR $[-1,5]$ %). The interaction is centered around zero for regression path duration and regressions out. The increased reading times and regressions for conditions that have subject or animate distractors indicate that syntactically and semantically similar distractors can interfere during long-distance dependency formation.

4.4 Simulation Study

The reliability of computational cognitive models critically depends on the availability of appropriate methods for statistical inference (Engbert et al., 2022; Schütt et al., 2017). We previously applied a broader principled Bayesian workflow (Schad et al., 2021) for the baseline SWIFT model in Rabe et al. (2021), which is used as the eye-movement platform in SEAM.

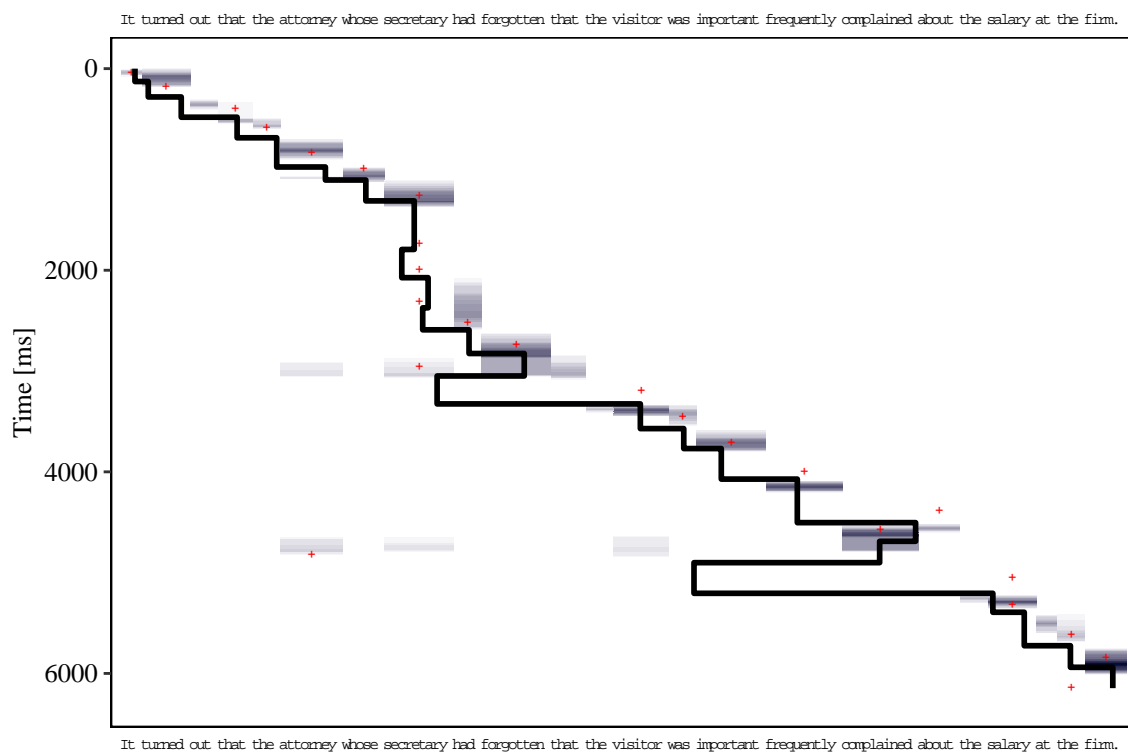
In Figures 4.4 and 4.5, we visualized the word activation field and eye trajectory for a simulated trial in SWIFT and SEAM, respectively. As can be seen, SEAM behaves similarly to SWIFT throughout most of the trial. However, the models' behaviors start to diverge when the verb *complained* is processed and triggers a retrieval in SEAM. During the retrieval phase, word activations of previous words that have been encoded as memory chunks increase. Words with better cue match for the retrieval approach the activation threshold faster than those with lower cue match. If a saccade is triggered during the retrieval phase, the reactivated words can attract regressions.

Without proper checks, it is not self-evident that Bayesian model fitting of SEAM can be carried out in the same way as for SWIFT. However, we expect that our implementation of SEAM will exhibit correct inference because it meets the following three critical conditions: First, for all observables that were taken into account (i.e., fixation positions and durations), a model likelihood has already been implemented in SWIFT (Seelig et al., 2020). Secondly, both SWIFT and LV05 are dynamic in the sense that they describe activation values as a function of time, which allows us to let them interact dynamically without a significant modification of their initial conceptualization. Thirdly, the dynamics of eye movements and sentence processing interact in the integrated SEAM model and will thus affect the observable temporal and spatial aspects of fixation sequences due to the activation coupling of the constituent SWIFT and LV05 components. The coupling via word activations permits indirect fitting of model parameters related to memory retrieval, as long as they have some

Figure 4.4*Example Simulation in SWIFT*

Note. SWIFT simulation for Example (1). The bold black line is the simulated fixation location (x-axis) as a function of time (y-axis). Saccades are horizontal displacements of the black line. Word activations are depicted by gradients in the background, with darker shades referring to higher activation. The target selection preceding each executed saccade is depicted by a red cross, marking both the time and intended saccade target. Target selection is based on the relative word activations at the respective time point of saccade programming. Saccade timers, which are also components of the internal states, are omitted for brevity. For more details, see Rabe et al. (2021) and Seelig et al. (2020).

Figure 4.5
Example Simulation in SEAM



Note. SEAM simulation for Example (1). As in Figure 4.4, the black line is the simulated fixation location (x-axis) as a function of time (y-axis), gradients in the background are word activations, and red crosses are selected targets. Note that, in comparison to Figure 4.4, processing of *forgotten* and *complained* triggers retrievals, which prolongs processing of the trigger and reactivates potential retrieval candidates. In this simulation, during both retrievals, regressive saccades are programmed and executed.

probabilistic effect on the outcome variables captured by SWIFT.

Given these properties, we tested the computational faithfulness of SEAM using the Markov Chain Monte Carlo (MCMC) sampling algorithm DREAM_{ZS} (Laloy & Vrugt, 2012) based on profile log-likelihoods and model parameter recovery, similar to the approach taken in Rabe et al. (2021). The DREAM_{ZS} (Laloy & Vrugt, 2012; ter Braak & Vrugt, 2008; Vrugt et al., 2009) sampler has previously been successfully used with complex dynamical models of eye-movement control, including SWIFT for reading (Rabe et al., 2021) and SceneWalk for scene viewing (Schwetlick et al., 2020, 2022).

After confirming the computational faithfulness of the model, we fitted the model to a training subset of the experimental data and compared predictions for a withheld test portion using relevant global summary statistics and the predicted experimental effects of similarity-based interference²² described in the previous section.

4.4.1 Method

Data assimilation. In eye-movement research, the experimental (observed) data are fixation sequences consisting of time-ordered sequential observations. In such a case, the identification of model parameters is possible within the field of *data assimilation* (Engbert et al., 2022; Reich & Cotter, 2015). Data assimilation refers to the integration of complex mathematical models with time-series data (see Morzfeld & Reich, 2018, for an introduction). In this framework, the SWIFT model has previously been implemented for Bayesian model fitting (Seelig et al., 2020). Rabe et al. (2021) showed that, in a principled Bayesian workflow (Schad et al., 2021), SWIFT can be reliably fitted to simulated and experimental data even with many free parameters and sparse data that resulted from splitting by participant and experimental condition.

Sequential likelihood. The time-ordered nature of fixational eye movements make them a suitable target for data assimilation (Engbert et al., 2002). To exploit the sequential information of the data, some of those models use *sequential likelihoods* for parameters $\theta \in \Theta$ such that

$$L_M(\theta | X_n) = P_M(x_1 | \theta) \prod_{i=2}^n P_M(x_i | X_{i-1}, \theta), \quad (4.13)$$

where $X_n = (x_1, \dots, x_n)$ is the entire sequence of n events and $P_M(x_i | X_{i-1}, \theta)$ is the likelihood of the i -th event of the sequence given all previous events $X_{i-1} = (x_1, \dots, x_{i-1})$.

Successful examples of applying data assimilation for visual tasks are, for example, Sce-

²²We will focus on effects of subjecthood and animacy cues, since those were of main interest in the experimental study. Additional features/cues such as clause locality were also encoded but are not of primary interest here. The full memory and retrieval schedules are available in the model supplement at <https://github.com/mmrabe/SEAM-2023-Paper/tree/main/SEAM/DATA>.

neWalk (Schwetlick et al., 2020, 2022) for scene viewing and SWIFT (Rabe et al., 2021; Seelig et al., 2020) for reading. There, each event of the sequence, x_i , is a fixation. Since the location and temporal onset of the first fixation are typically known due to the experimental paradigm, e.g., sequences always starting at a fixation cross, the likelihood for x_1 is given by $P_M(x_1 | \theta) = 1$. SceneWalk and SWIFT further decompose the likelihood into spatial and temporal components, since each fixation has a spatial location on the screen and a duration.

As SEAM is based on SWIFT and we only changed the latent transition rates rather than the saccade execution itself, we can easily use the data assimilation methods implemented for SWIFT. This is especially useful because we fit the model on a by-participant basis and hence only have little data for parameter estimation. The decomposition of temporal and spatial likelihood components is also theoretically interesting since we can expect the modification of the transition rates to affect both the temporal control and target selection of the (simulated) saccadic eye movements.

Profile likelihoods. As SEAM modifies model dynamics and thus the likelihood function of SWIFT, a reevaluation of the *profile log-likelihoods* is crucial. Those are generated by first simulating data with known parameter values, and then systematically varying parameter values and inspecting the likelihood of the data for each value. Ideally, the likelihood of the data should be highest for the true parameter values. In order to assess whether the modifications introduced in SEAM are appropriately captured in its likelihood, it should be ensured that the newly introduced free parameters affect the outcome likelihood. Thus, the behavior of the likelihood as a function of each of the new parameters represents a necessary condition for identifiability and statistical inference of the full model (Rabe et al., 2021; Seelig et al., 2020).

Parameters were inspected if they were going to be fitted later on and/or were added in this model implementation compared to the reference SWIFT implementation (Rabe et al., 2021). This was the case for a total of 11 parameters (see Figure 4.6). Parameters μ_1 and S_{\max} were also inspected even though they were not selected to be fitted to the recovery and experimental data. This is because the parameters themselves are identifiable, as can be seen in Figure 4.6, but they are not independent from other model parameters in terms of an effect on model behavior. All other shown model parameters are also fitted to simulated data for parameter recovery as well as to experimental data.

Parameter estimation and recovery. As a last step for the verification of the computational faithfulness of the approach, we applied a sampling algorithm to simulated data with known true parameter values in order to ensure the validity of the computational approach. We generated 100 unique data sets with different sets of true parameters θ^* randomly sampled from the prior distribution later used for parameter estimation. Parameters would be considered successfully recovered if the correlation between true and recovered parameters was sufficiently high and the normalized root mean squared error (NRMSE) was sufficiently

low.

Summary statistics and experimental effects. Even though we are using an objective likelihood-based approach for model fitting, it is important that simulated and empirical data are in good agreement at the level of relevant summary statistics, especially with regard to comparability with competitor models and theory testing (S. Roberts & Pashler, 2000). Because the goal for SEAM is to explain both spatial and temporal aspects of eye movements in reading, we consider a number of different spatial and temporal summary statistics frequently used in reading research. For the spatial dimension, we are looking at several fixation probabilities, that is, probabilities to fixate (or skip) specific words under different conditions. For the quantification of the temporal aspects of the model fit, we evaluate different fixation durations, that is, average reading times under different conditions.

A subset of the experimental test data set is withheld from parameter estimation, and this held-out set will then be compared on the basis of summary statistics against predicted data from SEAM and SWIFT using estimated parameters. Specifically, we first split the experimental data into a training and test subset, fitting the model to 70% of the data (training set) of each participant and condition, subsequently predicting eye trajectories for the other 30% (test set). For each withheld trial, we generated a fixation sequence using the HPDI (highest posterior density interval) midpoint of the sampled posterior distribution of a given participant and parameter (Rabe et al., 2021). We also present the predictions of SEAM and SWIFT for the experimental memory interference effects, which can be similarly derived from the simulated and experimental data alike.

4.4.2 Results

Profile likelihoods. We evaluated the likelihood for a typically sized simulated data set where all parameters had been set to default values²³ (see Appendix D). For each parameter, the respective true value, that is, the value used for simulating the data set, is shown with a vertical dashed red line. Then, for each parameter, for 50 equidistant parameter values in the intervals shown, the likelihood for the data given the model was evaluated. Ideally, the likelihood should be maximal around the true value.

In Figure 4.6 we observe that the likelihood peaks, as expected, around the true value for most of the parameters. This means that (i) the parameters affect the likelihood and (ii) the likelihood may be used to recover their values. Individual likelihood evaluations are represented by dots. The plotted line smooths are just for guidance and do not represent the true likelihoods. The important observation here is that the highest evaluated likelihoods are always relatively close to the true value, even for the case of μ_2 , where the smoothed lines

²³These values vary slightly from the defaults used in Rabe et al. (2021) and Seelig et al. (2020). They are not to be understood as universally valid defaults but as fixed values wherever they are not fitted, and are merely reported here for reasons of transparency and reproducibility.

falsely suggest a flat likelihood.

Since not every fixation involves a retrieval, the new SEAM parameters can only have a very limited effect on the likelihood. Therefore, effects observed in the likelihood function are less pronounced than for the established SWIFT parameters such as processing span δ_0 . The fact that that higher likelihood evaluations nevertheless cluster around the true values is an indication that the parameters are identifiable, but their fitted values should be interpreted with caution.

For one of the parameters, μ_2 , the likelihood does not peak at all, which is probably because μ_2 only affects the model's behavior in rare instances. As μ_2 only determines the threshold value of a retrieval trigger, the likelihood is only affected for the small subset of words that trigger retrievals. By contrast, μ_3 affects the threshold of multiple words at the same time, i.e., all words previously processed. Also note that the profile likelihoods as well as the parameter recovery reported below are based on simulated data sets comparable in size to the experimental data of Mertzen et al. (2023). We would expect μ_2 to exhibit a more pronounced effect on the likelihood for larger data sets with more retrieval events. Despite the noise level of the profile likelihood of μ_2 , we decided to fit μ_2 as a free parameter. This means that different plausible values from the prior are considered throughout the sampling procedure instead of keeping μ_2 fixed at a (possibly implausible) default value.

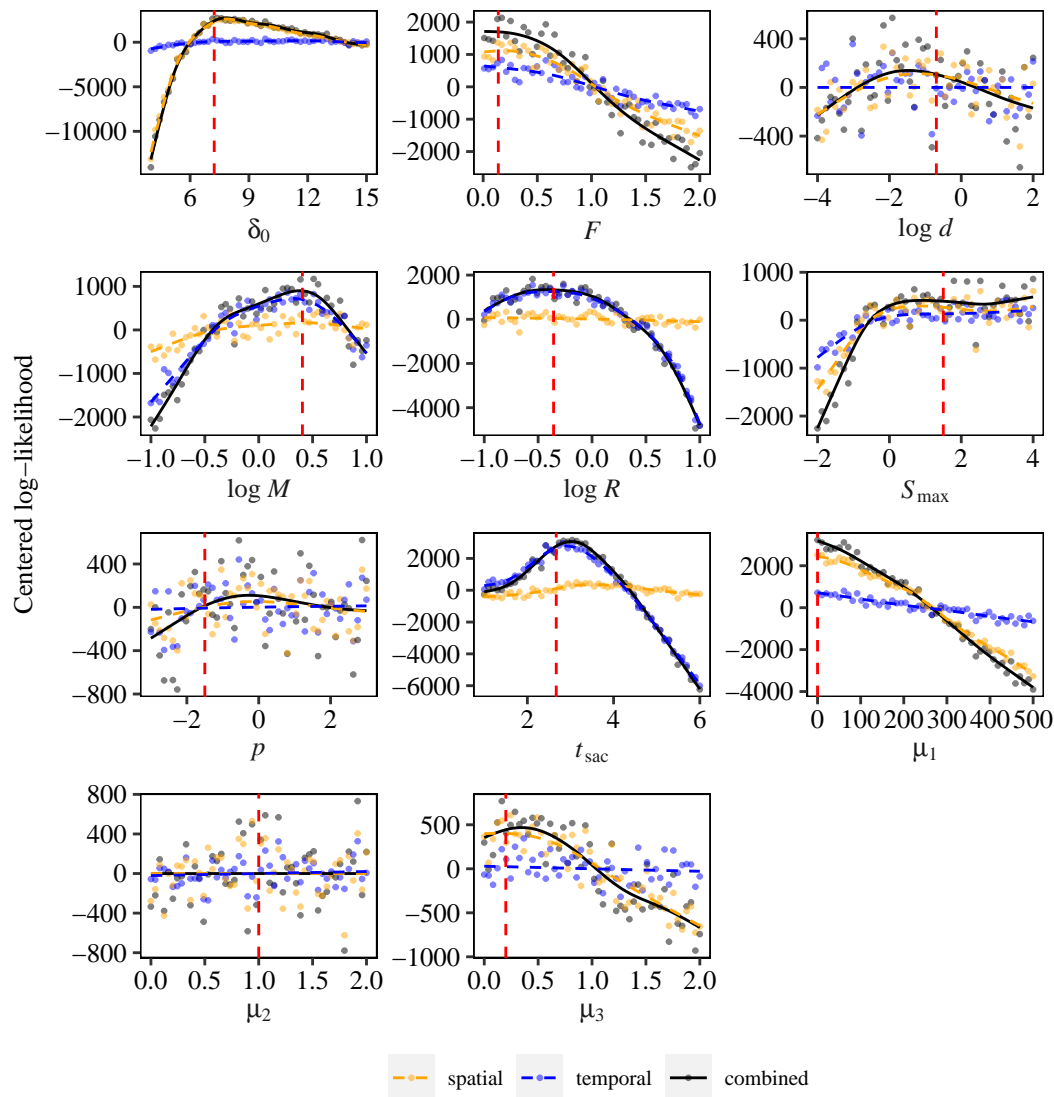
Parameter recovery. Analogous to the inspection of the profiles log-likelihoods, we simulated data from the known model but generated 50 data sets, each with a unique combination of random parameter values within the bounds of the previously inspected intervals, effectively sampling from the prior distribution. Then, we fitted the model to each of the data sets, using uninformative uniform priors over the bounds shown in Figure 4.6. Each fit is represented with one point per panel in Figure 4.7, showing 95% credible intervals (CrIs) on the y-axis and the true parameter value on the x-axis. Ideally, CrIs would be narrow intervals spanning around the identity diagonal.

