GAZING INTO A DISCRETE WORLD

MIXTURE MODELS OF COGNITION & BEHAVIOR

ŠIMON KUCHARSKÝ

Gazing into a Discrete World

Mixture models of Cognition & Behavior

Šimon Kucharský

This dissertation was typeset by the author using LaTeX 2ε, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 12 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface.

The PhD dissertation with this look & feel has was inspired by a template created by Hoàng-Ân Lê that can be found online at github.com/lhoangan/template-uva-thesis. The template was originally created and released under the permissive AGPL license by Jordan Suchow and is available at github.com/suchow/Dissertate.

The concept for the art on the front cover was created by Šimon Kucharský with assistance from DALL-E, and was ultimately turned into print-ready artwork by Iva van der Maas.

ISBN 978-94-93330-78-8

Gazing into a Discrete World

Mixture models of Cognition & Behavior

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 29 mei 2024, te 16.00 uur

door Šimon Kucharský
geboren te Praha

Promotiecommissie

Promotores:	dr.	I. Visser	Universiteit van Amsterdam
	prof. dr.	M.E.J. Raijmakers	Universiteit van Amsterdam
Copromotores:	prof. dr.	E.M. Wagenmakers	Universiteit van Amsterdam
Overige leden:	prof. dr.	A.J. Heathcote	University of Newcastle, Australia
	prof. dr.	H.M. Huizenga	Universiteit van Amsterdam
	dr.	D. Matzke	Universiteit van Amsterdam
	dr.	I.T.C. Hooge	Utrecht Universiteit
	prof. dr.	M.C. Frank	Stanford University
	dr.	J.M. Haaf	Universiteit van Amsterdam

Faculteit der Maatschappij- en Gedragswetenschappen

The work described in this thesis has been carried out within the Baby Lab Amsterdam and the JASP team at the University of Amsterdam. The research was supported by NWO talent grant no. 406.10.559.

University of Amsterdam

To my beloved family
Mé milé rodině
Tang cho gia đình yêu

Gazing into a Discrete World
Mixture models of Cognition & Behavior

Summary

This thesis titled Gazing into a Discrete World presents perspectives on modeling human behavior, focusing on alternative sources of data such as eye-tracking and response times, with a special focus dedicated to substantive questions about qualitative patterns in individual differences and development. Further, it advocates for a closer alignment between design and analysis of experiments, and their theoretical underpinnings.

The thesis is structured into three distinct parts. The first part delves into identifying and analyzing discrete behavioral patterns, particularly through eye movement data, emphasising model-based approaches for understanding visual attention. The second part addresses challenges in empirical research, with a focus on developmental psychology, offering remedies for imperfections and methodological advancements. The third part focuses on making correct inferences under uncertainty, highlighting the significance of Bayesian methods and developing openly available software tools for applied researchers.

The thesis contributes with advancements in integration of eye-tracking into cognitive-behavioral modeling, improvements in developmental psychology research, and provides openly available Bayesian tools.

Gazing into a Discrete World
Mixture models of Cognition & Behavior

Samenvatting

Dit proefschrift getiteld Gazing into a Discrete World presenteert perspectieven op het modelleren van menselijk gedrag, waarbij de nadruk ligt op alternatieve gegevensbronnen zoals eye-tracking en reactietijden, met een speciale focus op inhoudelijke vragen over kwalitatieve patronen in individuele verschillen en ontwikkeling. Verder wordt er gepleit voor een betere afstemming tussen het ontwerp en de analyse van experimenten en hun theoretische onderbouwing.

Het proefschrift bestaat uit drie afzonderlijke delen. Het eerste deel gaat in op het identificeren en analyseren van discrete gedragspatronen, in het bijzonder door middel van oogbewegingsdata, waarbij de nadruk ligt op modelgebaseerde benaderingen voor het begrijpen van visuele aandacht. Het tweede deel gaat in op uitdagingen in empirisch onderzoek, met een focus op ontwikkelingspsychologie, en biedt oplossingen voor onvolkomenheden en methodologische verbeteringen. Het derde deel richt zich op het maken van correcte gevolgtrekkingen onder onzekerheid, waarbij het belang van Bayesiaanse methoden wordt benadrukt en open software voor onderzoekers worden ontwikkeld.

Het proefschrift levert een bijdrage aan de integratie van eye-tracking in cognitief-gedragsmatige modellering, verbeteringen in ontwikkelingspsychologisch onderzoek en biedt openlijk beschikbare Bayesiaanse hulpmiddelen.

Introduction 3

Gaze HMM 13

WALD-EM 67

Cognitive strategies 117

HMM EAM 151

Sample size planning 209

Design choices 257

Rethinking habituation 287

Partial correlation 333

Binary classification 359

Statistics with JASP 393

Conclusion 421

Contents

Summary vii

Samenvatting viii

Thumb marks ix

Introduction 3

I Discrete Patterns of Behavior 11

1 Gaze HMM 13

1.1 Introduction . . . 14
1.2 Developing gazeHMM . . . 20
1.3 Simulation Study . . . 28
1.4 Validation Study . . . 49
1.5 Conclusion & Discussion . . . 60

2 WALD-EM 67

2.1 Introduction . . . 68
2.2 Conceptual WALD-EM model . . . 73
2.3 Concrete WALD-EM model . . . 79
2.4 Application: Infant scene viewing . . . 87
2.5 Benefits of joint modeling . . . 105
2.6 Conclusion & Discussion . . . 110

3 Cognitive strategies **117**
3.1 Introduction 118
3.2 Clustering transition matrices 124
3.3 Application: Deductive Mastermind 125
3.4 Application: Progressive Matrices 138
3.5 Conclusion & Discussion 144

4 HMM EAM **151**
4.1 Introduction 152
4.2 Model 156
4.3 Simulation study 162
4.4 Application: Dutilh et al. (2011) study 179
4.5 Conclusion & Discussion 187

Appendix **195**
4.A Derivation of the simplified LBA model 195
4.B Parameter estimates 199