We can see that the 95% CrIs almost always include the true value but are relatively wide, especially for the added parameters F , d , μ_2 , and μ_3 . Nevertheless, the agreement is generally good, as can be seen in the low normalized root mean square error or NRMSE values²⁴ and high correlations between true parameter values and CrI midpoints. This suggests that in general, true parameter values of simulated data sets can be recovered sufficiently well or at least with an acceptable level of uncertainty. As before, we note that parameter values, especially point estimates, should be interpreted with caution.

The reason for the high uncertainty for the new parameters is very similar to that for the profile log-likelihoods: Over the course of the entire fixation sequence, there are only very few retrieval events where these parameters could possibly have an effect on model

²⁴The NRMSE is the mean root mean squared deviation from the true value across all samples of the posterior, normalized on the sample range.

Figure 4.6
Example Profile Log-Likelihoods



Note. Centered profile log-likelihoods $\log L_M(\boldsymbol{\theta} | X)$ for a simulated data set X with known/true parameters $\boldsymbol{\theta}^*$. Profiles are generated by varying one parameter (dimension) of $\boldsymbol{\theta}$ at a time while holding the others constant at their respective *true* parameter value. True parameter values are denoted by the vertical red line. Dots in the background are individual stochastic pseudo-likelihood evaluations, each with a spatial and temporal likelihood component, and their combination (sum). Curves are GAM smooths on those individual evaluations.

behavior. Additionally, even when there is a retrieval, it is not guaranteed that it actually affects the activation of the currently fixated word, as the eyes may, for instance, already have continued past the retrieval trigger. Given these limitations, the recovery performance is surprisingly good, and the high correlations between true and recovered parameters appear very promising.

Summary statistics. So far, we have demonstrated that SEAM, like SWIFT in its most current version (Rabe et al., 2021), can be successfully fitted to simulated data: The true parameter values are in the vicinity of profile log-likelihood peaks and are contained within parameter recovery CrIs. This means that if we assume the true underlying cognitive architecture to be similar to SEAM, we can reliably use fitted parameters (or their credible intervals) to make inferences about it. However, as the true underlying cognitive architecture is unknown, such checks are per se impossible on experimental data. Instead, we compare simulated and experimental behavior on the basis of relevant summary statistics. For this, as explained earlier, we first split the experimental data into a training and test subset, fitting the model to 70% of the data of each participant and condition (training set), subsequently predicting eye trajectories for the other 30% (test set).

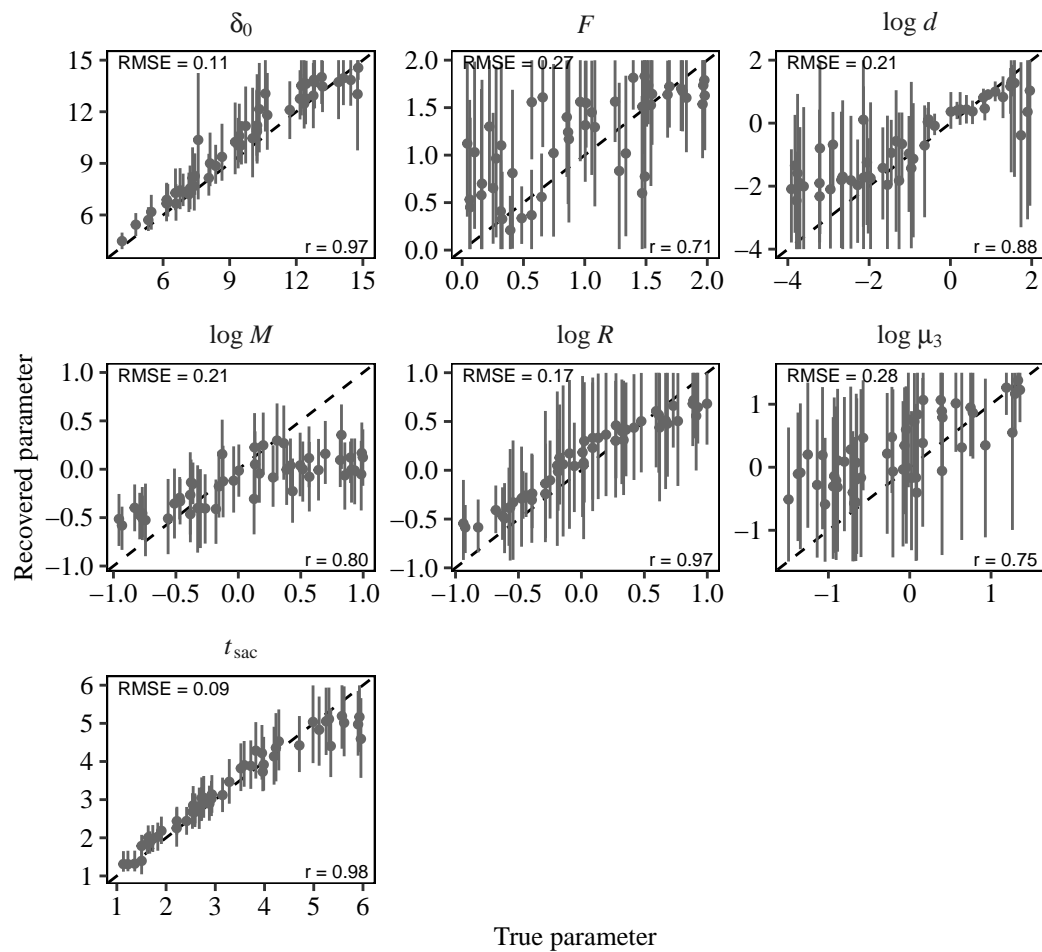
Rabe et al. (2021) had previously noted that SWIFT, with the cross-validation method described above, is unable to make reliable predictions for regressive eye movements. However, given that SEAM now incorporates processes for cue-based memory encoding and retrieval, and given that memory retrieval processes are specifically hypothesized to trigger regressions by modulating the activation of retrieval candidates, in SEAM we should see an improvement in regression-related statistics such as incoming/outgoing regression probabilities, as well as regression path durations. These are also two important dependent measures in which effects were found in the experimental data set (see Section 4.3 for a short summary; see Mertzen et al., 2023, for details).

In Figure 4.8, we show the comparison of summary statistics between experimental data and simulated data from the baseline SWIFT model (without memory retrieval) and SEAM (with memory retrieval). In all cases, SEAM predicts regression-related fixation probabilities and fixation durations more reliably than SWIFT. It is also noteworthy that not only the average across all word frequency bins but even word-frequency effects on summary statistics are reliably predicted.

Experimental effects of memory interference. Arguably the most critical test for the SEAM architecture is to evaluate whether the model can predict differences in summary statistics between experimental conditions in the design of Mertzen et al. (2023), which manipulates effects of memory retrieval on reading.

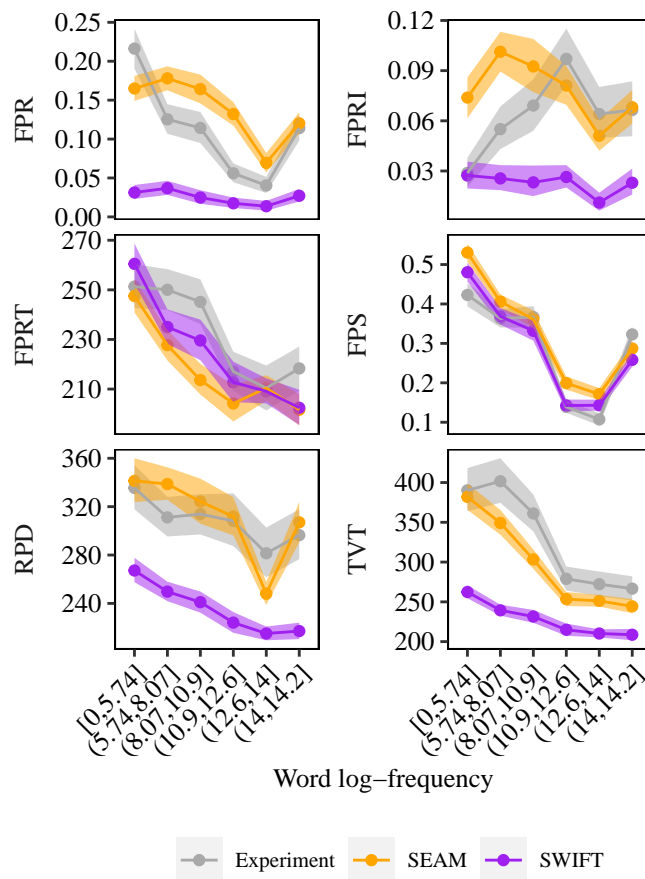
Based on a different experimental design, Rabe et al. (2021) were previously successful in demonstrating that SWIFT can be used to predict and explain differences in reading behavior when fitted to each participant and experimental condition separately. In our study

Figure 4.7
Parameter Recovery of SEAM Parameters



Note. Results of a parameter recovery for 50 simulated data sets, for which the parameters were randomly drawn from a uniform distribution with the bounds shown on the x-axes. 95% credible intervals (CrIs) are shown as error bars, centered around a point which is the mean of their lower and upper bounds. The diagonal is the identity line. Parameter recoveries with error bars spanning around the diagonal predict the true value within their CrI. Moreover, each panel shows the correlation between the true value and the point estimate as well as the normalized root mean squared error (NRMSE) of the CrI vs. the true value.

Figure 4.8
Spatial and Temporal Summary Statistics



Note. Displayed are relevant regression-related fixation probabilities and durations as estimated grand means from a linear mixed-effects model. Ribbons are 95% CrIs around the point estimate. Fixation targets (words) are grouped by log corpus frequency bins. *FPR* = First-pass outgoing regression probability, *FPRI* = First-pass incoming regression probability, *FPS* = First-pass skipping probability, *FPRT* = First-pass reading time (gaze duration), *RPD* = Regression path duration (go-past time), *TVT* = Total viewing/reading time. Words were not grouped into regions and all words of the sentences were considered.

Table 4.2

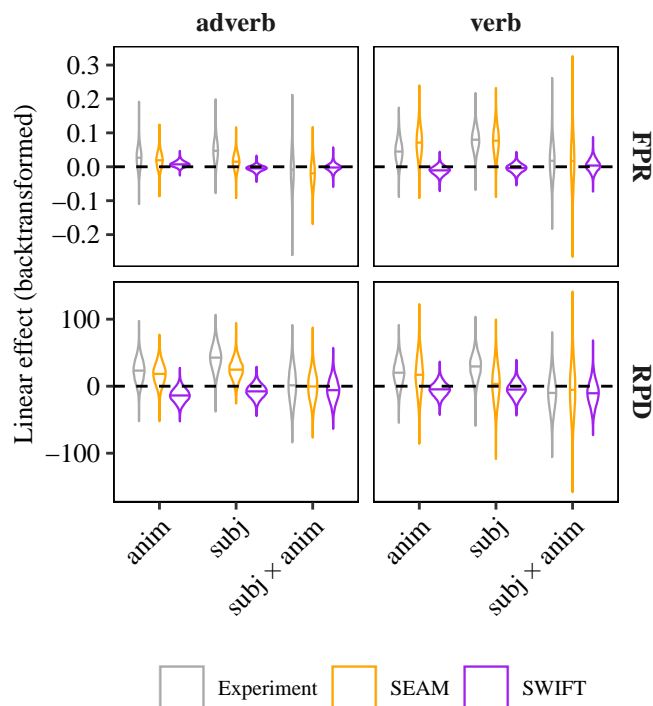
Summary of Empirical Vs. model Estimates From SEAM and SWIFT of the Subjecthood and Animacy Effects on Regression Path Durations and First-Pass Regressions

| Region of interest | Empirical estimates | | SEAM | | SWIFT | |
|-------------------------------------|---------------------|-----------|-----------|-----------|-----------|-----------|
| | subj | anim | subj | anim | subj | anim |
| Regression-path duration (ms) | | | | | | |
| pre-critical | [7, 77] | [-12, 60] | [-5, 55] | [-14, 50] | [-27, 12] | [-33, 6] |
| critical verb | [-6, 64] | [-16, 57] | [-49, 58] | [-37, 70] | [-26, 16] | [-25, 15] |
| First-pass regressions (percentage) | | | | | | |
| pre-critical | [-2, 11] | [-4, 9] | [-3, 6] | [-3, 7] | [-2, 1] | [-1, 3] |
| critical verb | [2, 15] | [-2, 11] | [0, 16] | [-1, 16] | [-3, 2] | [-4, 1] |

Note. Shown are the 95% credible intervals of the estimated effects from the data and from the two models. The empirical estimates are from the held-out data (30% of the data). *subj* = Effect of subjecthood, *anim* = Effect of animacy.

Figure 4.9

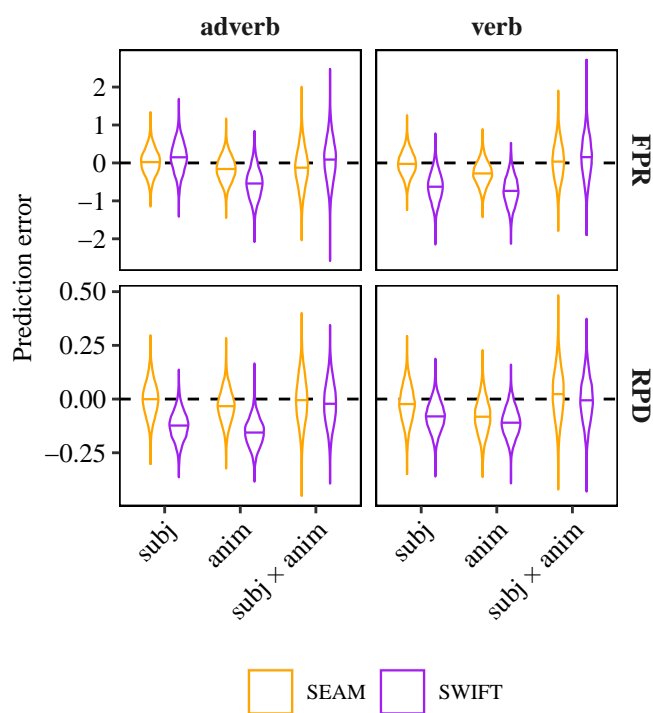
Posterior Distributions of Estimated Experimental Effects



Note. Experimental effects on outgoing first-pass regression probabilities (FPR, top row) and first-pass regression path durations (RPP, bottom row), as found in the experimental data (gray), baseline SWIFT (purple), and SEAM (orange). Violin plots are posterior distributions of mixed-effects models.

Figure 4.10

Distribution of Absolute Prediction Errors for Estimated Experimental Effects



Note. Prediction errors of experimental effects on outgoing regression probabilities (top row) and regression-path durations (bottom row). Violin plots are paired differences of posterior distributions of baseline SWIFT vs. experimental data (purple), and SEAM vs experimental data (orange).

presented here, however, we are only fitting one model at a time to each participant's data across all conditions, thereby considerably reducing the degrees of freedom. If the model is able to predict differences between experimental conditions, these do not originate from different parameter values for each condition but from the model dynamics, which are affected by the different feature specifications of the memory chunks across conditions. Therefore, capturing differences between conditions is a direct test of SEAM's added memory module. To illustrate the gain in empirical fit over baseline SWIFT, we also report predictions from SWIFT for reference. In SWIFT, no differences between experimental conditions are expected, because SWIFT has no parameters that could account for the processing cost of memory retrievals.

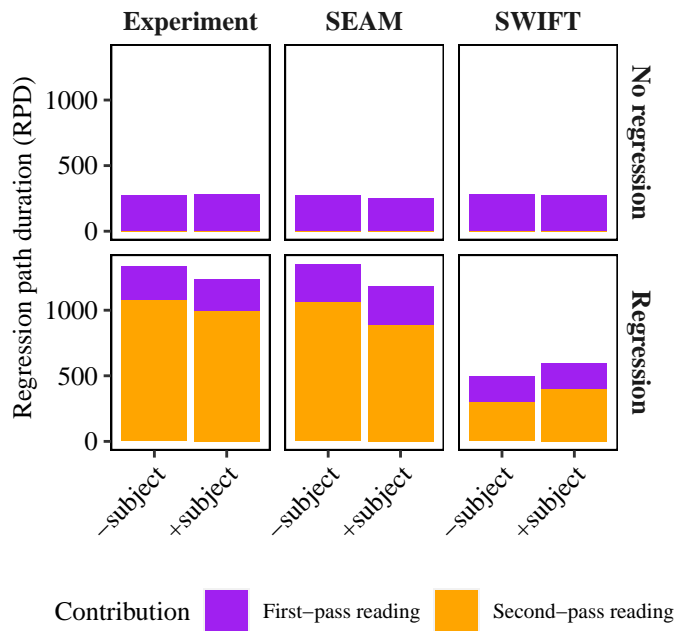
In order to evaluate the empirical fit of SEAM and baseline SWIFT, we conducted the same set of analyses for the observed experimental data and for the data predicted by SEAM and by SWIFT, after fitting each of the models to the training data sets. For both sets of data, we conducted a Bayesian mixed-effects regression for regression-path durations and outgoing regression probabilities as predicted by region and experimental condition (syntactic/semantic interference).

Table 4.2, and Figures 4.9 and 4.10 summarize the comparisons between the held-out empirical data and the predictions of SEAM and SWIFT. In order to interpret these comparisons, we compare SEAM and SWIFT against the empirical estimates from the held-out data using a region of practical equivalence (ROPE) approach (Freedman et al., 1984; Kruschke, 2014; Spiegelhalter et al., 1994) rather than formal model comparison methods such as k-fold cross validation, Bayes factors, or the like (for tutorial introductions to these topics, see Nicenboim et al., 2023). The ROPE approach is a graphical model comparison method that involves comparing model predictions against observed estimates from data; overlap in the posterior distribution of estimates provides an informal basis for deciding whether a model approximately matches observed estimates. In this approach, there is no notion of *statistical significance*; rather, the focus is on whether the model predictions are approximately consistent with the data. One important reason for taking this informal model comparison approach is the fact that the held-out data are relatively sparse. For this reason, the present evaluation should be seen rather as a proof-of-concept rather than a comprehensive evaluation. Such an evaluation would require significant amounts of benchmark data (for examples of such extensive evaluations, see Engelmann et al., 2019; Nicenboim et al., 2020; Yadav et al., 2022) and must be left for future work.

Table 4.2, and Figures 4.9 and 4.10 show that the predictions for the experimental effects of animacy (semantic interference) and subjecthood (syntactic interference) in the experimental data are generally more in agreement with SEAM than with SWIFT: the violin plots in Figure 4.9 from SEAM have a better overlap than the observed data than the predictions from SWIFT. This is true in both the pre-critical and critical regions, in both the first-pass

Figure 4.11

Effect of Subjecthood on the Contributions of First-Pass Reading and Rereading Times to Regression Path Durations

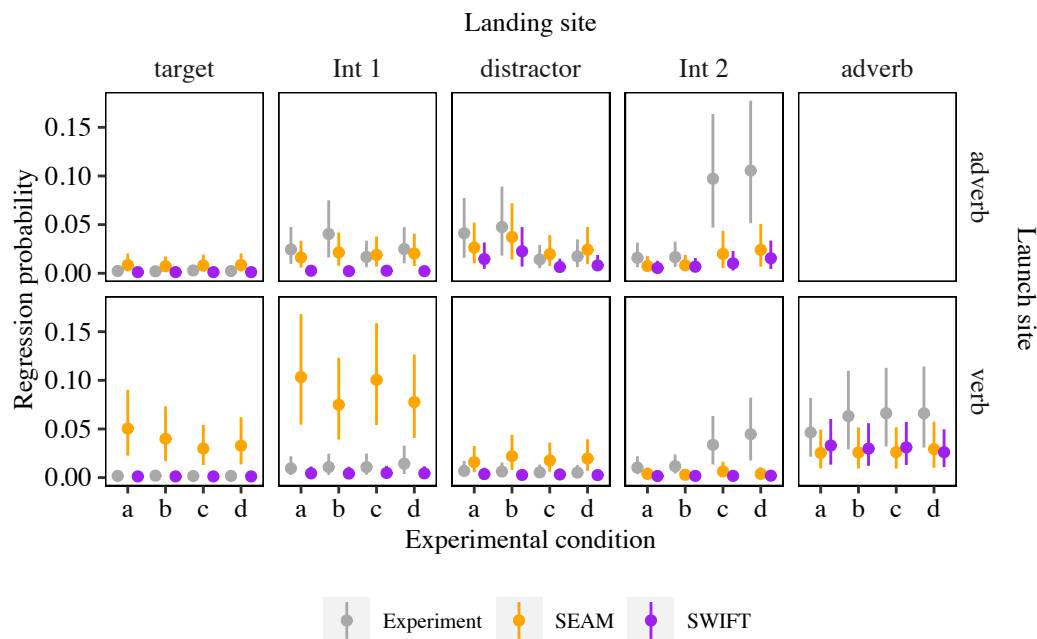


Note. Columns are sums of first-pass reading time on the launch site and rereading times on refixated previous regions until the (simulated) eye leaves to the right of the initial launch site.

regression and regression path duration measures. One exception is the subjecthood effect at the critical verb (see the bottom right panel in Figure 4.9); SEAM predicts essentially no effect of subjecthood, just like SWIFT. This is mainly because the regression paths predicted by SEAM are somewhat too short, i.e. return too early, in the +subject conditions (see Figure 4.11). We return to this in the Discussion section.

Given that SWIFT does not have any mechanism that accounts for cue-based memory retrieval, it is expected that the model predicts no effects of memory interference. Notice that the violin plots for the data as well as the SEAM and SWIFT predictions shown in Figures 4.9 and 4.10 are relatively wide; this is due to the fact that only 30% of the test portion of experimental data (the held-out data) are compared to the model predictions.

A main motivation of SEAM was to develop a model in which low-level psychological and high-level linguistic processes interact. The integration of the LV05-based memory module is expected to affect eye movements especially in cases of demanding dependency resolution and there, particularly strongly if there is high ambiguity between the correct dependents and distractors. Even though we already know that the Mertzen et al. (2023) data do not provide unequivocal evidence in support of this hypothesis, we can look at the distribution of regressions across trials conditional on launch and landing sites in order to investigate

Figure 4.12*Conditional Means of Experimental and Simulated Regression Probabilities*

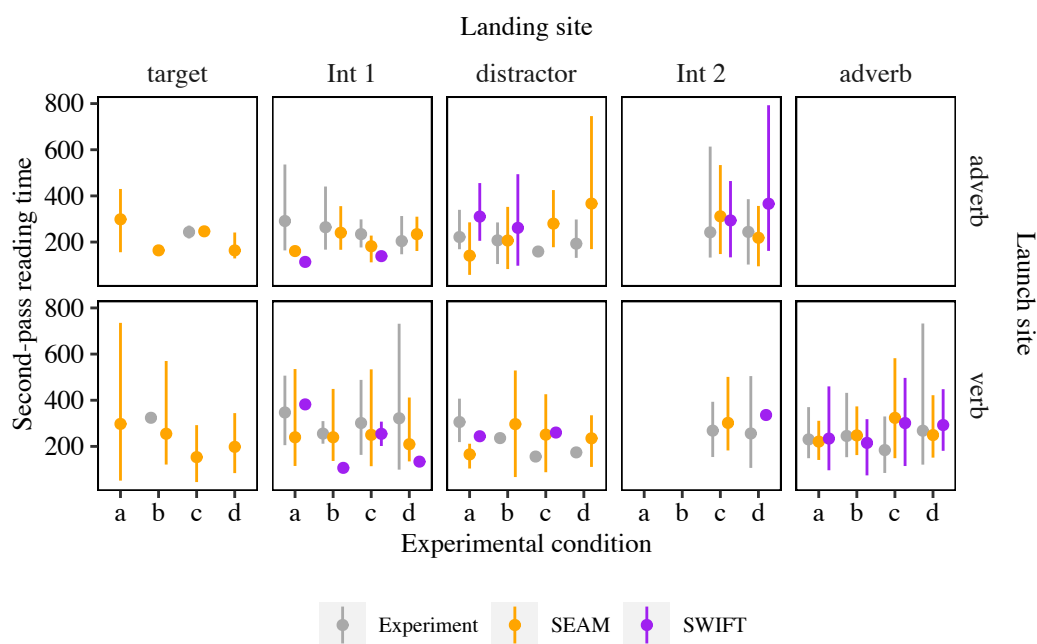
Note. Shown are estimates and 95% CRIs from nested logistic regressions. Each panel shows the mean proportion of trials with a critical regression given the launch site (rows) and landing site (columns). Estimates have been backtransformed to the linear scale. Regions Int 1 and Int 2 are intervening regions between target and distractor, and between distractor and adverb, respectively. Conditions a–d refer to the four conditions –subj/–anim, –subj/+anim, +subj/–anim, and +subj/+anim, respectively, as shown in Example 3.

where regressions from the (pre-)critical region tend to land in the experimental data and in the simulations. Figure 4.12 and Appendix E show that regressions in general have a tendency to land on the preceding word.²⁵ In these cases, SEAM is in better agreement with the experimental data than SWIFT. For regressions launched from the verb, however, SEAM currently predicts too many regressions on average, although the experimental effects (i.e., differences between conditions, see Figure E) are still in agreement with the experimental data. As there are generally very few regressions, both in the experimental and in the simulated data, analysis of regression durations is problematic but Figure 4.13 shows that they are also generally in good agreement with each other.

As SEAM and SWIFT are nested models,²⁶ the fact that SEAM but not SWIFT can predict different summary statistics is a first indicator that the differences in predictive power

²⁵Note that in conditions a and b in Figure 4.12, the distractor immediately precedes the pre-critical adverb, while conditions c and d have an intervening region between the distractor and pre-critical adverb.

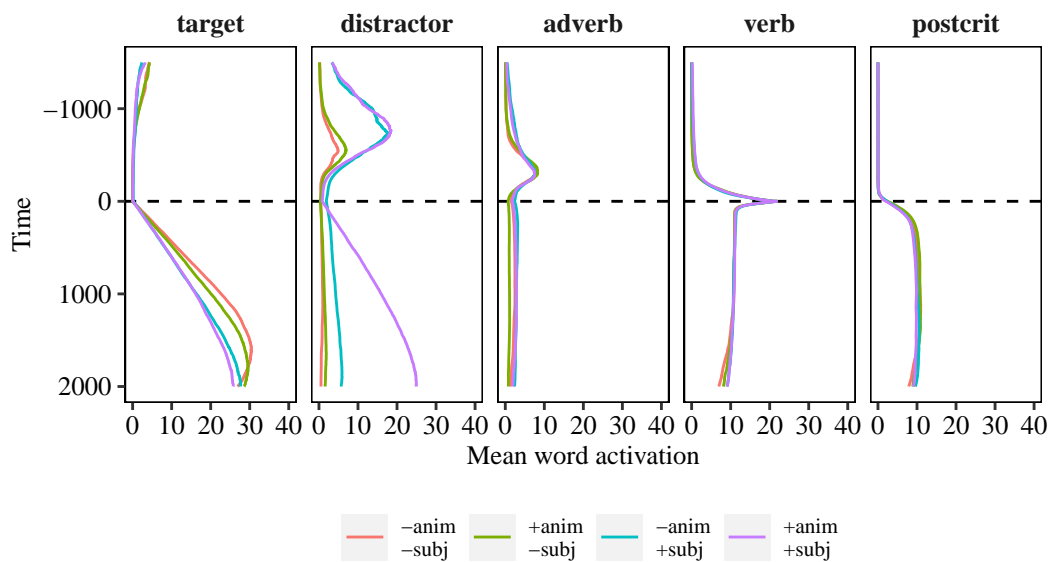
²⁶SWIFT is nested within the more complex SEAM, and the effects of the memory retrieval submodule on the word activations can be completely turned off by setting $F = 0$ and $\mu_3 = 0$ because that immediately ends any started retrieval and does not reactivate any previously encoded words.

Figure 4.13*Conditional Means of Experimental and Simulated Regression Durations*

Note. Shown are means and 95% quantiles of the raw data. Each panel shows the mean second-pass reading time (regression duration) following a critical regression given the launch site (rows) and landing site (columns). Regions Int 1 and Int 2 are intervening regions between target and distractor, and between distractor and adverb, respectively. Conditions a–d refer to the four conditions –subj/–anim, –subj/+anim, +subj/–anim, and +subj/+anim, respectively, as shown in Example 3.

Figure 4.14

Effects of Experimental Condition on SEAM Word Activations at Encoding of the Critical Verb



Note. SEAM word activation of the target, distractor, pre-critical adverb, critical verb, and post-critical region of a sentence grouped by experimental condition. Activations are averaged across 500 independent simulations of the same item in all four conditions. For each simulation, $t = 0$ is adjusted to the time of the start of post-lexical processing of the critical verb, that is, the start of the retrieval.

between the models may be due to the added memory retrieval submodule. To verify this and to attempt an explanation of the differences in observed behavior, we look at the differences in the internal model dynamics under the different experimental conditions.

In particular, we can examine the word activation field, which is the main driver for target selection probabilities in SEAM and SWIFT (Equation 4.8), including regressive saccades. In Figure 4.14, we show word activations in SEAM, averaged across 500 independent simulations, using the mean estimated model parameters across all model fits. Before averaging across simulations, all word activations are centered on the temporal dimension so that $t = 0$ is the time when the activation of the critical verb reaches its maximum, that is, when post-lexical processing of the critical verb starts and triggers the memory retrieval.

First, it is important to note that the activations of the critical verb, when normalized in time, do not vary substantially between experimental conditions. Although some conditions seem to have a slower decrease than others, overall the curves are very similar in all conditions. When the retrieval starts at $t = 0$, retrieval candidates are reactivated, with their memory activation $A'_{j,m}(t)$ modulating the transition rate $w'_j(t)$ (see Equation 4.12) of word/memory chunk j . While the activation for the target word seems to be very similar over time between conditions as well, there is some variability in the time course of the activations of the distractor noun and of the adverb around the retrieval.

Regarding the adverb, the main reason it is reactivated during retrieval is that it has the highest base-level activation $B(t)$, as it was most recently encoded/accessed in memory before the retrieval started. The later processing of object noun distractors also attenuates the processing that the adverb receives, which leads to weaker reactivation of the adverb during the retrieval.

We can also observe that the distractor word activations prior to the retrieval peak earlier for the two conditions where the distractor is a subject noun, that is, in the conditions where there is syntactic interference. This effect is not related to the retrieval at the critical verb (which has not started at this time), but is due to the distractor appearing earlier in the sentence when it is a syntactic subject. Interestingly, the distractor noun only significantly peaks during the retrieval in the +animate/+subject condition, that is, when both features match the retrieval cues. The distractor thus only attracts regressions when both the animacy and subjecthood features match, i.e., when there is both syntactic and semantic interference. Despite this difference in word activations, there is no significant difference between the proportions of observable targeted regressions from the critical verb to the distractor noun between any of the experimental conditions. This is true for the experimental data as well as for the data simulated by SEAM and SWIFT, as shown in Figure 4.12 and Appendix E.

As the estimates show, there is no indication in the experimental data that the distractor is regressed to more often in the +animate/+subject condition. The distractor's activation pattern in Figure 4.14 is simply a consequence of the hard-coded assumption in LV05 that

it has the highest feature match in this condition. Interestingly, however, the predicted data from SEAM do not show an increase in incoming regressions to the distractor either. An increase in word activation thus does not necessarily translate into a change in observed eye movements. The lack of a direct effect on distractor refixations is likely due to oculomotor error, which is more influential for long-range saccades, and due to upcoming words having even higher activations than the distractor.

Based on results from preliminary post-hoc analyses, the overestimation of the average regression probability from the critical verb to preceding regions (see Figure 4.12) is also probably due to the hard-coded retrieval schedules. Even though the times of memory encoding, and therefore the base-level activation, are stochastic and completely governed by the eye's trajectory, the feature match is deterministic. Future work could investigate alternative links between memory activation $A'_{j,m}(t)$ and word activation $a_j(t)$ or transition rates $w'_j(t)$, as they currently implement a very strong linking hypothesis.

In this context, we also note that – as far as we are aware – the only study that has previously looked at word-level rereading as a function of similarity-based interference is Lee et al. (2007). The authors report longer rereading times for a sentence-initial region containing both the retrieval target and the distractor in their high-interference conditions, but the Korean sentences used in their study were relatively short compared to those used by Mertzen et al. (2023). In future work, shorter sentences should be a fruitful testing ground for SEAM. If SEAM generates more linguistically mediated targeted regressions in shorter sentences, this would be in line with human data (Inhoff & Weger, 2005).

Summary. We showed that both SEAM and SWIFT can be fitted to the Mertzen et al. (2023) experimental data set. In contrast to SWIFT, however, SEAM's predictions are in good agreement with the overall and by-frequency regression probabilities and regression-path durations. SEAM shows the more specific memory interference effects, that is, differences in regression probabilities and regression-path durations due to differences in the animacy and subjecthood of a distractor noun.