II Addressing Imperfections **207**

5 Sample size planning **209**
5.1 Introduction 210
5.2 Planning for BST and BFDA 219
5.3 Illustration: Rule learning 226
5.4 Conclusion & Discussion 247

Appendix **251**
5.A Determining informed priors 251

6 Design choices **257**
6.1 Introduction 258
6.2 The Current Study 268
6.3 Method 271
6.4 Results 283
6.5 Discussion 283

Appendix **285**

7 Rethinking habituation **287**
7.1 Introduction 288
7.2 Habituation as a tool 289
7.3 A way forward 301
7.4 Practical considerations 318
7.5 Conclusion, Discussion, & Recommendations 324

Appendix **329**
7.A Extended habituation model of Thomas and Gilmore (2004) . 329

III Learning under Uncertainty **331**

8 Partial correlation **333**
8.1 Introduction 334
8.2 Bayesian inference for partial correlation 336
8.3 Properties of the Bayes factor 339
8.4 Two examples 341
8.5 Concluding remarks 345

Appendix **349**
8.A Derivation of the main results 349
8.B Rewriting the trace 356

9 Binary classification **359**
9.1 Introduction 360
9.2 Example 1: Parameters Known Exactly 363
9.3 Example 2: Parameters Subject to Uncertainty 380
9.4 Concluding Comments 388

10 Statistics with JASP **393**
10.1 Introduction 394
10.2 Characteristic Features of JASP 395
10.3 Example 1: Descriptive Statistics 400

10.4 Example 2: Bayesian T-Test 401
10.5 Example 3: The Learn Stats Module 404
10.6 Example 4: The Learn Bayes Module 405
10.7 Example 5: Distributions 409
10.8 Example 6: A Network Analysis 410
10.9 Example 7: Time Series & Forecasting 411
10.10 Integration with R 414
10.11 Why not just use R? 419

Conclusion **421**

List of Contributions **433**

A Word of Thanks **441**

References **442**

We live in a society where everybody knows everything
and it is a shame to say "I don't know"

– Arsène Wenger

Every day is important, every act is important. The secret is that people have to believe what you are trying to deliver.

–Mikel Arteta

Introduction

Discrete switches in behavior are ubiquitous phenomena observed across various domains of life. These transitions, which can result in markedly different functions and outcomes, are integral to the dynamics and progression of various systems. They have been thoroughly examined in disciplines like biology, physics, social sciences, and engineering.

Within the realm of psychology, such discrete shifts are evident across a broad spectrum of human activities, from the most fundamental cognitive processes like changing attention focus, through the complexities of reasoning and decision-making, up to learning and development. These shifts can as well occur within a single individual or as patterns between different individuals.

In cognitive experiments, human behavior is traditionally captured through responses. While this data might suffice to identify distinct cognitive modes in some instances, it may fall short in others. For example, two individuals using different strategies to solve a task might reach the same conclusion, even if the underlying cognitive processes may be markedly different. Consequently, it is often desirable to seek additional forms of data, such as eye movements or response times, to discern new behavioral patterns.

Eye-tracking data, which monitor the trajectory of a participant's gaze over time as a proxy for visual attention, is one such valuable data source. Thus, it holds significant promise for deepening our comprehension of human behavior.

The optimal approach to studying human behavior involves the application of robust theories, the construction of generative models, and a close integration of statistical analysis with theoretical frameworks. The ambition of this thesis is to formulate models that characterize human behavior at various analytical levels, with a particular focus on analyzing eye movement data to detect

discrete behavioral shifts.

Yet, the foundation of a strong theoretical framework, a concept for a generative model, or an immediately available and practical statistical model for estimation within reasonable constraints is not always at hand. At times, the wide-ranging nature of empirical evidence can make it even more challenging to determine an appropriate starting point for modeling. Therefore, this thesis also includes initiatives that supply practical tools for researchers to gather and analyze their data. Additionally, it offers descriptions of current research practices in specific paradigms, all aimed at guiding the field toward a more structured modeling of human behavior.

Structure

This thesis consists of three parts, each showcasing the complex dimensions of empirical research in this field. The first part addresses substantive questions about discrete switches observed in human behavior, placing a spotlight on discerning these patterns through eye-tracking data as an indicative measure of visual attention, and devising statistical models that align with the prevailing theoretical insights into these phenomena. The second part pivots to tackle methodological hurdles and the practical barriers in conducting high-quality research, with a particular emphasis on our experience with popular paradigms in developmental psychology research. The third part concentrates on the science of making accurate inferences amidst uncertainty, and it is particularly dedicated to equipping applied researchers with statistically principled methods for their research endeavors.

Discrete Patterns of Behavior

In the first part we concern ourselves with uncovering and analysing behavior that may or may not exhibit discrete patterns, and explore methods how eye movement data can be used in addition to response behavior.

Human visual attention is characterised and manifested by discrete events such as fixations and saccades. Before we turn our attention towards higher cognitive tasks, we delve into the world of identifying these events from raw-eye tracking data. Chapter 1 presents a new, model-based approach for identi-

fying discrete events of eye movements, such as fixations and saccades, from a continuous stream of data from the eye-tracker.

Once eye movement events are identified, we can study them to better understand visual behavior. Chapter 2 uses fixation data to model systematic patterns in eye movement behavior. We argue that both temporal and spatial aspects of visual attention need to be taken into account when studying these phenomena, and show the benefits of generative and analytic models that are built on strong theoretical foundations.

Finally, chapters 3 and 4 are concerned with higher level cognitive reasoning and discrete behavior that might emerge as a result of using different strategies as a means to solve cognitive tasks. In chapter 3, we use eye-tracking methods to detect cognitive strategies in solving higher level cognitive tasks (Mastermind game and a Matrix reasoning task). However, not always is eye movement data necessary to detect distinct modes of behavior. Chapter 4 presents a dynamic model of evidence accumulation in speeded decision tasks that can detect participants switching between guessing and a stimulus controlled mode, using only responses and response times to identify them.