Given that the compared models SEAM and SWIFT only differ in the supplemental cue-based memory retrieval processes contributed by the LV05 component, we can attribute the better performance of SEAM in these metrics to LV05 principles with the four additional parameters that were fit to the training data from Mertzen et al. (2023) (F , d , μ_2 , and μ_3). It is also noteworthy that these parameters were estimated based on a restricted training data set for each participant, and that the model can make reasonable predictions on the held-out test data for all experimental conditions with a single model fit for each participant.

Furthermore, even though the models are compared to each other and to the experimental data using summary statistics and predicted experimental effects, neither SWIFT nor SEAM was directly optimized to reproduce these measures. Instead, both models were fitted directly to the raw, unbiased fixation sequences of each participant. Therefore, the models can make

reasonably accurate predictions for summary statistics and experimental effects although they are not specifically fitted to them.

4.4.3 Discussion

We showed that adding a memory interference mechanism in the SWIFT architecture—resulting in the SEAM model—allows us to bring together eye-movement control theory and a psycholinguistic account of dependency completion. We demonstrated that the key regressive eye-movement related patterns in an experimental psycholinguistic data set can be accounted for by the SEAM architecture. Specifically, we showed that first-pass regressions and regression path duration patterns that occur due to the interference manipulation in the Mertzen et al. (2023) data can be accounted for by SEAM, but not by SWIFT; in SEAM, as in the data, both syntactic and semantic interference have an impact on the two dependent measures at the pre-critical region and the critical verb.

The main results of our simulations are summarized in Table 4.2 and Figures 4.9, 4.10, and 4.14. There were three interesting patterns in the SEAM fit that deserve discussion. First, as shown in Figure 4.9, at the critical verb, regression path durations from SEAM show essentially no effect of subjecthood; this is surprising because the data do show such an effect. At the same time, in SEAM, first-pass regressions at the verb show a clear subjecthood effect. This is because even though regressions were triggered at the verb, which should itself increase the mean RPD, regression paths predicted by SEAM return too early in the +subject conditions, thereby masking the effect on RPD²⁷.

The second interesting pattern relates to the effects observed at the pre-critical adverb region (*the attorney whose secretary had forgotten [...] frequently complained*, see Example 3). Recall that in the original LV05 model, sentences are processed in strictly serial order. Effects of similarity-based interference at the pre-critical adverb are thus unexpected under this model: Given the assumption that the verb is the retrieval trigger, there should be no retrieval-related effects before it is read. Nevertheless, Mertzen et al. (2023) did observe interference effects at the pre-critical adverb (others have found similar patterns in the pre-critical region; see Lago et al., 2021; Van Dyke, 2007). Mertzen and colleagues discuss several possible reasons for these effects: Differential processing spillover from previous regions due to differences in sentence complexity between conditions, lingering memory interference during encoding of the noun phrases, and predictive processing of the verb. A final important possibility considered by Mertzen et al. (2023) is parafoveal preview of the verb while the adverb is being processed, so that the verb can trigger the retrieval prior to

²⁷As the regression path duration is the sum of gaze durations on the current word and on all preceding regions until the (simulated) eye leaves to the right of the current word, an effect in regression path durations could be due to (a) an effect on gaze durations on preceding regions, (b) an effect on gaze duration on the current word, (c) a combination of both. Likewise, a null effect could be a masked effect of gaze durations on the launch site vs. gaze durations on preceding regions.

being fixated. Our SEAM simulations are partly consistent with this last account: In 25% of our simulations, the verb reaches the retrieval stage while the adverb is being fixated. However, there is also processing spillover in the form of residual word activation in SEAM. Especially in the +subject conditions, where there is an additional retrieval in the embedded sentence at *was important*, and the activation of the retrieval target may not have fully decayed when the adverb is read, leading to more regressions. Based solely on the Mertzen et al. (2023) data and the small sample size of the held-out data, it is difficult to quantify the relative contributions of preview and spillover, and we leave this issue to future research. Nevertheless, SEAM provides a promising starting point for tackling possible pre-critical retrieval effects.

A third noteworthy pattern occurs in Figure 4.14; the +subject/+animate condition causes a large increase in the distractor's word activation after the critical verb is encoded. This suggests that the probability of the distractor to attract regressions should be much higher in that condition than the sum of the +subject/–animate and –subject/+animate conditions. Even though the combination of the two retrieval cues is additive at the level of the LV05 memory activation (see Equation 4.1), the exponential transformation of $A(t)$ in Equation (4.11) significantly amplifies it. Nevertheless, the superadditive effect on the distractor's activation when it matches both retrieval cues does not generate any detectable overadditive effects in the analyzed regression-related dependent measures (regression path duration and first-pass regressions). As discussed in the previous section, the spike in activation does not necessarily translate into observed regressions, partly because the large distance between the verb and the distractor amplifies the influence of oculomotor error. With less complex sentences, it is thus possible that SEAM would show effects on the observed regression probabilities.

4.5 General Discussion

From the very beginning of eye-movement research in reading, a dominant idea has been that the eye and mind are tightly coupled (e.g., Just & Carpenter, 1980). After psycholinguists started looking at fixation patterns in reading as a function of language comprehension difficulty, an important idea that was expressed in a now-classic paper by Frazier and Rayner (1982) was the selective reanalysis hypothesis: this was the idea that increased comprehension difficulty (e.g., due to garden-pathing) leads to targeted regressions to a preceding region that caused the processing difficulty. Although the strongest version of selective reanalysis, and thus of the eye-mind assumption, is difficult to uphold given subsequent investigations (e.g., Mitchell et al., 2008; von der Malsburg & Vasishth, 2011), it is nevertheless well-established that increased regressions are triggered when language processing difficulty occurs (e.g., Clifton et al., 2007), and that rereading can aid comprehension (Schotter et al., 2014). We assume that the mixed evidence in the psycholinguistic literature regarding se-

lective rereading (see Paape et al., 2022, for a review) may be the result of a more indirect linkage between higher-level sentence processing and saccade targeting: In our model, retrieval events during dependency completion affect the activation values of previous words in the sentence. Words with higher activation will tend to attract saccades, but due to the inherent stochasticity of the eye-movement control system and oculomotor error, subtle linguistic manipulations do not necessarily engender measurable effects at typical sample sizes.

Most of the psycholinguistic work carried out on reading until now has side-stepped the underlying complex latent processes involved in reading, and instead focused only on key events involved in linguistic dependency completion. Abstracting away from these underlying latent reading processes has had many advantages, a major one being that it allows us to focus exclusively on the psycholinguistically interesting aspects of processing at the level of the sentence representation. On the other hand, the simplification comes at a cost, because interactions between constraints on eye-movement control and language comprehension end up being ignored.

Interestingly, cognitive psychology has gone in a completely different direction than psycholinguistics: there, the focus has been on spelling out detailed process models of eye-movement control that rely primarily on relatively low-level drivers of eye movements, such as frequency and word length. Models of eye-movement control such as E-Z Reader (Reichle et al., 1998) and SWIFT (Engbert et al., 2005) have shown excellent performance in explaining benchmark data in reading, without modeling the higher-level cognitive processes such as linguistic dependency completion in any great detail.

One major gap in the literature is that these two threads—psycholinguistic explanations of reading difficulty versus cognitive psychology models of reading—have only rarely been considered to be joint actors in explaining key effects observed in experimental data from psycholinguistics. Our paper makes an attempt to fill this gap: using data from a classic similarity-based interference design, we demonstrate one way in which an eye-movement control model, SWIFT, can be extended to include dependency completion processes. We show that such an extended model (SEAM) can produce regressive eye movements triggered by retrieval that occurs during linguistic dependency completion. Developing such models is the only way to unpack the latent processes involved in reading and to investigate how *low*- and *high*-levels of cognitive processes interface dynamically. To our knowledge, SEAM is the only model to date that extends a complete model of eye-movement control with a detailed model of linguistic dependency completion, using data from a planned experiment in psycholinguistics and rigorous statistical inference.

Apart from using SWIFT as the eye-movement module, SEAM differs in important ways from previous integrative models of eye movement control and higher-level sentence processing. For instance, Über-Reader (Reichle, 2021), whose eye movement module is highly similar to that of E-Z Reader (Reichle et al., 1998), has a parsing module that builds syntac-

tic structure, but each parsing step is assumed to take the same amount of time. In SEAM, by contrast, completing syntactic dependencies takes a variable amount of time that is determined by the LV05 equations (which originally come from ACT-R). Furthermore, regressive saccades are not captured by Über-Reader, but are modeled dynamically in SEAM.

Another integrative model proposed by Dotlačil (2021), whose eye movement module is also based on E-Z Reader, makes use of ACT-R Equations, but in a different way from SEAM: In Dotlačil's model, the latency with which a given dependent word is integrated into the sentence's syntactic representation depends on the retrieval time for the dependent words and additionally on the retrieval time for the relevant parsing rule from declarative memory. SEAM does not assume retrieval of parsing rules, which are assumed to be represented as procedural knowledge, as in the LV05 model. Another salient difference between the models is that regressions in Dotlačil's model are only triggered when parsing failure occurs, while regressions in SEAM are driven by the dynamic target selection processes taken over from SWIFT. As a final comparison, the model of Engelmann et al. (2013) and (Vasishth & Engelmann, 2022) combines an LV05 sentence processing module with eye movement control based on EMMA (Salvucci, 2001), but also does not provide a detailed model of saccade targeting, unlike SEAM.

There are of course several limitations to the present work. First and foremost, the current implementation of SEAM and its evaluation are only a proof-of-concept. Because of the absence of large-scale data sets with psycholinguistically interesting manipulations, it is difficult to present a comprehensive evaluation of the proposed SEAM architecture. However, such an investigation is in principle possible to carry out, given (i) the progress on Bayesian inference for process-based models and (ii) the fact that more and more researchers are releasing data and code associated with their published papers. We expect that in future work, more comprehensive evaluations of architectures like ours can be carried out, using large-scale data from a broad range of phenomena in psycholinguistics. At a minimum, such an investigation would need to include cross-linguistic data from garden-path sentences of different types (e.g., Frazier, 1979), predictability manipulations (e.g., Levy, 2008), the full spectrum of similarity-based interference effects (e.g., Jäger et al., 2017), underspecification effects (e.g., Swets et al., 2008), etc. This would be a sizable project, but one which would significantly advance our understanding of how eye-movement control and parsing interface during reading.

A second limitation is that, due to the computational complexity of investigating such a detailed model of reading, formal model comparison between the baseline SWIFT model and the SEAM model is difficult to carry out. We avoided overfitting the models to data by separating the empirical data into a training set and a held-out set, and evaluating the model fit only on the held-out set. This is already a significant advance over conventional approaches to model evaluation; in both cognitive psychology and psycholinguistics, it is common to

evaluate a model on the same data that it is trained on. In principle, it is possible to go even further than we did in this paper, and to evaluate predictive performance by using k -fold cross-validation. This would involve creating k (usually, in machine learning, $k = 10$) subsets of the data to train on, and then use the k held-out data sets for evaluation; this would allow us to compute a quantitative measure of average fit, such as expected log pointwise density (e.g., Gelman et al., 2014). We did not carry out such a quantitative evaluation because it would have been computationally extremely costly. For example, just the pure SWIFT model discussed in Rabe et al. (2021) required a high-performance computing environment, and the total computing time was approximately 10,000 core hours, amounting to 3.5 hours run time on 72 independent parallel nodes with 40 cores per node. Our goal in the present work was to get as close as possible to the underlying processes involved in reading, but obviously this comes with an unavoidable computational cost.

4.6 Conclusion

We present an integrated model of eye-movement control and linguistic dependency completion while reading. The model, called SEAM, is an integration of the SWIFT model of eye-movement control and the Lewis-Vasishth model of sentence processing. SEAM is evaluated using experimental data from a similarity-based interference experiment. We show that the SEAM model can account for empirically observed regressive eye movements; in the model, regressive eye movements are shown to be triggered by retrieval processes that result from higher-level dependency completion during sentence parsing. To our knowledge, this is the first demonstration of how eye-movement control and sentence comprehension processes can interact in explaining data from a psycholinguistically controlled experiment.

Chapter 5

General Discussion

In this dissertation, I presented results from three modeling studies embedded in the domains of eye movements and reading. The studies address several key issues of contemporary reading models, including parameter inference, model comparison, and the integration of eye-movement models with linguistic dependency completion processes. Collectively, they contribute to advancing the field of computational cognitive science, particularly in the context of eye-movement control during reading, and motivate further methodological and theoretical work.

5.1 Summary

In Chapter 2, we highlighted three main limitations in existing approaches to parameter inference and model comparison for dynamical cognitive models of reading: (1) neglect of the time-ordered nature of observations, (2) lack of likelihood implementations, and (3) inability to explain interindividual differences. These limitations were addressed through the development of a likelihood-based framework for the SWIFT model. By combining approximated spatial and temporal likelihood components, we demonstrated the feasibility of likelihood-based Bayesian inference for individual subjects. Interindividual differences in reading behavior correlated with differences in estimated model parameters, providing a significant advancement for understanding complex reading tasks.

The formulation of a likelihood for a complex dynamical model like SWIFT is expected to be non-trivial. Nevertheless, in Chapter 2, we showed that it is possible to implement dynamical cognitive models using sequential likelihoods, embedded in the broader theoretical approach of data assimilation. Even when the analytical form of the likelihood is intractable, methods such as probability density approximation or pseudo-marginal likelihoods are worthwhile. Generally, this approach is also applicable to models outside the realm of reading and even besides eye-movement research.

The subsequent Chapter 3 focused on the application of the SWIFT model to demanding experimental data, emphasizing the importance of Bayesian inference and likelihood functions in contrast to other types of parameter estimation. We proposed a more realistic model of oculomotor error, demonstrated the reliable estimation of model parameters for individual subjects, and were able to capture interindividual differences in reading behavior across various experimental conditions. This work represents a significant step toward integrating

cognitive modeling of eye-movement control with the prediction of interindividual differences.

The final study presented in Chapter 4 delved into a broader theoretical context of eye-movement research in reading, emphasizing the interplay between eye and mind in the investigation of syntactic processing difficulty during reading. Highlighting the different theoretical angles of linguistic explanations of reading difficulty and psychological models of eye-movement control, SEAM bridges this gap by combining eye-movement control with linguistic dependency completion processes. It should be noted that SEAM is a proof-of-concept, and comprehensive evaluations across various psycholinguistic phenomena are needed to validate its architecture.

5.2 Statistical Rigor Through Bayesian Inference

After introducing a likelihood model for SWIFT, all three studies have highlighted that likelihood-based Bayesian inference is a powerful and theoretically grounded approach for parameter estimation in cognitive models. Unlike measures based on deviance between experimental and simulated summary statistics, likelihood-based parameter inference offers several advantages:

1. There is no need to aggregate experimental data across participants and/or items in order to calculate summary statistics. Instead, especially given sequential likelihoods, the model can be fitted directly to the unaggregated observations, which greatly improves the precision and spares the selection and weighting of different summary statistics.
2. Even though summary statistics are not used for model fitting, when aggregating across simulated data, the models are still able to produce reliable predictions at the level of summary statistics.
3. The implementation of likelihoods is necessary for using Bayesian parameter inference methods. Especially when working with models with highly dimensional parameter space, the regularization of the parameter search is critical for efficient and timely inference and results. This also allows for efficient hierarchical model fitting within a Bayesian framework (see Section 5.8 for more details).

Critically, further research should also address the efficiency of simulated, pseudo-marginal likelihood, proposed in Chapter 2. Although it was demonstrated in all studies that parameter inference is reliable, given the results of profile-likelihood evaluations and parameter recoveries, the computation of the approximate likelihood is costly. Particularly, a replacement of simulation-based likelihood components, i.e. the temporal likelihood approximated

by kernel density estimation, could potentially increase the computational performance of the algorithm.

5.3 Individual Differences in Eye Movements During Reading

Reading is subject to a variety of individual factors, e.g. age, experience, domain knowledge, etc. Therefore, we can expect and have indeed observed considerable interindividual differences in both the observed summary statistics and estimated model parameters. In particular, the following observations and conclusions can be made with regard to individual differences:

1. Modeling individual differences allows us to move beyond a one-size-fits-all approach and gain a deeper understanding of the complexity of reading behavior. It acknowledges that readers vary in their cognitive processes and strategies, and this variability can be essential to understanding the mechanisms underlying reading.
2. Considering individual differences in model parameters also allows us to estimate individual estimates for particular participants. In turn, this also makes it easier to detect and reliably predict commonalities across participants.
3. Understanding individual differences in reading behavior can have practical applications in education and clinical settings. Tailoring interventions or support to an individual's reading profile can be more effective in improving reading skills or addressing reading-related disorders.

While the discussed research represents significant progress in modeling individual differences, there are challenges to overcome. Hierarchical Bayesian approaches and the integration of larger datasets are mentioned as potential future directions to increase the reliability of estimates for individual subjects.

5.4 Integration of Cognitive Processes During Reading

The integration of eye-movement control models with linguistic dependency completion processes represents a significant step toward bridging the gap between cognitive psychology and psycholinguistics. While cognitive psychology models have focused on low-level drivers of eye movements, psycholinguistics has explored comprehension difficulty and regressions during reading. The model SEAM, introduced in Chapter 4, aims to bring these two threads of research together. There are a number of implications for theory and further research:

1. Integrative models that consider processes at different “levels” such as perception, higher cognition, language, and oculomotor execution, provide a more comprehensive understanding of reading. By incorporating accounts for different aspects of the

behavioral response, researchers can improve the predictive power of their models, explaining not only isolated patterns in the data but also complex reading behaviors that span over the entire sentence or beyond.