In conclusion, the first part of this thesis represents a comprehensive exploration into the realm of discrete behavioral patterns, with a special emphasis on eye movement data as a key to understanding complex visual attention mechanisms. From proposing a model-based approach for identifying specific eye movement events to applying fixation data for modeling systematic patterns in visual behavior, this work underscores the significance of integrating both temporal and spatial aspects of visual attention. Additionally, it emphasizes the importance of robust theoretical foundations for generative and analytic models. The chapters dedicated to higher cognitive reasoning and distinct behavioral modes further expand our understanding, demonstrating the varied applications of eye-tracking methodologies in detecting cognitive strategies and the utility of dynamic models in identifying nuanced aspects of decision-making processes. This part of the thesis not only contributes to our understanding of visual behavior and provides distinction between qualitative and quantitative inter- and intra-individual differences, but also highlights the potential of eye-tracking data in revealing deeper insights into human cognition.

Addressing Imperfections

Conducting empirical research presents a host of challenges that researchers must navigate to ensure the validity and reliability of their findings. One primary hurdle is sample size planning, where determining the appropriate number of participants is critical to achieving statistical power while maintaining practical and ethical constraints. Too small a sample can lead to underpowered studies that cannot detect true effects, while excessively large samples may waste resources and potentially expose subjects to more intensive experimental designs than necessary. Another significant challenge is the need to synthesize empirical findings across studies, which requires systematic methods and rigorous meta-analytical techniques to integrate results, often from diverse contexts and varying methodological quality. This synthesis is essential to build a coherent understanding of phenomena and to guide future research directions. Lastly, optimizing research design paradigms is crucial for obtaining clear and interpretable results. This involves careful consideration of the research questions, and the use of designs that minimize biases and confounds. Each of these steps requires meticulous planning and a deep understanding of methodological principles to ensure that empirical research contributes meaningful knowledge to the field.

The second part of the thesis presents a critical and still often understated reality in the realm of empirical research: the inherent imperfections and challenges that accompany the execution of any study. This acknowledgment stems from our wealth of experience with specific experimental paradigms, their limitations, and the need for continual methodological refinement. It underscores the thesis's commitment to not just conduct empirical research but also to critically evaluate and enhance the processes that underpin it.

In this pursuit, three chapters are presented with dual objectives. The first is a critical examination of widely-used experimental paradigms, mainly the habituation paradigm, pinpointing their deficiencies and proposing innovative ways to augment them. This introspective approach is grounded in the belief that understanding and addressing the flaws of current methodologies is as important as the empirical investigations themselves.

The second objective is to offer practical methods and guidance for researchers striving to elevate their empirical work. This includes strategies for more effec-

tive research design, data collection, and analysis, all aimed at bolstering the robustness and reliability of research outcomes.

Collectively, these studies serve a dual purpose: they reflect on the existing practices, offering a candid critique, guiding researchers towards methodological excellence. The overarching goal is to foster a research environment that not only produces results but also refines the tools and techniques that allow science to progress in a reliable and replicable manner.

Requiring sufficient sample sizes is a considerable problem in science in general. This problem is particularly tricky to solve in developmental research where resources (e.g., infants and their attention) are scarce. Chapter 5 aims to tackle this problem by discussing Bayesian sequential sampling designs which allow researchers collecting data until they accrue sufficient evidence, or run out of time, money, or patience, and as a result help researchers avoid wasting precious resources.

Chapters 6 and 7 are focused on a popular experimental paradigm in infant research that uses habituation as a pivotal tool to study various phenomena. This method relies heavily on the precise measurement and identification of an infant's habituation response to stimuli. While several techniques to assess habituation have been proposed, the field lacks a consensus on the most effective approaches for specific research objectives. Chapter 6 offers a comprehensive, collaborative effort through a systematic review and meta-analysis, aiming to describe current practices in infant habituation research and to quantify the typical effect sizes observed with these methods. This large-scale collaborative project is currently still ongoing, after its preregistration was accepted as part of a Registered Report. The chapter presents this preregistration, and so does not contain results. Chapter 7, on the other hand, argues that in addition to learning from the current practices in habituation research, one needs to venture beyond existing methodologies, introducing novel approaches to study habituation in itself. Further understanding of the habituation process will also allow proposing alternative research designs that employ habituation as an investigative tool for other areas of interest in infant research.

Overall, this part of the thesis emphasizes the importance of continuous reflection and improvement in research methodologies, aiming to enhance the quality and reliability of empirical studies, particularly in developmental psy-

chology.

Learning under Uncertainty

While our understanding of empirical phenomena always evolves, one constant remains: the presence of uncertainty. It permeates every aspect of research, from hypothesis formation to data interpretation. Recognizing this uncertainty is a necessary, fundamental aspect of scientific progress. Thus, a pivotal question of scientific advancement is learning in the presence of uncertainty.

Bayesian reasoning emerges as a powerful ally in this context. It provides a formal framework for updating our beliefs in light of new evidence, a process at the core of scientific inquiry. By integrating prior knowledge with current data, Bayesian methods offer a coherent and adaptable approach. This paradigm is not just about reaching conclusions in a principled manner; it is about quantifying the degree of confidence in these conclusions and adjusting them as new evidence becomes available.

Unfortunately, computational challenges often stand as formidable barriers in adopting Bayesian methods. Furthermore, the uncertainty in constructing appropriate priors and the potential for misinterpretation of Bayesian results can add even more hurdles. These challenges can be especially daunting for practitioners who may not have extensive training in advanced statistical methods or access to powerful computing resources. Therefore, the development of Bayesian methods that are accessible and user-friendly for the practical researcher is of great importance. Simplifying Bayesian analysis without compromising its rigor can democratize its application, enabling a wider range of scientists to harness its full potential. This involves creating more intuitive software tools, and providing better guidance on interpreting results using Bayesian tools. By making Bayesian methods more approachable and feasible in a practical context, we can broaden their impact across various fields of scientific inquiry.