2. SEAM is a proof-of-concept model. In addition to the added Lewis and Vasishth (2005) memory-retrieval model, other models could be integrated to account for additional aspects. For example, the concept of predictability in SWIFT and SEAM, which is fixed a property of the word fixed in the corpus, could be dynamically determined by means of surprisal (Levy, 2008). Moreover, lower-level cognitive processes with regard to word identification could partially or fully replace the “lexical” processing stage of words.
3. While the introduction of models like SEAM is promising, with increased complexity, there is a need for comprehensive evaluations using large-scale data sets and various psycholinguistic phenomena. Researchers should strive to validate these integrative models across a wide range of reading scenarios, such as garden-path sentences, predictability manipulations, and interference effects. This validation will help establish the credibility and generalizability of these models.

5.5 Effects of Similarity-Based Interference

In Chapter 4, it was shown that SEAM has an overall better predictive power for the considered experimental data set than the baseline SWIFT model. That was particularly true for “later” reading measures such as regression-path durations (go-past durations), total reading times, or re-rereading times. This suggests that the effects, for which SEAM can account, occur later during reading and may not be well described by earlier (first-pass) reading measures, which is in line with previously published literature on syntactic and semantic effects on eye movements during reading (e.g., see Huestegge & Bocianski, 2010; Weiss et al., 2018).

The model did have difficulties predicting the precise distribution of landing sites of retrieval-induced regressions. Presumably, this was the case because (1) retrievals were hard-coded, (2) not all words in the sentence were fully coded in the predefined memory schedule, (3) reading a sentence with a retrieval always triggered a retrieval, and (4) cue-feature matches were identical across participants and trials, i.e. the same word is identically encoded and retrieved across all trials and participants. However, it is plausible that words are encoded with different features and cues between participants or across trials, e.g. depending on reading skill, language proficiency, task motivation, or overall accuracy. Consequently, there is no guarantee that interference even occurs at all for a given experimental trial, whereas the model always assumes it. An improved version of SEAM could therefore

also consider a more complete feature schedule across all words of the sentence and a noisy encoding or retrieval of features during dependency resolution to add more stochasticity.

5.6 Accessibility of Model Implementations

Another issue to consider is the extreme model complexity of SEAM and the complex implementation in the C programming language. Already simpler models that do not require compiling the model code before using it, have usability and accessibility issues, making it difficult for other researchers to use the models for predicting data, let alone model comparison or fitting. This can also make it more difficult to evaluate the models against other data sets or using different inferential methods, which can threaten the scientific value of the underlying theories.

Therefore, simpler model implementations should be provided for interested researchers to become acquainted with key concepts of the model. Building on the simplified R-based *toySWIFT* version proposed by Engbert and Rabe (2023), which greatly simplifies key characteristics of SWIFT and offers an accessible interface for model simulations and parameter inference, a simplified version of SEAM could be equally helpful in advancing the scientific outreach. It is also possible to establish models within larger frameworks, in which components of the models can be selectively switched on and off or integrated with other approaches.

5.7 Model Comparison of Complex Models

Integrated reading models are powerful but also complex. Given the increasing number of competitors, a model comparison is in order to evaluate their relative usefulness. Ideally, likelihood-based approaches could be entertained but the lack of a likelihood implementation for some of these models complicates this endeavor. Even if a likelihood has been implemented, though, likelihood-based model comparison is computationally costly and, due to the objectivity of the likelihood, does not permit model comparison with regard to very specific patterns of the predicted behavioral response, e.g. comparing which model makes “better” predictions for regression probabilities under different experimental manipulations. Moreover, reading models are often evaluated on different corpora, sometimes using very simplistic stimuli. This makes a direct comparison based on published metrics alone impossible.

Instead, assuming availability of model implementations, it may be worthwhile to establish a reading benchmark similar to the MIT Saliency Benchmark (<https://saliency.tuebingen.ai/>; Kümmerer et al., 2018) that compares predictions of reading models. Modelers would retrieve a standardized training data and sentence corpus and return their predictions for sen-

tences that were withheld from the training subset. Those would then be compared against the true withheld data along some set of “gold-standard” summary statistics that are often reported in the field, such as different fixation durations and probabilities, within-word landing positions etc.

5.8 Hierarchical Bayesian Modeling

It has been demonstrated in this dissertation and elsewhere that there are considerable individual differences in reading, which manifests both in interindividual variability in observed summary statistics as well as in fitted model parameters. A clear disadvantage of the approach applied in the presented studies is that each individual participant’s data were fitted separately, without considering the data of other participants. This means that, effectively, for n participants, all fitted parameters $\theta = \{\theta_1, \dots, \theta_n\}$ were fitted exactly n times. This is inefficient for at least three reasons:

1. Parameter values θ can be expected to differ between participants but only with some variance (σ_θ^2 , random effect) around a mean (μ_θ , fixed effect). Estimating those distributional *hyperparameters* in addition to the participant-level parameter values can help regularizing the estimates. For example, we can expect the processing span of a participant to vary maybe between 4–14 letters around the fovea. However, if all participants cluster around an average of ~ 7 characters, it is more likely that an extreme estimate closer to the edges of the prior is due to inefficiencies of the sampler than a truly extreme parameter value. In such cases, the fitted parameters of the other participants can help shrink the outlier toward a more reliable estimate.
2. The parameters could be correlated across participants. For example, a participant with a very narrow processing span δ could be expected to make more frequent saccades, i.e. have a shorter global timer t_{sac} . Like the variances described above, these correlations can be fitted as part of the participant-level parameter covariance matrix.
3. For some participants, data quality or quantity will be lower-than-average. In those cases, their parameter fits could “borrow” statistical power from other participants. This generally applies to all parameters fitted across subjects but could be especially useful for parameters that are not expected to vary between subjects, i.e. where the hyperprior is maximally regularizing.

Of course, the introduction of a hierarchical prior adds some complexity to the sampling algorithm. Assume for simplicity that the n true parameter values of the single parameter $\theta = \{\theta_1, \dots, \theta_n\}$ are normally distributed. Then, the MCMC sampler, instead of sampling

proposals for θ directly from a specified prior, will first sample the distributional hyperparameters

$$\mu_\theta \sim \mathcal{N}(\nu_\mu, \Sigma_\mu^2) \text{ and} \quad (5.1)$$

$$\sigma_\theta^2 \sim \text{IG}(a_\sigma, b_\sigma), \quad (5.2)$$

where ν_μ , Σ_μ^2 , a_σ , and b_σ define the shape of the (known or assumed) hyperpriors of the hyperparameters, which is Gaussian for the fixed effect μ_θ and inverse-gamma for the random effect σ_θ^2 . Subsequently, the sampler will generate a proposal

$$\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2) \quad (5.3)$$

with mean μ_θ (fixed effect) and variance σ_θ^2 . The proposal then consists of the n samples of θ and its distributional hyperparameters μ_θ and σ_θ^2 . Consequently, the posterior of the parameters, given the data $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ and the hyperpriors, is

$$P(\theta, \mu_\theta, \sigma_\theta^2 \mid \mathbf{X}, \nu_\mu, \Sigma_\mu^2, a_\sigma, b_\sigma) \propto \prod_{i=1}^n \left[P_\theta(\theta_i \mid \mu_\theta, \sigma_\theta^2) L_M(\theta_i \mid \mathbf{X}_i) \right] \times P_\mu(\mu_\theta \mid \nu_\mu, \Sigma_\mu^2) P_\sigma(\sigma_\theta^2 \mid a_\sigma, b_\sigma), \quad (5.4)$$

where L_M is the participant-level likelihood (see also Equations 1.3 and 1.7), P_θ is the participant-level Gaussian prior conditional on the sampled hyperparameters μ_θ and σ_θ^2 , P_μ is the Gaussian hyperprior for μ_θ , and P_σ is the inverse-gamma hyperprior for σ_θ^2 .

Note that this approach assumes that hierarchical priors are normally distributed. Although there is no theoretical constraint to a Gaussian distribution, it has at least the following four practical reasons:

1. The hyperparameters μ_θ and σ_θ^2 , which are fitted in addition to the participant-level parameters θ , are intuitively interpretable as empirical characteristics of θ across participants, namely the mean and variance (or relatedly, the standard deviation).
2. While analytical probability density functions exist for numerous other distributions, it has to be ensured that the sampled hyperparameters represent the empirical (not population-level) distributional characteristics of θ . That means, that μ_θ must match exactly the mean and σ_θ^2 must match exactly the variance of the sampled θ values. For non-Gaussian distributions, there are few if any tractable and numerically stable solutions for such purposes.
3. Using a Gaussian distribution for more than one hierarchical parameters permits the

use of a covariance matrix instead of a scalar variance parameter for σ_{θ}^2 . This permits sampling and estimation of parameter correlations across participants.

4. Generally, if parameters are to have finite lower and/or upper bounds, the sampled θ can be transformed using analytical, monotonic functions, for example

$$\exp : \mathbb{R} \rightarrow \{\theta' \in \mathbb{R} : 0 \leq \theta' \leq +\infty\} \text{ or} \quad (5.5)$$

$$\tanh : \mathbb{R} \rightarrow \{\theta' \in \mathbb{R} : -1 \leq \theta' \leq 1\} . \quad (5.6)$$

Given the individual differences and high dimensionality of SWIFT and SEAM, the hierarchical approach may be particularly useful for fitting those models to appropriate experimental data. Presumably, as a first step, a proof-of-concept hierarchical fitting of the simplified toySWIFT (Engbert & Rabe, 2023) would be a powerful demonstration that should be considered for future research.

5.9 Conclusion

In summary, the work presented in this dissertation offers valuable insights into the integration of likelihood-based Bayesian inference, interindividual differences in reading behavior, and the dynamic interaction between eye-movement control and linguistic processing. Further research and validation are needed to fully explore the potential of these models and methodologies in advancing our understanding of reading processes. The SEAM model in particular affords more validation with regard to other syntactic phenomena, e.g., more specific effects such as local coherence effects or classical effects such as garden-path effects. Moreover, further research should address the methodological issues of model comparison, accessibility, and hierarchical modeling, which would be major advancements in the field of reading models and, consequently, in our understanding of the fundamental processes involved during reading.

References

- Adeli, H., Vitu, F., & Zelinsky, G. J. (2016). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *The Journal of Neuroscience*, *37*(6), 1453–1467. <https://doi.org/10.1523/jneurosci.0825-16.2016>
- Aerila, J.-A., & Merisuo-Storm, T. (2017). Emergent readers and the joy of reading: A Finnish perspective. *Creative Education*, *08*(15), 2485–2500. <https://doi.org/10.4236/ce.2017.815171>
- Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87. <https://doi.org/10.1007/BF00337259>
- Amari, S. V., & Misra, R. B. (1997). Closed-form expressions for distribution of sum of exponential random variables. *IEEE Transactions on Reliability*, *46*(4), 519–522.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates.
- Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*(3), 313–341. https://doi.org/10.1207/s15516709cog0000_22
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. <https://doi.org/10.1037/0033-295X.111.4.1036>
- Anderson, J. R. (1990). *Cognitive psychology and its implications*. Freeman.
- Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461. <https://doi.org/10.3758/bf03214334>
- Andrieu, C., & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, *37*(2), 697–725. <https://doi.org/10.1214/07-aos574>
- Ankenier, C. S., Sekicki, M., & Staudte, M. (2018). The influence of visual uncertainty on word surprisal and processing effort. *Frontiers in Psychology*, *9*. <https://doi.org/10.3389/fpsyg.2018.02387>
- Apgar, J. F., Witmer, D. K., White, F. M., & Tidor, B. (2010). Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular BioSystems*, *6*(10), 1890. <https://doi.org/10.1039/b918098b>

- Azuma, M., Yaoi, K., Minamoto, T., Osaka, M., & Osaka, N. (2014). Effect of memory load on eye movement control: A study using the reading span test. *Journal of Eye Movement Research*, 7. <https://doi.org/10.16910/jemr.7.5.3>
- Balling, L. W., & Kizach, J. (2017). Effects of surprisal and locality on Danish sentence processing: An eye-tracking investigation. *Journal of Psycholinguistic Research*, 46(5), 1119–1136. <https://doi.org/10.1007/s10936-017-9482-2>
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3), 445–459. <https://doi.org/10.3758/bf03193014>
- Barredo, E. A., Viray, M. C., & Galimba, S. M. B.-o. (2022). Reading ability of grade 7 students in Rebokon Agricultural and Vocational High School and its effect to their academic performance. *International Journal of Secondary Education*, 10(2), 74. <https://doi.org/10.11648/j.ijsedu.20221002.12>
- Becker, W., & Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. *Vision Research*, 19(9), 967–983. [https://doi.org/10.1016/0042-6989\(79\)90222-0](https://doi.org/10.1016/0042-6989(79)90222-0)
- Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 4(3), 91–99. [https://doi.org/10.1016/s1364-6613\(99\)01440-0](https://doi.org/10.1016/s1364-6613(99)01440-0)
- Bélisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on Rd. *Journal of Applied Probability*, 29(4), 885–895. <https://doi.org/10.2307/3214721>
- Boakye, N. A. (2017). Extensive reading in a tertiary reading programme: Students' accounts of affective and cognitive benefits. *Reading & Writing*, 8(1). <https://doi.org/10.4102/rw.v8i1.153>
- Boehm, U., Evans, N. J., Gronau, Q. F., Matzke, D., Wagenmakers, E.-J., & Heathcote, A. J. (2023). Inclusion Bayes factors for mixed hierarchical diffusion decision models. *Psychological Methods*. <https://doi.org/10.1037/met0000582>
- Boston, M. F., Hale, J. T., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1), 1–12. <https://doi.org/10.16910/jemr.2.1.1>
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. <https://doi.org/10.3758/brm.41.4.977>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>

- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. <https://doi.org/10.32614/rj-2018-017>
- Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. <https://doi.org/10.18637/jss.v100.i05>
- Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432–459. <https://doi.org/10.1037/0033-295x.100.3.432>
- Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. <https://doi.org/10.1137/0916069>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan. *Journal of Statistical Software*, *76*(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Chandra, J., Krügel, A., & Engbert, R. (2020). Modulation of oculomotor control during reading of mirrored and inverted texts. *Scientific Reports*, *10*, Article 4210. <https://doi.org/10.1038/s41598-020-60833-6>
- Chasteen, A. L., & Pratt, J. (1999). The effect of inhibition of return on lexical access. *Psychological Science*, *10*(1), 41–46. <https://doi.org/10.1111/1467-9280.00104>
- Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? Evidence from garden-path and local coherence structures. *Quarterly Journal of Experimental Psychology*, *70*(7), 1380–1405.
- Clifton, C., Staub, A., & Rayner, K. (2007). Eye Movements in Reading Words and Sentences. In R. V. Gompel, M. Fisher, W. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain*. Elsevier.
- Coelho, C. A. (1998). The generalized integer gamma distribution—a basis for distributions in multivariate statistics. *Journal of Multivariate Analysis*, *64*(1), 86–102. <https://doi.org/10.1006/jmva.1997.1710>
- Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornič (Ed.), *Attention and performance VI* (pp. 535–555). Routledge.
- Daily, L. Z., Lovett, M. C., & Reder, L. M. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science*, *25*(3), 315–353. https://doi.org/10.1207/s15516709cog2503_1
- Daneman, M., Reingold, E. M., & Davidson, M. (1995). Time course of phonological activation during reading: Evidence from eye fixations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 884–898. <https://doi.org/10.1037/0278-7393.21.4.884>
- Ditchburn, R. W., & Ginsborg, B. L. (1952). Vision with a stabilized retinal image. *Nature*, *170*(4314), 36–37. <https://doi.org/10.1038/170036a0>

- Dotlačil, J. (2018). Building an ACT-R reader for eye-tracking corpus data. *Topics in Cognitive Science*, *10*(1), 144–160. <https://doi.org/10.1111/tops.12315>
- Dotlačil, J. (2021). Parsing as a cue-based retrieval model. *Cognitive Science*, *45*(8), Article e13020. <https://doi.org/10.1111/cogs.13020>
- Engbert, R. (2021). *Dynamical Models in Neurocognitive Psychology*. Springer Nature Publishing. <https://doi.org/10.1007/978-3-030-67299-7>
- Engbert, R., & Kliegl, R. (2001). Mathematical models of eye movements in reading: A possible role for autonomous saccades. *Biological Cybernetics*, *85*(2), 77–87. <https://doi.org/10.1007/PL00008001>
- Engbert, R., & Kliegl, R. (2011). Parallel graded attention models of reading. In S. P. Liv-ersedge, I. D. Gilchrist, & S. Everling (Eds.), *Oxford Handbook of Eye Movements* (pp. 787–800). Oxford University Press.
- Engbert, R., & Krügel, A. (2010). Readers use Bayesian estimation for eye movement control. *Psychological Science*, *21*(3), 366–371. <https://doi.org/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. [https://doi.org/10.1016/s0042-6989\(01\)00301-7](https://doi.org/10.1016/s0042-6989(01)00301-7)
- Engbert, R., & Nuthmann, A. (2008). Self-consistent estimation of mislocated fixations during reading. *PLoS One*, *3*(2), Article e1534. <https://doi.org/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. <https://doi.org/10.1037/0033-295x.112.4.777>
- Engbert, R., & Rabe, M. M. (2023). Tutorial on dynamical modeling of eye movements in reading. <https://doi.org/10.31234/osf.io/dsvmt>
- Engbert, R., Rabe, M. M., Schwetlick, L., Seelig, S. A., Reich, S., & Vasishth, S. (2022). Data assimilation in dynamical cognitive science. *Trends in Cognitive Sciences*, *26*(2), 99–102. <https://doi.org/10.1016/j.tics.2021.11.006>
- Engbert, R., Sinn, P., Mergenthaler, K., & Trukenbrod, H. (2015). Microsaccade toolbox. <http://read.psych.uni-potsdam.de/>
- Engelmann, F. (2015). ACTR-in-R [GitHub repository]. <https://github.com/felixengelmann/ACTR-in-R/tree/ee01519>
- Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, *43*(12), Article e12800. <https://doi.org/10.1111/cogs.12800>