Chapter 8 is dedicated to advancing the availability of well-calibrated default Bayesian procedures for practical researchers, thereby enriching the statistical toolkit with analyses that are both easily accessible, applicable, and interpretable. This chapter builds upon prior research in Bayesian Pearson's cor-

relation (Ly, Marsman, & Wagenmakers, 2018), introducing an extension that enables researchers to perform Bayesian partial correlation analyses. It provides analytical solutions for the Bayes factor test concerning the null hypothesis of a partial correlation, along with the posterior distribution of the partial correlation coefficient, under the assumption that the alternative hypothesis is true. Deriving analytical solution has a couple of advantages. First, analytic methods are typically faster than numerical methods, and in general require less computational resources. In the current case, it also allows us to conduct full Bayesian inference with just a couple of summary statistics of the data (instead of using the full raw data), which makes it easy to apply in cases where the original data is not available. The analytic solution for a partial correlation also shows an interesting connection to the Pearson's correlation, showing that the inference for partial correlation is just its generalization. Insights from this connection allowed us to study the properties of the Bayes factor and specify under what conditions it satisfies desiderata for Bayes factors proposed by Jeffreys (1961). This development marks a significant step in equipping researchers with more robust and sophisticated Bayesian tools for their analytical needs.

In addition to developing novel Bayesian analyses, enhancing the understanding of Bayesian reasoning is also important. Chapter 9 presents a tutorial, primarily aimed at medical students and professionals, for interpreting binary classification procedures – such as medical tests – where the posterior probability of having a disease can be computed using *prevalence*, *sensitivity*, and *specificity*. Such examples are often used to introduce Bayes' theorem to students, illustrating key concepts such as prior probability, posterior probability, conditional and total probability. However, in the real world, *prevalence*, *sensitivity*, and *specificity* are often unknown, and therefore are associated with uncertainty. This chapter guides the reader through an example using the `Binary classification` module in JASP, focusing on how such uncertainty is taken into account from a principled Bayesian perspective.

Chapter 9 taps into an area of methodological research where potentially complex analyses and procedures are made available to the practitioners through easy to use, intuitive software. Chapter 10 expands on this idea and provides an overview of JASP, a user-friendly, open source software that makes a wide range of statistical procedures available through a couple of simple mouse clicks. This

chapter represents a body of work dedicated to development and maintenance of this software.

All chapters together provide insights into the discrete patterns of behavior, addressing methodological imperfections and embracing the inherent uncertainty in cognitive and developmental research, ultimately offering a structured, theory-driven perspective on human cognition and behavior, and advocates for responsible, transparent, and reliable implementation of research methods.

Part I

Discrete Patterns of Behavior

If you closed your eyes, you couldn't tell the difference.

–Phil Brown

Chapter 1

Characterising Eye Movement Events with an Unsupervised Hidden Markov Model

This chapter is published as Lüken, M., Kucharský, Š., and Visser, I. (2022). Characterising eye movement events with an unsupervised hidden Markov model. *Journal of Eye Movement Research*, *15*(1). doi: 10.16910/jemr.15.1.4

Abstract

Eye-tracking allows researchers to infer cognitive processes from eye movements that are classified into distinct events. Parsing the events is typically done by algorithms. Previous algorithms have successfully used hidden Markov models (HMMs) for classification but can still be improved in several aspects. To address these aspects, we developed `gazeHMM`, an algorithm that uses an HMM as a generative model, has no critical parameters to be set by users, and does not require human coded data as input. The algorithm classifies gaze data into fixations, saccades, and optionally postsaccadic oscillations and smooth pursuits.

We evaluated gazeHMM's performance in a simulation study, showing that it successfully recovered HMM parameters and hidden states. Parameters were less well recovered when we included a smooth pursuit state and/or added even small noise to simulated data. We applied generative models with different numbers of events to benchmark data. Comparing them indicated that HMMs with more events than expected had most likely generated the data. We also applied the full algorithm to benchmark data and assessed its similarity to human coding. For static stimuli, gazeHMM showed high similarity and outperformed other algorithms in this regard. For dynamic stimuli, gazeHMM tended to rapidly switch between fixations and smooth pursuits but still displayed higher similarity than other algorithms. Concluding that gazeHMM can be used in practice, we recommend parsing smooth pursuits only for exploratory purposes.

Future HMM algorithms could use covariates to better capture eye movement processes and explicitly model event durations to classify smooth pursuits more accurately.

1.1 Introduction

Eye-tracking is often used to study cognitive processes involving attention and information search based on recorded gaze position (Schulte-Mecklenbeck et al., 2017). Before these processes can be studied, the raw gaze data is classified into events that are distinct in their physiological patterns (e.g., duration), underlying neurological mechanisms, or cognitive functions (Leigh & Zee, 2015). Basic events are fixations, saccades, smooth pursuits, and post-saccadic oscillations (PSOs). Classifying raw eye-tracking data into these events reduces their complexity and is usually the first step towards cognitive interpretation (Salvucci & Goldberg, 2000). The classification is typically done by algorithms, which is considered faster, more objective, and

reproducible compared to human coding (Andersson, Larsson, Holmqvist, Stridh, & Nyström, 2017). Hein and Zangemeister (2017) give a comprehensive overview of different classification algorithms (for a structured review on classifying saccades, see also Stuart et al., 2019).

The aim of the current study is to develop a generative, unsupervised model for characterising, describing and understanding eye movement data. Below we discuss the requirements for such a model. One such requirement is obviously that it can reliably classify eye movement events.

To motivate our decision to add another algorithm to this array of classification tools, it is useful to briefly discuss the properties and goals of those tools. On one hand, many classification algorithms use non-parametric methods to differentiate between eye movement events[1]. A classic example is the "Velocity-threshold" algorithm (Stampe, 1993), which classifies[2] samples with a velocity above a fixed threshold as saccades (see also Larsson, Nystrom, & Stridh, 2013; Larsson, Nyström, Andersson, & Stridh, 2015; Nyström & Holmqvist, 2010). On the other hand, many parametric methods have been developed recently. Some of them require human-labeled training data as input and can therefore be termed as *supervised* (Hastie, Tibshirani, & Friedman, 2017). For example, Bellet, Bellet, Nienborg, Hafed, and Berens (2019) trained a convolutional neural network (CNN) on eye-tracking data from humans and macaques and achieved saccade classifications that were highly similar to those of human coders (for other supervised algorithms, see Startsev, Agtzidis, & Dorr, 2019; Zemblys, Niehorster, & Holmqvist, 2019; Zemblys, Niehorster, Komogortsev, & Holmqvist, 2018). Due to their high agreement with human coders, one might call the supervised approaches "state-of-the-art". However, the requirement of labeled training data is a disadvantage of supervised methods because the labeling process can easily become costly and time-consuming (Zemblys et al., 2019). More importantly, supervised methods also (implicitly) treat human-labeled

[1] We use the terms parametric/non-parametric to distinguish between models that assume population distribution properties with finite number of population (unknown) parameters vs. models that do not assume population distributions, assume distribution with infinite number of parameters, or leave the population parameters undefined (Geisser & Johnson, 2006; Zacks, 2014).

[2] We use the terms classification and event classification throughout this paper but see discussion about the appropriateness of those terms as compared with event detection in (R. S. Hessels, Niehorster, Nyström, Andersson, & Hooge, 2018).

training data as a reliable gold standard, an assumption that may be unwarranted (see discussion in Hooge, Niehorster, Nyström, Andersson, & Hessels, 2018). The reliance on training data also makes supervised methods inflexible: When test data strongly deviates from the training data, the classification performance can decrease substantially (e.g., Startsev et al., 2019). Furthermore, when the required events for test data differ from the hand-coded events in the training data, the latter would need to be recoded, causing additional costs. In contrast, *unsupervised* classification algorithms do not require labeled training input. Instead, they learn parameters from the characteristics of the data themselves (Hastie et al., 2017). In consequence, they are also more flexible in classifying data from different individuals, tasks, or eye-trackers (e.g., R. S. Hessels, Niehorster, Kemner, & Hooge, 2017; Houpt, Frame, & Blaha, 2018).

Besides discriminating between supervised and unsupervised methods, algorithms can vary in whether they are explicitly modeling the data generating process and are thus able to simulate new data. To our knowledge, these *generative models* have been rarely used to classify eye movement data (cf. Mihali, van Opheusden, & Ma, 2017; Wadehn, Weber, Mack, Heldt, & Loeliger, 2020). Classifiers with generative assumptions have the advantage that their parameters can be easily interpreted in terms of the underlying theory. In the context of eye movements, they can also help to explain or confirm observed phenomena: For instance, their parameters can indicate that oscillations only occur after but not before saccades. When the goal is to understand eye movement events and improve their classification based on this understanding, this aspect is an advantage over non-parametric or supervised methods. Moreover, generative models can challenge common theoretical assumptions and bring up new research questions (Epstein, 2008). For example, they might suggest that oscillations also occur before saccadic eye movements (as mentioned in Nyström & Holmqvist, 2010) or that the assumption that eye movements are discrete events (e.g., saccades and PSOs cannot overlap) does not hold (as discussed in Andersson et al., 2017).

We argue that the recent focus on supervised approaches misses an important facet of eye movement event classification: Supervised methods are trained on human-labeled data and can predict human classification well. This is an important milestone for applicants that are interested in automating human

classification. However, since human classification may not be as reliable, valid, and objective as assumed (Andersson et al., 2017; Hooge et al., 2018), supervised approaches will also reproduce these flaws. Instead, we suggest taking a different avenue and developed an unsupervised, generative algorithm to set a starting point for more explicit parametric modeling of common eye movement events (cf. Mihali et al., 2017). By relying on likelihood-based goodness-of-fit measures, we aim to achieve a classification that reaches validity through model comparison instead of making the classification more human-like. A model-based approach can also improve the reliability because it will lead to the same classification given the correct settings, whereas human annotation can depend on implicit, idiosyncratic thresholds that may be hard to reproduce (see Hooge et al., 2018).

One class of generative models that are used in eye movement classification are HMMs. They estimate a sequence of hidden states (i.e., a discrete variable that cannot be directly observed) that evolves parallel to the gaze signal. Each gaze sample depends on its corresponding state. Each state depends on the previous but not on earlier states of the sequence (Zucchini, MacDonald, & Langrock, 2016). Further, HMMs can be viewed as unsupervised models that can learn the hidden states and parameters of the emission process from the observed data alone, and as such do not in principle need labeled training data. They are suitable models for eye movement classification because the hidden states can be interpreted as eye movement events and gaze data are dependent time series (i.e., one gaze sample depends on the previous). HMMs can be applied to individual or aggregated data (or both, see Houpt et al., 2018) and are thus able to adapt well to interindividual differences in eye movements.

On this basis, several classification algorithms using HMMs have been developed: One instance is described in Salvucci and Goldberg (2000) and combines the HMM with a fixed threshold approach (named "Identification by HMM" [I-HMM]). Samples are first labeled as fixations or saccades, depending on whether their velocity exceeds a threshold, and then reclassified by the HMM. Pekkanen and Lappi (2017) developed an algorithm that filters the position of gaze samples through naive segmented linear regression (NSLR). The algorithm uses an HMM to parse the resulting segments into fixations, saccades, smooth pursuits, and PSOs based on their velocity and change in angle (named

NSLR-HMM). Another version by Mihali et al. (2017) uses a Bayesian HMM to separate microsaccades (short saccades during fixations) from motor noise based on sample velocity (named "Bayesian Microsaccade Detection" [BMD]). Moreover, Houpt et al. (2018) applied a hierarchical approach developed by Fox and colleagues that describes sample velocity and acceleration through an autoregression (AR) model, computes the regression weights through an HMM, and estimates the number of events with a beta-process (BP) from the data (named BP-AR-HMM).

Several studies have tested the performance of HMM algorithms against other classification methods: I-HMM has been deemed as robust against noise, behaviorally accurate, and showing a high sample-to-sample agreement to human coders (Andersson et al., 2017; Komogortsev, Gobert, Jayarathna, Koh, & Gowda, 2010; Salvucci & Goldberg, 2000). However, the agreement was lower when compared to an algorithm using a Bayesian mixture model (Kasneci, Kasneci, Kübler, & Rosenstiel, 2014; Tafaj, Kasneci, Rosenstiel, & Bogdan, 2012). NSLR-HMM showed even higher agreement to human coding than I-HMM (Pekkanen & Lappi, 2017) but was outperformed for saccades by the CNN algorithm by Bellet et al. (2019).