- 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. <https://doi.org/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. <https://doi.org/10.1137/1114019>
- Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, 109(3), 545–572. <https://doi.org/10.1037/0033-295x.109.3.545>
- Eskenazi, M. A., & Folk, J. R. (2016). Regressions during reading: The cost depends on the cause. *Psychonomic Bulletin Review*, 24(4), 1211–1216. <https://doi.org/10.3758/s13423-016-1200-9>
- Findlay, J. M., & Gilchrist, I. D. (2003). *Active Vision: The Psychology of Looking and Seeing*. Oxford University Press.
- 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. <https://doi.org/10.1017/s0140525x99002150>
- Fletcher, R. (1964). Function minimization by conjugate gradients. *The Computer Journal*, 7(2), 149–154. <https://doi.org/10.1093/comjnl/7.2.149>
- Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies* [Doctoral dissertation, University of Massachusetts].
- Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2), 178–210. [https://doi.org/10.1016/0010-0285\(82\)90008-1](https://doi.org/10.1016/0010-0285(82)90008-1)
- Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, 26(5), 505–526. [https://doi.org/10.1016/0749-596x\(87\)90137-9](https://doi.org/10.1016/0749-596x(87)90137-9)
- Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, 40(3), 575–586. <https://doi.org/10.2307/2530902>
- Fum, D., Missier, F. D., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, 8(3), 135–142. <https://doi.org/10.1016/j.cogsys.2007.07.001>
- Gardiner, C. (1985). *Handbook of Stochastic Processes*. Springer-Verlag, New York.
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997–1016. <https://doi.org/10.1007/s11222-013-9416-2>
- Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman; Hall/CRC.

- Georgiou, G. K., & Das, J. (2016). What component of executive functions contributes to normal and impaired reading comprehension in young adults? *Research in Developmental Disabilities, 49-50*, 118–128. <https://doi.org/10.1016/j.ridd.2015.12.001>
- Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.). Continuum Press.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76. [https://doi.org/10.1016/s0010-0277\(98\)00034-1](https://doi.org/10.1016/s0010-0277(98)00034-1)
- Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O’Neil (Eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*. MIT Press.
- Gilchrist, I. (2011). Saccades. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The oxford handbook of eye movements* (pp. 86–94). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199539789.013.0005>
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. Chapman; Hall/CRC.
- Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics, 22*(4), 403–434. [https://doi.org/10.1016/0021-9991\(76\)90041-3](https://doi.org/10.1016/0021-9991(76)90041-3)
- Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(6), 1304–1321. <https://doi.org/10.1037/0278-7393.32.6.1304>
- Haken, H., Kelso, J. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51*(5), 347–356. <https://doi.org/10.1007/bf00336922>
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97–109. <https://doi.org/10.1093/biomet/57.1.97>
- Heister, J., Würzner, K.-M., Bubbenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). DlexDB—eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau, 62*(1), 10–20. <https://doi.org/10.1026/0033-3042/a000029>
- Henderson, J. M., & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 417–429. <https://doi.org/10.1037/0278-7393.16.3.417>
- Holcomb, P. J., Grainger, J., & O’Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of cognitive neuroscience, 14*, 938–950. <https://doi.org/10.1162/089892902760191153>

- Holmes, W. R. (2015). A practical guide to the probability density approximation (PDA) with improved implementation and error characterization. *Journal of Mathematical Psychology*, 68-69, 13–24. <https://doi.org/10.1016/j.jmp.2015.08.006>
- Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2023). *Surprisal does not explain syntactic disambiguation difficulty: Evidence from a large-scale benchmark*. <https://doi.org/10.31234/osf.io/z38u6>
- Huestegge, L., & Bocianski, D. (2010). Effects of syntactic context on eye movements during reading. *Advances in Cognitive Psychology*, 6(-1), 79–87. <https://doi.org/10.2478/v10053-008-0078-0>
- Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, 40(6), 431–439. <https://doi.org/10.3758/bf03208203>
- Inhoff, A. W., & Topolski, R. (1994). Use of phonological codes during eye fixations in reading and in on-line and delayed naming tasks. *Journal of Memory and Language*, 33(5), 689–713. <https://doi.org/10.1006/jmla.1994.1033>
- Inhoff, A. W., & Weger, U. W. (2005). Memory for word location during reading: Eye movements to previously read words are spatially selective but not precise. *Memory & Cognition*, 33(3), 447–461.
- Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., & Vasisht, S. (2015). Teasing apart retrieval and encoding interference in the processing of anaphors. *Frontiers in Psychology*, 6, Article 506. <https://doi.org/10.3389/fpsyg.2015.00506>
- Jäger, L. A., Engelmann, F., & Vasisht, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, 94, 316–339. <https://doi.org/10.1016/j.jml.2017.01.004>
- Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasisht, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, 111, Article 104063. <https://doi.org/10.1016/j.jml.2019.104063>
- Javal, L. É. (1878). Essai sur la physiologie de la lecture. *Annales d'Oculistique*, 79, 97–117.
- Jones, A. C., Folk, J. R., & Brusnighan, S. M. (2012). Resolving syntactic category ambiguity: An eye-movement analysis. *Journal of Cognitive Psychology*, 24(6), 672–688. <https://doi.org/10.1080/20445911.2012.679925>
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329–354. <https://doi.org/10.1037/0033-295X.87.4.329>
- Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2), 262–284. <https://doi.org/10.1080/09541440340000213>

- 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. <https://doi.org/10.1037/0096-3445.135.1.12>
- Kolers, P. A. (1976). Reading a year later. *Journal of Experimental Psychology*, *2*(5), 554–565. <https://doi.org/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. [https://doi.org/10.1016/0010-0285\(75\)90011-0](https://doi.org/10.1016/0010-0285(75)90011-0)
- Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1648–1662. <https://doi.org/10.1037/xlm0000128>
- Krügel, A., & Engbert, R. (2010). The launch-site effect for skipped words during reading. *Vision Research*, *50*, 1532–1539. <https://doi.org/10.1016/j.visres.2010.05.009>
- 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. <https://doi.org/10.1080/13506285.2014.894166>
- Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
- Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2018). Saliency benchmarking made easy: Separating models, maps and metrics. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision – ECCV 2018* (pp. 798–814). Springer International Publishing.
- Lago, S., Acuña Fariña, C., & Meseguer, E. (2021). The reading signatures of agreement attraction. *Open Mind*, *5*, 132–153. https://doi.org/10.1162/opmi_a_00047
- 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), Article W01526. <https://doi.org/10.1029/2011wr010608>
- Land, M. F. (2011, August). Oculomotor behaviour in vertebrates and invertebrates. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The oxford handbook of eye movements* (pp. 2–15). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199539789.013.0001>
- Law, K., Stuart, A., & Zygalakis, K. (2015). *Data Assimilation*. Springer.
- Lee, Y., Lee, H., & Gordon, P. C. (2007). Linguistic complexity and information structure in Korean: Evidence from eye-tracking during reading. *Cognition*, *104*(3), 495–534. <https://doi.org/10.1016/j.cognition.2006.07.013>
- Levy, R. P. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. <https://doi.org/10.1016/j.cognition.2007.05.006>

- Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989–2001. <https://doi.org/10.1016/j.jmva.2009.04.008>
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*(3), 375–419. https://doi.org/10.1207/s15516709cog0000_25
- Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10*(10), 447–454. <https://doi.org/10.1016/j.tics.2006.08.007>
- Locascio, G., Mahone, E. M., Eason, S. H., & Cutting, L. E. (2010). Executive dysfunction among children with reading comprehension deficits. *Journal of Learning Disabilities, 43*(5), 441–454. <https://doi.org/10.1177/0022219409355476>
- Luce, R. D., & Raiffa, H. (1989). *Games and decisions: Introduction and critical survey*. Courier Corporation.
- Luke, S. G., Darowski, E. S., & Gale, S. D. (2018). Predicting eye-movement characteristics across multiple tasks from working memory and executive control. *Memory & Cognition, 46*, 826–839. <https://doi.org/10.3758/s13421-018-0798-4>
- Luke, S. G., & Henderson, J. M. (2013). Oculomotor and cognitive control of eye movements in reading: Evidence from mindless reading. *Attention, Perception, & Psychophysics, 75*(6), 1230–1242. <https://doi.org/10.3758/s13414-013-0482-5>
- Luke, S. G., & Henderson, J. M. (2016). The influence of content meaningfulness on eye movements across tasks: Evidence from scene viewing and reading. *Frontiers in Psychology, 7*. <https://doi.org/10.3389/fpsyg.2016.00257>
- Marin, J.-M., & Robert, C. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer Science & Business Media.
- Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review, 101*(1), 3–33. <https://doi.org/10.1037/0033-295x.101.1.3>
- Martinez-Conde, S., & Macknik, S. L. (2011, August). Microsaccades. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The oxford handbook of eye movements* (pp. 96–114). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199539789.013.0006>
- Martinez-Conde, S., Macknik, S. L., Troncoso, X. G., & Hubel, D. H. (2009). Microsaccades: a neurophysiological analysis. *Trends in Neurosciences, 32*(9), 463–475. <https://doi.org/10.1016/j.tins.2009.05.006>
- Masson, M. E. J., Rabe, M. M., & Kliegl, R. (2017). Modulation of additive and interactive effects by trial history revisited. *Memory & Cognition, 45*(3), 480–492. <https://doi.org/10.3758/s13421-016-0666-z>

- Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, 81(12), 899–917. <https://doi.org/10.1037/h0037368>
- Mätzig, P., Vasishth, S., Engelmann, F., Caplan, D., & Burchert, F. (2018). A computational investigation of sources of variability in sentence comprehension difficulty in aphasia. *Topics in Cognitive Science*, 10(1), 161–174. <https://doi.org/10.1111/tops.12323>
- McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of initial eye fixations on words. *Vision Research*, 28(10), 1107–1118. [https://doi.org/10.1016/0042-6989\(88\)90137-x](https://doi.org/10.1016/0042-6989(88)90137-x)
- McCutchen, D., & Perfetti, C. A. (1982). The visual tongue-twister effect: Phonological activation in silent reading. *Journal of Verbal Learning and Verbal Behavior*, 21(6), 672–687. [https://doi.org/10.1016/s0022-5371\(82\)90870-2](https://doi.org/10.1016/s0022-5371(82)90870-2)
- Mertzen, D. (2022). *A cross-linguistic investigation of similarity-based interference in sentence comprehension* [Doctoral dissertation]. Universität Potsdam. <https://doi.org/10.25932/publishup-55668>
- Mertzen, D., Paape, D., Dillon, B., Engbert, R., & Vasishth, S. (2023). Syntactic and semantic interference in sentence comprehension: Support from English and German eye-tracking data. *Glossa Psycholinguistics*, 2(1). <https://doi.org/10.5070/g60111266>
- Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition*, 30(4), 551–561.
- 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. <https://doi.org/10.1063/1.1699114>
- Meyer, D., & Pietzner, V. (2022). Reading textual and non-textual explanations in chemistry texts and textbooks – a review. *Chemistry Education Research and Practice*, 23(4), 768–785. <https://doi.org/10.1039/d2rp00162d>
- Miller, G. A. (1962). Some psychological studies of grammar. *American Psychologist*, 17(11), 748–762. <https://doi.org/10.1037/h0044708>
- Miller, G. A., & Isard, S. (1964). Free recall of self-embedded English sentences. *Information and Control*, 7(3), 292–303. [https://doi.org/10.1016/s0019-9958\(64\)90310-9](https://doi.org/10.1016/s0019-9958(64)90310-9)
- Mitchell, D. C., Shen, X., Green, M. J., & Hodgson, T. L. (2008). Accounting for regressive eye-movements in models of sentence processing: A reappraisal of the selective reanalysis hypothesis. *Journal of Memory and Language*, 59(3), 266–293. <https://doi.org/10.1016/j.jml.2008.06.002>
- Morris, R. K., Rayner, K., & Pollatsek, A. (1990). Eye movement guidance in reading: The role of parafoveal letter and space information. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2), 268–281. <https://doi.org/10.1037/0096-1523.16.2.268>

- Morzfeld, M., & Reich, S. (2018). Data assimilation: Mathematics for merging models and data. *Snapshots of modern mathematics from Oberwolfach*, (11). <https://doi.org/10.14760/SNAP-2018-011-EN>
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1), 90–100. [https://doi.org/10.1016/s0022-2496\(02\)00028-7](https://doi.org/10.1016/s0022-2496(02)00028-7)
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4), 308–313. <https://doi.org/10.1093/comjnl/7.4.308>
- Nicenboim, B., Schad, D. J., & Vasishth, S. (2023). *Introduction to Bayesian data analysis for cognitive science* [Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series]. <https://vasishth.github.io/bayescogsci/>
- Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, 142, Article 107427. <https://doi.org/10.1016/j.neuropsychologia.2020.107427>
- 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. <https://doi.org/10.1016/j.visres.2008.10.022>
- Nuthmann, A., Engbert, R., & Kliegl, R. (2005). Mislocated fixations during reading and the inverted optimal viewing position effect. *Vision Research*, 45(17), 2201–2217. <https://doi.org/10.1016/j.visres.2005.02.014>
- Nuthmann, A., Engbert, R., & Kliegl, R. (2007). The IOVP effect in mindless reading: Experiment and modeling. *Vision Research*, 47(7), 990–1002. <https://doi.org/10.1016/j.visres.2006.11.005>
- O'Regan, J. K. (1979). Saccade size control in reading: Evidence for the linguistic control hypothesis. *Perception & Psychophysics*, 25(6), 501–509. <https://doi.org/10.3758/bf03213829>
- O'Regan, J. K. (1990). Eye movements and reading. *Reviews of oculomotor research*, 4, 395–453. <https://doi.org/10.1177/0956797610378686>
- Paape, D., Vasishth, S., Paape, D., & Vasishth, S. (2022). Is reanalysis selective when regressions are consciously controlled? *Glossa Psycholinguistics*, 1(1).
- Palestro, J. J., Sederberg, P. B., Osth, A. F., Zandt, T. V., & Turner, B. M. (2018). *Likelihood-free methods for cognitive science*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-72425-6>
- Parker, A. J., Kirkby, J. A., & Slattery, T. J. (2020). Undersweep fixations during reading in adults and children. *Journal of Experimental Child Psychology*, 192, 104788. <https://doi.org/10.1016/j.jecp.2019.104788>

- Paterson, K. B., Liversedge, S. P., & Davis, C. J. (2009). Inhibitory neighbor priming effects in eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 43–50. <https://doi.org/10.3758/pbr.16.1.43>
- Perea, M., & Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 767–779. <https://doi.org/10.1037/0096-1523.24.3.767>
- Pollatsek, A., Hyönä, J., & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 820–833. <https://doi.org/10.1037/0096-1523.26.2.820>
- Pollatsek, A., Perea, M., & Binder, K. S. (1999). The effects of "neighborhood size" in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 1142–1158. <https://doi.org/10.1037/0096-1523.25.4.1142>
- 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. <https://doi.org/10.1016/j.cogpsych.2005.06.001>
- Pollatsek, A., Tan, L. H., & Rayner, K. (2000). The role of phonological codes in integrating information across saccadic eye movements in chinese character identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 607–633. <https://doi.org/10.1037/0096-1523.26.2.607>
- Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and performance*, *32*, 531–556.
- Rabe, M. M., Chandra, J., Krügel, A., Seelig, S. A., Vasishth, S., & Engbert, R. (2021). A Bayesian approach to dynamical modeling of eye-movement control in reading of normal, mirrored, and scrambled texts. *Psychological Review*, *128*(5), 803–823. <https://doi.org/10.1037/rev0000268>
- Rabe, M. M., Paape, D., Mertzen, D., Vasishth, S., & Engbert, R. (2023). *SEAM: An integrated activation-coupled model of sentence processing and eye movements in reading*. <https://doi.org/10.48550/arXiv.2303.05221>
- 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. <https://doi.org/10.21105/joss.02134>
- Rabovsky, M., Conrad, M., Álvarez, C. J., Paschke-Goldt, J., & Sommer, W. (2019). Attentional modulation of orthographic neighborhood effects during reading: Evidence from event-related brain potentials in a psychological refractory period paradigm. *PLOS ONE*, *14*(1), e0199084. <https://doi.org/10.1371/journal.pone.0199084>

- Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes, 23*(7-8), 942–971. <https://doi.org/10.1080/01690960802069730>
- Rayner, K. (1978). Eye movements in reading and information processing. *Psychological bulletin, 85*, 618–660.
- Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology, 7*(1), 65–81. [https://doi.org/10.1016/0010-0285\(75\)90005-5](https://doi.org/10.1016/0010-0285(75)90005-5)
- Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception, 8*(1), 21–30. <https://doi.org/10.1068/p080021>
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422. <https://doi.org/10.1037/0033-2909.124.3.372>
- Rayner, K. (2009). Eye movements in reading: Models and data. *Journal of eye movement research, 2*, 1–10.
- Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 720–732. <https://doi.org/10.1037/0096-1523.30.4.720>
- Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior, 22*(3), 358–374. [https://doi.org/10.1016/s0022-5371\(83\)90236-0](https://doi.org/10.1016/s0022-5371(83)90236-0)
- Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*(3), 191–201. <https://doi.org/10.3758/bf03197692>
- Rayner, K., Juhasz, B., Ashby, J., & Clifton, C. (2003). Inhibition of saccade return in reading. *Vision Research, 43*(9), 1027–1034. [https://doi.org/10.1016/s0042-6989\(03\)00076-2](https://doi.org/10.1016/s0042-6989(03)00076-2)
- Rayner, K., & Morris, R. K. (1992). Eye movement control in reading: Evidence against semantic preprocessing. *Journal of Experimental Psychology: Human Perception and Performance, 18*(1), 163–172. <https://doi.org/10.1037/0096-1523.18.1.163>
- Rayner, K., Pollatsek, A., & Binder, K. S. (1998). Phonological codes and eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 476–497. <https://doi.org/10.1037/0278-7393.24.2.476>
- Rayner, K., & Reichle, E. D. (2010). Models of the reading process. *WIREs Cognitive Science, 1*(6), 787–799. <https://doi.org/https://doi.org/10.1002/wcs.68>

- Rayner, K., Well, A. D., & Pollatsek, A. (1980). Asymmetry of the effective visual field in reading. *Perception & Psychophysics*, *27*(6), 537–544. <https://doi.org/10.3758/bf03198682>
- Rayner, K., White, S. J., Johnson, R. L., & Liversedge, S. P. (2006). Reading words with jumbled letters. *Psychological Science*, *17*(3), 192–193. <https://doi.org/10.1111/j.1467-9280.2006.01684.x>
- Reich, S., & Cotter, C. (2015). *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press.
- Reichle, E. D. (2011). Serial-attention models of reading. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The oxford handbook of eye movements* (pp. 767–786). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199539789.013.0042>
- Reichle, E. D. (2021). *Computational models of reading*. Oxford University Press.
- 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. <https://doi.org/10.1037/0033-295x.105.1.125>
- 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. <https://doi.org/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. [https://doi.org/10.1016/S0042-6989\(99\)00152-2](https://doi.org/10.1016/S0042-6989(99)00152-2)
- Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–476. <https://doi.org/10.1017/s0140525x03000104>
- Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science*, *21*(9), 1300–1310. <https://doi.org/10.1177/0956797610378686>
- Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21. <https://doi.org/10.3758/pbr.16.1.1>
- Reilly, R., & Radach, R. (2002). Glenmore: An interactive activation model of eye movement control in reading. *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02*. <https://doi.org/10.1109/iconip.2002.1202810>
- 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. <https://doi.org/10.1016/j.cogsys.2005.07.006>

- Risse, S. (2014). Effects of visual span on reading speed and parafoveal processing in eye movements during sentence reading. *Journal of Vision*, *14*(8), Article 11. <https://doi.org/10.1167/14.8.11>
- Risse, S., & Seelig, S. (2019). Stable preview difficulty effects in reading with an improved variant of the boundary paradigm. *Quarterly Journal of Experimental Psychology*, *72*(7), 1632–1645. <https://doi.org/10.1177/1747021818819990>
- Robert, C., & Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Science & Business Media.
- Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120. <https://doi.org/10.1214/aoap/1034625254>
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. <https://doi.org/10.1037/0033-295X.107.2.358>
- Ross, J., Morrone, M., Goldberg, M. E., & Burr, D. C. (2001). Changes in visual perception at the time of saccades. *Trends in Neurosciences*, *24*(2), 113–121. [https://doi.org/10.1016/s0166-2236\(00\)01685-4](https://doi.org/10.1016/s0166-2236(00)01685-4)
- Rusich, D., Arduino, L. S., Mauti, M., Martelli, M., & Primativo, S. (2020). Evidence of semantic processing in parafoveal reading: A rapid parallel visual presentation (RPVP) study. *Brain Sciences*, *11*, 28. <https://doi.org/10.3390/brainsci11010028>
- Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, *1*(4), 201–220. [https://doi.org/10.1016/S1389-0417\(00\)00015-2](https://doi.org/10.1016/S1389-0417(00)00015-2)
- Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126. <https://doi.org/10.1037/met0000275>
- Schad, D. J., & Engbert, R. (2012). The zoom lens of attention: Simulating shuffled versus normal text reading using the SWIFT model. *Visual Cognition*, *20*(4-5), 391–421. <https://doi.org/10.1080/13506285.2012.670143>
- 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*, Article 104038. <https://doi.org/10.1016/j.jml.2019.104038>
- Schilling, H., Rayner, K., & Chumbley, J. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory and Cognition*, *26*(6), 1270–1281. <https://doi.org/10.3758/BF03201199>
- Schotter, E. R., Tran, R., & Rayner, K. (2014). Don't believe what you read (only once) comprehension is supported by regressions during reading. *Psychological Science*, *25*(6), 1218–1226.

- 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. <https://doi.org/10.1037/rev0000068>
- Schwetlick, L., Backhaus, D., & Engbert, R. (2022). A dynamical scan-path model for task-dependence during scene viewing. *Psychological Review*. <https://doi.org/10.1037/rev0000379>
- Schwetlick, L., Rothkegel, L. O. M., Trukenbrod, H. A., & Engbert, R. (2020). Modeling the effects of perisaccadic attention on gaze statistics during scene viewing. *Communications Biology*, *3*, Article 727. <https://doi.org/10.1038/s42003-020-01429-8>
- Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.
- 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*, Article 102313. <https://doi.org/10.1016/j.jmp.2019.102313>
- Sereno, S. C., & Rayner, K. (2000). Spelling-sound regularity effects on eye fixations in reading. *Perception & Psychophysics*, *62*(2), 402–409. <https://doi.org/10.3758/bf03205559>
- Shockley, E. (2019). PyDREAM: A Python implementation of the MT-DREAM(ZS) algorithm from Laloy and Vrugt 2012 [GitHub repository]. <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. <https://doi.org/10.1080/01621459.1967.10482935>
- Sisson, S. A., & Fan, Y. (2011). Likelihood-free Markov chain Monte Carlo, 313–335.
- Slattery, T. J., & Parker, A. J. (2019). Return sweeps in reading: Processing implications of undersweep-fixations. *Psychonomic Bulletin & Review*, *26*(6), 1948–1957. <https://doi.org/10.3758/s13423-019-01636-3>
- Snell, J., & Grainger, J. (2019). Readers are parallel processors. *Trends in Cognitive Sciences*, *23*(7), 537–546. <https://doi.org/10.1016/j.tics.2019.04.006>
- Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological Review*, *125*(6), 969–984. <https://doi.org/10.1037/rev0000119>
- Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, *12*(3), 175–200. <https://doi.org/10.20982/tqmp.12.3.p175>

- Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 357–416. <https://doi.org/10.2307/2983527>
- Starr, M. S., & Rayner, K. (2001). Eye movements during reading: Some current controversies. *Trends in Cognitive Sciences*, *5*(4), 156–163. [https://doi.org/10.1016/S1364-6613\(00\)01619-3](https://doi.org/10.1016/S1364-6613(00)01619-3)
- Stock, L., Krüger-Zechlin, C., Deeb, Z., Timmermann, L., & Waldthaler, J. (2020). Natural reading in Parkinson's disease with and without mild cognitive impairment. *Frontiers in Aging Neuroscience*, *12*. <https://doi.org/10.3389/fnagi.2020.00120>
- Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*, *36*(1), 201–216. <https://doi.org/10.3758/MC.36.1.201>
- ter Braak, C. J. (2006). A Markov Chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. *Statistics and Computing*, *16*(3), 239–249. <https://doi.org/10.1007/s11222-006-8769-1>
- ter Braak, C. J., & Vrugt, J. A. (2008). Differential evolution Markov chain with snooker updater and fewer chains. *Statistics and Computing*, *18*(4), 435–446.
- Turner, B. M., & Sederberg, P. B. (2013). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, *21*(2), 227–250. <https://doi.org/10.3758/s13423-013-0530-0>
- Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 407–430. <https://doi.org/10.1037/0278-7393.33.2.407>
- Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(5), 615–628. <https://doi.org/10.1017/s0140525x98001733>
- Van Kampen, N. G. (1992). *Stochastic Processes in Physics and Chemistry* (Vol. 1). Elsevier.
- van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. <https://doi.org/10.1080/17470218.2013.850521>
- Vasilev, M. R., & Angele, B. (2017). Parafoveal preview effects from word $n + 1$ and word $n + 2$ during reading: A critical review and Bayesian meta-analysis. *Psychonomic Bulletin & Review*, *24*(3), 666–689. <https://doi.org/10.3758/s13423-016-1147-x>
- Vasishth, S., & Engelmann, F. (2022). *Sentence comprehension as a cognitive process: A computational approach* [GitHub repository]. <https://vasishth.github.io/RetrievalModels/>

- Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, 23(11), 968–982. <https://doi.org/10.1016/j.tics.2019.09.003>
- Veldre, A., Yu, L., Andrews, S., & Reichle, E. D. (2020). Towards a complete model of reading: Simulating lexical decision, word naming, and sentence reading with Über-Reader. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. <https://hdl.handle.net/2123/22990>
- Vihola, M. (2012). Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5), 997–1008. <https://doi.org/10.1007/s11222-011-9269-5>
- 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. [https://doi.org/10.1016/s0042-6989\(01\)00166-3](https://doi.org/10.1016/s0042-6989(01)00166-3)
- Vitu, F., O'Regan, J. K., Inhoff, A. W., & Topolski, R. (1995). Mindless reading: Eye-movement characteristics are similar in scanning letter strings and reading texts. *Perception & Psychophysics*, 57(3), 352–364. <https://doi.org/10.3758/bf03213060>
- von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, 65, 109–127. <https://doi.org/10.1016/j.jml.2011.02.004>
- von der Malsburg, T., & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, 28(10), 1545–1578. <https://doi.org/10.1080/01690965.2012.728232>
- von Helmholtz, H. (1896). *Handbuch der physiologischen Optik* (2nd ed.). Leopold Voss.
- 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). <https://doi.org/10.1515/ijnsns.2009.10.3.273>
- Warren, T., White, S. J., & Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z Reader. *Cognition*, 111(1), 132–137. <https://doi.org/10.1016/j.cognition.2008.12.011>
- Weger, U. W., & Inhoff, A. W. (2007). Long-range regressions to previously read words are guided by spatial and verbal memory. *Memory & Cognition*, 35(6), 1293–1306.
- Weiss, A. F., Kretschmar, F., Schlesewsky, M., Bornkessel-Schlesewsky, I., & Staub, A. (2018). Comprehension demands modulate re-reading, but not first-pass reading behavior. *Quarterly Journal of Experimental Psychology*, 71(1), 198–210. <https://doi.org/10.1080/17470218.2017.1307862>
- Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310), 1102–1104.

- Yadav, H., Paape, D., Smith, G., Dillon, B. W., & Vasishth, S. (2022). Individual differences in cue weighting in sentence comprehension: An evaluation using Approximate Bayesian Computation. *Open Mind*, 6, 1–24. https://doi.org/10.1162/opmi_a_00052
- Yarbus, A. L. (2013). *Eye movements and vision*. Springer.
- Zhang, M., Liversedge, S. P., Bai, X., Yan, G., & Zang, C. (2019). The influence of foveal lexical processing load on parafoveal preview and saccadic targeting during chinese reading. *Journal of Experimental Psychology: Human Perception and Performance*, 45, 812–825. <https://doi.org/10.1037/xhp0000644>

Appendix A

Experimental Data and Sentence Material

All eye-tracking data used in our simulation studies originate from Risse and Seelig (2019), who collected data for an experiment that was a version of the $n + 1$ boundary paradigm (Rayner, 1975) to investigate effects of parafoveal word difficulty on fixation durations and distinguish them from preview benefit effects (see Vasilev & Angele, 2017, for a comprehensive review). Their data is available online at <https://doi.org/10.17605/OSF.IO/KZ483>.

In the experiment, 34 participants, mostly students of psychology at the University of Potsdam, read 114 single sentences presented on a computer screen while their eyes were being tracked. The simple structured German sentences consisted of six to 12 words with an average length of 9 words. Every sentence contained a gaze contingent invisible boundary before a specific target word. Before the eyes crossed the boundary, the preview of the target word could either be of low, high or medium frequency (i.e. high, low or medium difficulty respectively). During the saccade in which the boundary was crossed, the target word was always exchanged with the medium frequency word. Word frequencies were taken from the dlexDB database (Heister et al., 2011) based on *The DWDS corpus: A reference corpus for the German language of the 20th century* (Geyken, 2007).

Data treatment and preprocessing. The data were collected using an Eyelink II System (*SR Research, Osgoode/Ontario, Canada*) with a temporal resolution of 1,000 Hz. Since spatial resolution was preprocessed to letter accuracy. Within-letter position was randomized by added small random numbers to avoid artifacts from discretization. Basically, the data used here were treated by the same preprocessing as reported in the statistical analysis of the experiment. Additionally, fixation durations smaller than 25 ms were discarded (550 fixations in 338 trials). Trials that included fixation durations larger than 1,000 ms were discarded (45). Trials consisting of less than three fixations were also removed from the data-set. Additionally, re-readings signaled by regressions starting from the second last or last word of the sentence and all subsequent fixations were discarded (5,773 fixations). After preprocessing, 30,639 fixations from 3,422 trials were included in the data-set for estimation. The implementations of the model, the estimation algorithm, and scripts for analyses and plots, along with the corpus data and fixation sequences are available at <https://doi.org/10.17605/OSF.IO/XDKWQ>

Appendix B

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 (see Seelig et al., 2020, for an introduction) at time t is given as $n = (n_1, n_2, \dots, n_{4+N_W})$ where n_1, \dots, n_4 are saccade timers and n_5, \dots, 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

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

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 Equation (B.1); 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) and Seelig et al. (2020).

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

$$\Lambda_i(t) = L_i^{-\eta} \sum_{j=1}^{L_i} \lambda(\epsilon_{ij}(t)), \quad (\text{B.2})$$

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.,

$$\pi(m, t) = \frac{a_m(t)}{\sum_{j=1}^{N_w} a_j(t)}, \quad (\text{B.3})$$

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 (see Seelig et al., 2020, for details).

Appendix C

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 by 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 C.1), 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 $P_{\text{spat}}(d = 0) = 0$ 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 C.1 compares the theoretical distributions of saccade amplitudes following a Gamma vs. a 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.,

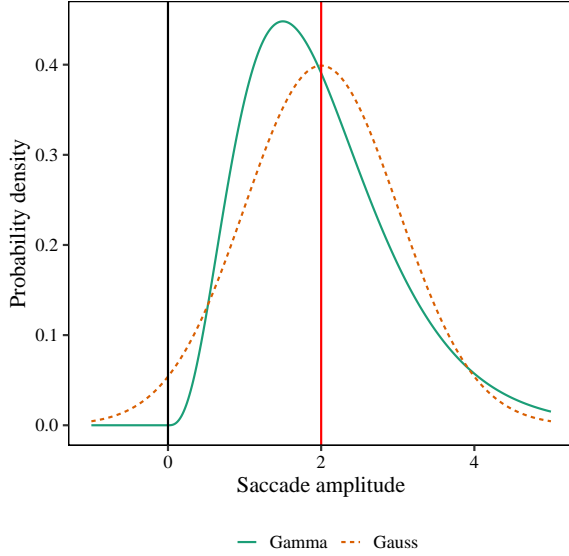
$$f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad (\text{C.1})$$

$$F(x; \alpha, \beta) = \int_0^x f(u; \alpha, \beta) du = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)} \quad (\text{C.2})$$

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

Figure C.1

Theoretical Distribution of Saccade Amplitudes Assuming a Gamma vs. Gaussian Distribution



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

an expected value (mean) and variance of

$$\mathbb{E}[D] = v_m + \epsilon_{\text{sre}} - x_{i-1} \quad (\text{C.3})$$

$$\text{Var}[D] = \sigma_{\text{sre}}^2, \quad (\text{C.4})$$

where x_{i-1} is the launch site²⁸ 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

$$\epsilon_{\text{sre}} = \text{sre}_1 - \text{sre}_2 \cdot (v_m - x_{i-1}) \quad (\text{C.5})$$

$$\sigma_{\text{sre}} = \text{omn}_1 + \text{omn}_2 \cdot |v_m - x_{i-1}|, \quad (\text{C.6})$$

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

²⁸Note 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.