In sum, HMMs seem to be a promising method for classifying eye movements in unsupervised settings. Nevertheless, the existing HMM algorithms each have at least one aspect in which they could be improved.

First, I-HMM relies on setting an appropriate threshold to determine the initial classification, which can distort the results (Blignaut, 2009; Komogortsev et al., 2010; Shic, Scassellati, & Chawarska, 2008). Second, the current implementation of NSLR-HMM requires human-coded data, which narrows its applicability to applications where supervised methods are also an option. It also inheres fixed parameters that prevent the algorithm to adapt to individual- or task-specific signals. Third, BMD limits the classification to microsaccades which are irrelevant in many applications and sometimes even considered as noise (Duchowski, 2017). The opposite problem was observed for BP-AR-HMM: It tends to estimate an unreasonable number of events from the data of which many are considered as noise events (e.g., blinks). Therefore, the authors suggest using it as an exploratory tool followed by further event classification (Houpt et al., 2018).

1.1.1 Goals

The goal of the project reported in this article is to move towards generative models of eye movement events. The purpose of generative models is to bring better understanding of the events they describe in a fully statistical framework, which enables likelihood-based comparisons and hypothesis tests, or to generate novel hypotheses. Such models can be also used for classification, even though that may not be their only or primary application.

In this article, we present a novel model of eye movement events, named gazeHMM, that relies on an HMM as a generative model.

The first step in developing a generative model that can be also used as a statistical model (e.g., to be fit to data), is to ensure its computational consistency, that is, whether the model is able to recover parameter values that were used to generate the data. Second, as classification is one of the possible applications of such model, it is important to evaluate the classification performance and ensure that the model does reasonably well identifying the eye movement events it putatively describes. We believe these two questions are the minimal requirements of a generative model in the current setting, and the current article brings just that — evaluation of the basic characteristics of a generative model that we developed.

Table 1.1 presents a selection of recently developed classification algorithms (i.e., the "state-of-the-art") and highlights the contribution of gazeHMM for the purpose of eye movement classification: First, our algorithm uses an unsupervised classifier and thus does not require human-coded training data. This independence also allows gazeHMM to adapt well to interindividual differences in gaze behavior. Second, gazeHMM uses a parametric model (i.e., an HMM) and relies on maximum likelihood estimation, which enables model comparisons and testing parameter constraints. This property has been rarely used in eye movement event models. Third, it classifies the most relevant eye movement events, namely, fixations, saccades, PSOs, and smooth pursuits. Additionally, gazeHMM gives the user the option to only classify the first two or the first three of these events, a feature that most other algorithms do not have. As a minor goal, we aimed to reduce the number of thresholds which users have to set to a minimum.

The following section describes gazeHMM and the underlying generative

			Event			
Algorithm	Unsupervised	Parametric	Fixations	Saccades	PSOs	Pursuits
gazeHMM	X	X	X	X	X	X
BP-AR-HMM	X	X				
NSLR-HMM		X	X	X	X	X
I2MC	X		X			
U'n'Eye		X		X		
IRF			X	X	X	
gazeNet		X	X	X	X	
CNN-BLTSM		X	X	X		X

Table 1.1: Recently Developed Algorithms for Eye Movement Classification. X means that an algorithm has the respective property or classifies the respective event. BP-AR-HMM = beta-process autoregressive HMM (Houpt et al., 2018); NSLR-HMM = naive segmented linear regression HMM (Pekkanen & Lappi, 2017); I2MC = identification by two-means clustering (R. S. Hessels et al., 2017); U'n'Eye by Bellet et al. (2019); IRF = identification by random forest (Zemblys et al., 2018); gazeNet by Zemblys et al. (2019); CNN-BLTSM = convolutional neural network bidirectonal long short-term memory (Startsev et al., 2019).

model in detail. Then, we present the parameter recovery of the HMM and show how the algorithm performs compared to other eye movement event classification algorithms concerning the agreement to human coding. Importantly, we did not compare gazeHMM to supervised algorithms due to the training requirements of these methods. Finally, we discuss these results and propose directions in which gazeHMM and other HMM algorithms could be improved.

1.2 Developing gazeHMM

As illustrated in Figure 1.1, most eye movement event classification algorithms consist of three steps (cf. R. S. Hessels et al., 2017): During *preprocessing*, features (such as velocity and acceleration) are extracted from the raw gaze positions. Often, a filtering or smoothing procedure is applied to the data, before or after the transformation, to separate the gaze signal from noise and artifacts

(Spakov, 2012). Then follows the *classification*, depending on the method and settings of the algorithm, each sample is labeled as a candidate for one of the predefined events. Lastly, as part of the *postprocessing*[3], the algorithm decides which candidates to accept, relabel, or merge (R. S. Hessels et al., 2017; Komogortsev et al., 2010).

1.2.1 Preprocessing

Algorithms require variables that describe gaze data (hereafter called *eye movement features*) to classify them into events. Many eye movement features have been proposed and used in previous algorithms (for examples, see Andersson et al., 2017; Zemblys et al., 2018), but most of them rely on thresholds or window ranges that have to be set by the user (e.g., the distance between the mean position in a 100 ms window before and after each sample, see Olsson, 2007). This can be problematic because such parameters are often set without theoretical justification and differ substantially between features or heavily depend on the eye-tracker's characteristics (e.g., sampling frequency, Andersson et al., 2017). In gazeHMM, we used velocity, acceleration, and sample-to-sample angle (synonymous to relative or change in angle Larsson et al., 2013) because they belong to the most basic features which do not require additional parameter settings.

Theoretically, these three features should separate eye movement events, depending on one's definitions (R. S. Hessels et al., 2018). In the present work, we assume eye-tracking applications with fixed head position (chin-rest), gazing at a fixed display with a stationary eye-tracker. Fixations typically show samples with low velocity and acceleration. Due to tremor, we assume that the angle between samples should not follow any direction but a uniformly random walk. In contrast, saccade samples usually have a high velocity and acceleration and roughly follow the same direction. PSO samples tend to have moderate velocity and high acceleration since they occur between saccades and low-velocity events (Larsson et al., 2013, 2015). They can be specifically distinguished by their

[3] R. S. Hessels et al. (2017) called step two the *search rule* and step three the *classification rule*. For non-parametric methods, this distinction might be accurate. However, for parametric methods, calling step two "classification" is more appropriate since the probabilistic classification is done here. Step three usually consists of some heuristic relabeling and correcting for classification errors.

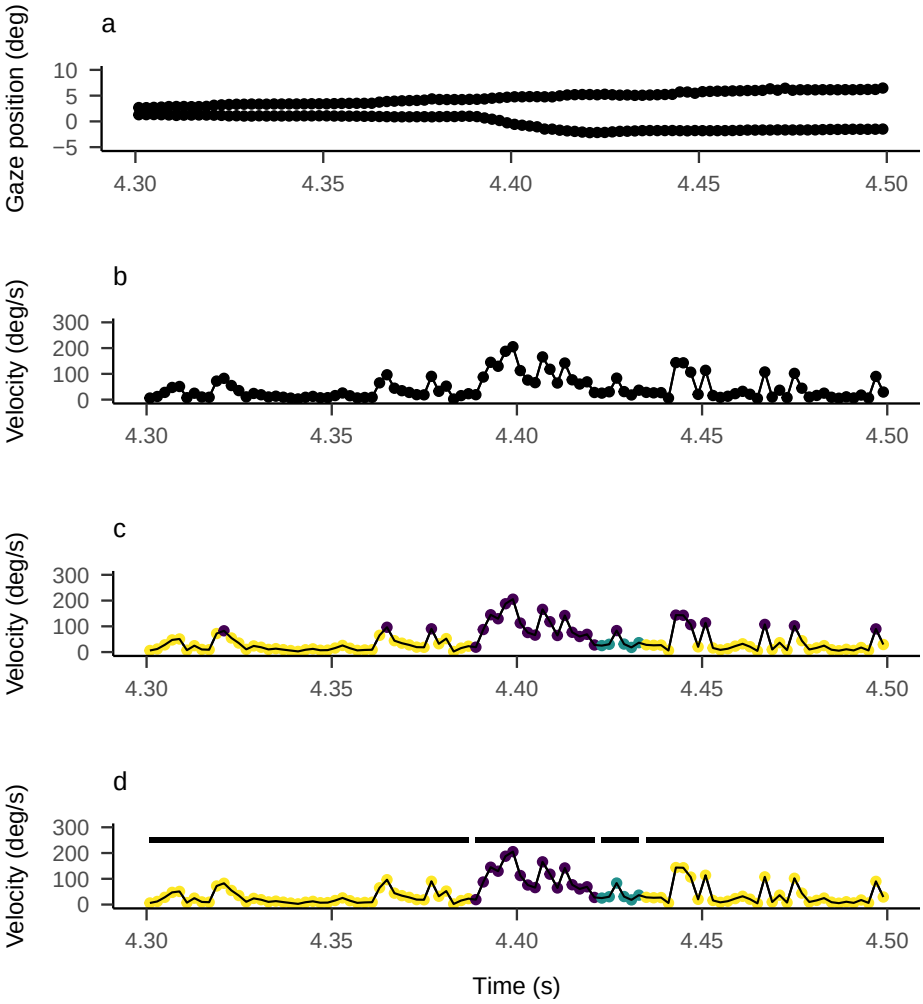


Figure 1.1: Example Workflow for Eye Movement Event Classification Algorithms. Workflow description: (a) the raw gaze signal in x (upper line) and y (lower line) coordinates; (b) the raw gaze signal is filtered and transformed into a velocity signal; (c) samples are classified as events (indicated by colors), and (d) relabeled. Sequences of samples belonging to the same event are merged (indicated by black segments). Data from Andersson et al. (2017).

change in direction clustered around 180 degrees (Pekkanen & Lappi, 2017). Importantly, the feature distribution during oscillations depends on the resolution of the gaze recording: Eye-trackers with higher sampling frequency yield more changes in direction and more samples in between those changes. Those samples in between typically follow the same direction. Thus, with high sampling frequencies, PSO samples might also cluster around a sample-to-sample angle of zero with outliers around 180 degrees. Lastly, smooth pursuit samples have a moderate velocity but low acceleration (due to the smoothness) and like saccades, they follow a similar direction (Larsson et al., 2013; Leigh & Zee, 2015). Other algorithms focus exclusively on classifying microsaccades (e.g., Mihali et al., 2017), but as stated earlier, these events were not in the scope of gazeHMM. The velocity and acceleration signals are computed from the raw gaze position by using a Savitzky-Golay filter (similar to Nyström & Holmqvist, 2010; Savitzky & Golay, 1964). The sample-to-sample angle is calculated as:

$$\alpha(t) = \arctan\left(\frac{y_{t+1} - y_t}{x_{t+1} - x_t}\right) - \arctan\left(\frac{y_t - y_{t-1}}{x_t - x_{t-1}}\right), \tag{1.1}$$

with $\alpha(t) := \alpha(t) + 2\pi$ for $\alpha(t) < 0$, and is therefore bound between 0 and 2π. Most of the missing data in eye movement data are due to blinks. In gazeHMM, we do not consider blinks as an additional event but rather as another source of noise. Therefore, the user can provide an indicator for samples that should be labeled as blinks (e.g., based on automated blink detection through the eye-tracker). Often, eye-trackers record a few samples with unreasonably high velocity and acceleration before losing the pupil signal when a blink occurs. Since these samples could distort the classification of saccades in the HMM, gazeHMM removes them heuristically. Before classifying the samples, it sets all samples within 50 ms before and after blink samples as missing. We note that this arbitrary setting is undermining our development goal of requiring as few user settings as possible. However, when we included blinks in the generative model itself, the classification of the other events became worse. Thus, we justify the heuristic blink removal by its accuracy, simplicity, and practicality. Furthermore, we experienced during the development that the default setting of 50 ms was appropriate for all data we examined.