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 Equations C.8 and C.9), so that the expected value is always in the direction of the respective intended target.

$$d = x_i - x_{i-1} \quad (C.7)$$

$$\mathbb{E}_F [D] = \max (v_m + \epsilon_{sre} - x_{i-1}, 0.5) \quad (C.8)$$

$$\mathbb{E}_B [D] = \max (x_{i-1} - v_m - \epsilon_{sre}, 0.5) \quad (C.9)$$

$$\text{Var} [D] = \sigma_{sre}^2 \quad (C.10)$$

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:

$$\alpha_{(.)} = \frac{(\mathbb{E}_{(.)} [D])^2}{\text{Var} [D]} \quad (C.11)$$

$$\beta_{(.)} = \frac{|\mathbb{E}_{(.)} [D]|}{\text{Var} [D]} \quad (C.12)$$

After the target k_i has been selected (see Seelig et al., 2020, for an introduction), 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:

$$D \sim \begin{cases} \Gamma_T (\alpha_F, \beta_F, x_{\max} - x_{i-1}) & \text{for } k_i > k_{i-1} \vee (k_i = k_{i-1} \wedge z > s) \\ -\Gamma_T (\alpha_B, \beta_B, x_{i-1}) & \text{otherwise} \end{cases} \quad (C.13)$$

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 Equation C.15). If not, a forward refixation is executed. R (see Equation C.14) 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 $L_{k_{i-1}}$.

$$R = \frac{x_{i-1} - b_{k_{i-1}-1}}{L_{k_{i-1}} + 1}; R \in [0, 1] \quad (\text{C.14})$$

$$s = \frac{R \cdot F(x_{i-1}; \alpha_B, \beta_B)}{(1-R) \cdot F(x_{\max} - x_{i-1}; \alpha_F, \beta_F) + R \cdot F(x_{i-1}; \alpha_B, \beta_B)} \quad (\text{C.15})$$

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 skip-pings (i.e., forward saccades), the likelihood is q_F (see Equation C.17). For regressions, the likelihood is determined with q_B (see Equation C.18). For refixations, the likelihood is the weighted sum of forward and backward saccade likelihoods, q_R (see Equation C.19).

$$\begin{aligned} q(k_i, l_i | m, F_{i-1}, \theta) &= q(x_i | m, x_{i-1}, \theta) \\ &= \begin{cases} q_F(x_i | m, x_{i-1}, \theta), & \text{for } k_i > k_{i-1} \\ q_B(x_i | m, x_{i-1}, \theta), & \text{for } k_i < k_{i-1} \\ q_R(x_i | m, x_{i-1}, \theta), & \text{for } k_i = k_{i-1} \end{cases} \end{aligned} \quad (\text{C.16})$$

$$q_F(x_i | m, x_{i-1}, \theta) = \frac{f(x_i - x_{i-1}; \alpha_F, \beta_F)}{F(x_{\max} - x_{i-1}; \alpha_F, \beta_F)} \quad (\text{C.17})$$

$$q_B(x_i | m, x_{i-1}, \theta) = \frac{f(x_{i-1} - x_i; \alpha_B, \beta_B)}{F(x_{i-1}; \alpha_B, \beta_B)} \quad (\text{C.18})$$

$$q_R(x_i | m, x_{i-1}, \theta) = \frac{R \cdot f(x_i - x_{i-1}; \alpha_F, \beta_F) + (1-R) \cdot f(x_{i-1} - x_i; \alpha_B, \beta_B)}{(1-R) \cdot F(x_{\max} - x_{i-1}; \alpha_F, \beta_F) + R \cdot F(x_{i-1}; \alpha_B, \beta_B)} \quad (\text{C.19})$$

Appendix D

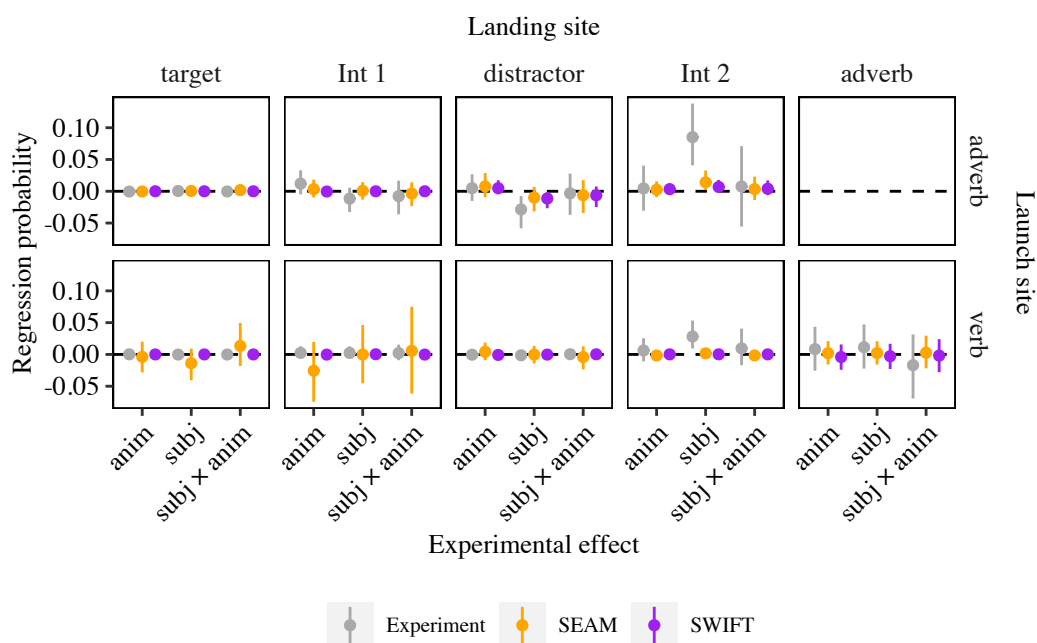
SEAM Model Parameters

| Parameter | Default | Description |
|-------------------|---------|--|
| F | 0.14 | Retrieval latency scaling factor |
| S_{\max} | 1.5 | Maximum activation strength |
| d | 0.5 | Exponential decay of memory activation |
| p | -1.5 | Additional penalty for word-category mismatch |
| μ_1 | 0.0 | Fixed production time (50 ms in original LV05 model) |
| μ_2 | 0.5 | Relative minimum activation of retrieval trigger during retrieval |
| μ_3 | 0.2 | Relative activation threshold for retrieval candidates |
| δ_0 | 7.23 | Non-dynamical (fixed) processing span width (in letter spaces) |
| δ_1 | 1.0 | Dynamical processing span width (in letter spaces) |
| asym | 1.0 | Relative width of the processing span to the left of the fixation location |
| η | 0.5 | Word-length exponent |
| α | 1.50 | Baseline word difficulty |
| β | 0.5 | Word-frequency effect on word difficulty |
| γ | 1.0 | Target selection exponent |
| minact | -5.0 | Minimum activation threshold of words for target selection |
| θ | 0.0 | Effect of predictability on processing speed |
| $t_{\text{sac}0}$ | 1.0 | Relative duration of the first fixation of the sequence |
| t_{sac} | 2.2 | Mean saccade interval (fixation duration) |
| h | 0.64 | Foveal inhibition factor |
| h_1 | 0.0 | Parafoveal inhibition factor |
| ppf | 0.0 | Inhibition from words to the left of the fixation location |
| ι | 1.0 | Transfer across saccades (activation loss during saccade) |
| M | 1.25 | Relative fixation duration of misplaced fixations |
| R | 0.8 | Relative fixation duration of well-placed refixations |
| κ_0 | 0.0 | Non-labile latency dependence on target distance (factor) |
| κ_1 | 0.0 | Non-labile latency dependence on target distance (exponent) |
| proc | 1.0 | Relative processing speed for postlexical processing |
| decay | 0.07 | Global decay of word activations during postlexical processing |
| τ_1 | 1.2 | Mean duration of the labile saccade program |

| Parameter | Default | Description |
|-----------------|---------|---|
| τ_n | 0.8 | Mean duration of the non-labile saccade program |
| τ_x | 0.2 | Mean duration of saccade execution |
| aord | 30 | Order of random walks for word activation |
| cord | 15 | Order of random walks for global saccade program |
| lord | 12 | Order of random walks for labile saccade program |
| nord | 10 | Order of random walks for non-labile saccade program |
| xord | 20 | Order of random walks for saccade execution |
| ocshift | 0.0 | Fixed oculomotor shift parameter |
| $omn_1^{(FS)}$ | 0.80 | Oculomotor noise intercept parameter for forward saccades |
| $omn_2^{(FS)}$ | 0.03 | Oculomotor noise slope parameter for forward saccades |
| $omn_1^{(SK)}$ | 0.80 | Oculomotor noise intercept parameter for skipping saccades |
| $omn_2^{(SK)}$ | 0.14 | Oculomotor noise slope parameter for skipping saccades |
| $omn_1^{(FRF)}$ | 0.80 | Oculomotor noise intercept parameter for forward refixations |
| $omn_2^{(FRF)}$ | 0.03 | Oculomotor noise slope parameter for forward refixations |
| $omn_1^{(BRF)}$ | 0.80 | Oculomotor noise intercept parameter for backward refixations |
| $omn_2^{(BRF)}$ | 0.03 | Oculomotor noise slope parameter for backward refixations |
| $omn_1^{(RG)}$ | 0.80 | Oculomotor noise intercept parameter for regressions |
| $omn_2^{(RG)}$ | 0.03 | Oculomotor noise slope parameter for regressions |
| $sre_1^{(FS)}$ | 5.0 | Saccadic range error intercept parameter for forward saccades |
| $sre_2^{(FS)}$ | 0.5 | Saccadic range error slope parameter for forward saccades |
| $sre_1^{(SK)}$ | 5.0 | Saccadic range error intercept parameter for skipping saccades |
| $sre_2^{(SK)}$ | 0.75 | Saccadic range error slope parameter for skipping saccades |
| $sre_1^{(FRF)}$ | 2.5 | Saccadic range error intercept parameter for forward refixations |
| $sre_2^{(FRF)}$ | 0.5 | Saccadic range error slope parameter for forward refixations |
| $sre_1^{(BRF)}$ | -2.5 | Saccadic range error intercept parameter for backward refixations |
| $sre_2^{(BRF)}$ | -0.5 | Saccadic range error slope parameter for backward refixations |
| $sre_1^{(RG)}$ | -2.5 | Saccadic range error intercept parameter for regressions |
| $sre_2^{(RG)}$ | -0.9 | Saccadic range error slope parameter for regressions |

Appendix E

Effects of Memory Interference on Experimental and Simulated Regression Probabilities



Note. Effects are estimates and 95% CrIs from logistic regressions. Each panel shows the effect on the proportion of trials with a critical regression given the launch site (rows) and landing site (columns). Estimates have been backtransformed to the linear scale. Regions Int 1 and Int 2 are intervening regions between target and distractor, and between distractor and adverb, respectively.

Author Publications

This bibliography lists all public scientific contributions of the author, as of the date of submission of the dissertation, grouped by type and sorted in descending chronological order. Items marked with \diamond were disseminated/presented before the onset of the author's doctoral studies. Items marked with \star are publications, on which this dissertation is based.

Unpublished Manuscripts

Engbert, R. & Rabe, M. M. (2023). *Tutorial on dynamical modeling of eye movements in reading*. PsyArXiv. <https://doi.org/10.31234/osf.io/dsvmt>

\star Rabe, M. M., Paape, D., Mertzen, D., Vasishth, S., & Engbert, R. (2023). *SEAM: An integrated activation-coupled model of sentence processing and eye movements in reading*. arXiv. <https://doi.org/10.48550/arXiv.2303.05221>

Rabe, M. M., Lindsay, D. S., & Kliegl, R. (2021). *ROC asymmetry is not diagnostic of unequal residual variance in Gaussian signal detection theory*. PsyArXiv. <https://doi.org/10.31234/osf.io/erzvp>

Rabe, M. M., Paape, D., Vasishth, S., & Engbert, R. (2021). *Dynamical cognitive modeling of syntactic processing and eye movement control in reading*. PsyArXiv. <https://doi.org/10.31234/osf.io/w89zt>

Published Journal Articles

Engbert, R., Rabe, M. M., Schwetlick, L., Seelig, S. A., Reich, S., & Vasishth, S. (2021). Data assimilation in dynamical cognitive science. *Topics in Cognitive Science*, 26(2), 99–102. <https://doi.org/10.1016/j.tics.2021.11.006>

\star Rabe, M. M., Chandra, J., Krügel, A., Seelig, S. A., Vasishth, S., & Engbert, R. (2021). A Bayesian approach to dynamical modeling of eye-movement control in reading of normal, mirrored, and scrambled texts. *Psychological Review*, 128(5), 803–823. <https://doi.org/10.1037/rev0000268>

- Engbert, R., Rabe, M. M., Kliegl, R., & Reich, S. (2021). Sequential data assimilation of the SEIR model for COVID-19. *Bulletin of Mathematical Biology*, 83, Article 1 (2021). <https://doi.org/10.1007/s11538-020-00834-8>
- 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(48), Article 2134. <https://doi.org/10.21105/joss.02134>
- *Seelig, S. A., Rabe, M. M., Malem-Shinitzki, 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, Article 102313. <https://doi.org/10.1016/j.jmp.2019.102313>
- ◊Masson, M. E. J., Rabe, M. M., & Kliegl, R. (2017). Modulation of additive and interactive effects by trial history revisited. *Memory & Cognition*, 45(3), 480–492. <https://doi.org/10.3758/s13421-016-0666-z>

Software

- Granziol, U., Rabe, M. M., Spoto, A., & Vidotto, G. (2023). *appRiori*: Code and obtain customized planned comparisons with ‘appRiori’ [R package]. <https://cran.r-project.org/package=appRiori>
- Rabe, M. M., Kliegl, R., & Schad, D. J. (2019). *designr*: Balanced factorial designs with crossed random and fixed factors [R package]. <https://cran.r-project.org/package=designr>
- Rabe, M. M., Schad, D. J., Vasishth, S., & Kliegl, R. (2019). *hypr*: Hypothesis matrix translation [R package]. <https://cran.r-project.org/package=hypr>

Conference Contributions and Invited Talks

- Rabe, M. M., Paape, D., Vasishth, S., & Engbert, R. (2021, November). *SEAM: An integrated activation-coupled model of sentence processing and eye movements in reading* [Poster]. Psychonomic Society 62nd Annual Meeting (virtual).
- Rabe, M. M. (2019, November). *Eye-movement control during reading: Bayesian parameter inference of the dynamical SWIFT model* [Invited talk]. University of Victoria, BC, Canada.
- Rabe, M. M., Chandra, J., Krügel, A., Seelig, S. A., & Engbert, R. (2019, November). *Bayesian inference of dynamical cognitive and oculomotor processes in the SWIFT model of reading* [Poster]. Psychonomic Society 60th Annual Meeting, Montréal, QC, Canada. <https://doi.org/10.17605/osf.io/mxq4n>

- Seelig, S. A., Rabe, M. M., Malem-Shinitski, N., Reich, S., & Engbert, R. (2019, September). *Bayesian parameter estimation for the SWIFT model of eye-movement control during reading* [Poster]. Cognitive Computational Neuroscience (CCN) 2019, Berlin, Germany. <https://doi.org/10.32470/CCN.2019.1369-0>
- Rabe, M. M., Chandra, J., Krügel, A., Seelig, S. A., & Engbert, R. (2019, August). *Bayesian inference of the SWIFT model: Reading mirrored, scrambled, and normal texts* [Poster]. European Conference on Eye Movements (ECEM) 2019, Alicante, Spain. <https://doi.org/10.17605/osf.io/s5jk2>
- Rabe, M. M., Seelig, S. A., Chandra, J., Vasishth, S., Reich, S., & Engbert, R. (2019, March). *Parameter inference and model comparison in dynamical cognitive models: SWIFT* [Poster]. 2nd SFB 1294 Data Assimilation Spring School, Dierhagen, Germany.
- ◊Rabe, M. M., Kliegl, R., & Lindsay, D. S. (2017, November). *A generalized linear mixed model approach to signal detection theory in recognition memory experiments* [Poster]. Psychonomic Society 58th Annual Meeting, Vancouver, BC, Canada.
- ◊Rabe, M. M., Kliegl, R., & Lindsay, D. S. (2017, May). *Generalized linear mixed model of signal detection theory* [Poster]. Northwest Cognition and Memory (NOWCAM), Burnaby, BC, Canada.
- ◊Fallow, K. M., Rabe, M. M., & Lindsay, D. S. (2015, November). *Recognition memory response bias for paintings, words, and faces* [Poster]. Psychonomic Society 56th Annual Meeting, Chicago, IL, USA.

Declaration of Authorship

I hereby certify that the dissertation I am submitting is entirely my own original work except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those regulations concerning disciplinary actions that may result from plagiarism. Any use of the works by any other author, in any form, is properly acknowledged at their point of use.

Potsdam, September 11, 2023

Place and Date

(see below)

Signature

Eigenständigkeitserklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Ich bestätigte die Kenntnisnahme der Richtlinien der Universität zum Umgang mit Plagiarismus, einschließlich solcher zu daraus folgenden Disziplinarmaßnahmen. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht.

Potsdam, 11. September 2023

Ort und Datum

Unterschrift