1.2.2 The Generative Model

We denote the three eye movement features by X, Y, and Z. Each feature was generated by a hidden state variable S. Given S, the HMM treats X, Y, and Z as conditionally independent. Conditional independence might not accurately resemble the relationship between velocity and acceleration (which are naturally correlated). This step was merely taken to keep the HMM simple and identifiable. In gazeHMM, S can take one of two, three, or four hidden states. By selecting appropriate default starting values for the states (see Table 1.4), the algorithm is nudged to associate them with the same eye movement events. We remark that gazeHMM does not guarantee a consistent correspondence between states and events (see the phenomenon of label switching in the simulation study discussion). However, when applying gazeHMM to eye movement data, we did not encounter any problems in this regard. Moreover, gazeHMM comes with tools for a 'sanity check' to confirm whether expected and estimated state characteristics match (i.e., the HMM converged to an appropriate solution). Given correct identification, the first state represents fixations, the second saccades, the third PSOs, and the fourth smooth pursuits. Thus, users can choose whether they would like to classify only fixations and saccades, or additionally PSOs and/or smooth pursuits. HMMs can be described by three submodels: An initial state model, a transition model, and a response model. The initial state model contains probabilities for the first state of the hidden sequence $\rho_i = P(S_1 = i)$, with i denoting the hidden state. In gazeHMM, the initial states are modeled by a multinomial distribution. The evolution of the sequence is in turn described by the transition model, which comprises the probabilities for transitioning between different states in the HMM. Typically, probabilities to transition from state i to j, $a_{ij} = P(S_{t+1} = j | S_t = i)$, are expressed in matrix form (Visser, 2011):

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{i1} & \dots & a_{ij} \end{bmatrix}.$$

Again, the transition probabilities for each state are modeled by multinomial distributions. The response model encompasses distributions describing the

response variables for every state in the model. Previous algorithms have used Gaussian distributions to describe velocity and acceleration signals (sometimes after log-transforming them). However, several reasons speak against choosing the Gaussian: First, both signals are usually positive (depending on the computation). Second, the distributions of both signals appear to be positively skewed conditionally on the states, and third, to have variances increasing with their mean. Thus, instead of using the Gaussian, it could be more appropriate to describe velocity and acceleration with a distribution that has these three properties. In gazeHMM, we use gamma distributions with a shape and scale parametrization for this purpose:

$$
\begin{aligned}
(X \mid S = i) &\sim \text{Gamma}(\alpha_{xi}, \beta_{xi}) \\
(Y \mid S = i) &\sim \text{Gamma}(\alpha_{yi}, \beta_{yi}),
\end{aligned}
$$

with i denoting the hidden state. When we developed gazeHMM, the gamma distribution appeared to fit eye movement data well, but we also note that it might not necessarily be the best fitting distribution for every type of eye movement data. We assume that the best fitting distribution will depend on the task, eye-tracker, and individual (see discussion). We emphasize that gazeHMM does not critically depend on the choice of distribution and other distributions than the gamma can be readily included in the model, for example the log-normal has the same required properties of being positive and positively skewed. To model the sample-to-sample angle, we pursued a novel approach in gazeHMM: A mixture of von Mises distributions (with a mean and concentration parameter) and a uniform distribution:

$$
\begin{aligned}
(Z \mid S = 1) &\sim U(0, 2\pi) \\
(Z \mid S = 2) &\sim \text{von Mises}(\mu_1, \kappa_1) \\
(Z \mid S = 3) &\sim \text{von Mises}(\mu_2, \kappa_2) \\
(Z \mid S = 4) &\sim \text{von Mises}(\mu_3, \kappa_3).
\end{aligned}
$$

Both the distributions and the feature operate on the full unit circle (i.e., between 0 and 2π), which should lead to symmetric distributions. Von Mises is a maximum entropy distribution on a circle under a specified location and

concentration, and can be considered an analogue to the Gaussian distribution in circular statistics (Mardia & Jupp, 2009). Because we assume fixations to change their direction similar to a uniformly random walk (Larsson et al., 2013, 2015), their sample-to-sample angle can be modeled by a uniform distribution. Thus, the uniform distribution should distinguish fixations from the other events. Taking all three submodels together, the joint likelihood of the observed data and hidden states can be expressed as:

$$L(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{S}|\lambda) = \rho_{S_1} f_{S_1}(X_1) f_{S_1}(Y_1) f_{S_1}(Z_1) \prod_{t=1}^{T-1} a_{S_t S_{t+1}} f_{S_{t+1}}(X_{t+1}) f_{S_{t+1}}(Y_{t+1}) f_{S_{t+1}}(Z_{t+1}), \quad (1.2)$$

with λ denoting the vector containing the initial state and transition probabilities as well as the response parameters. By summing over all possible state sequences, the likelihood of the data given the HMM parameters becomes (Visser, 2011):

$$L(\mathbf{X}, \mathbf{Y}, \mathbf{Z}|\lambda) = \sum_{\text{all S}} \rho_{S_1} f_{S_1}(X_1) f_{S_1}(Y_1) f_{S_1}(Z_1) \prod_{t=1}^{T-1} a_{S_t S_{t+1}} f_{S_{t+1}}(X_{t+1}) f_{S_{t+1}}(Y_{t+1}) f_{S_{t+1}}(Z_{t+1}). \quad (1.3)$$

The parameters of the HMM are estimated through maximum likelihood using an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 1997). The EM algorithm is generally suitable to estimate likelihoods with missing variables. For HMMs, it imputes missing with expected values and iteratively maximizes the joint likelihood of parameters conditional on the observed data and the expected hidden states (i.e., eye movement events Visser, 2011). When evaluating the likelihood of missing data, gazeHMM integrates over all possible values, which results in a probability density of one. The sequence of hidden states is estimated through the Viterbi algorithm (Forney Jr, 1973; Viterbi, 1967) by maximizing the posterior state probability. Parameters of the response distributions (except for the uniform distribution) are optimized on the log-scale (except for the mean param-

eter of the von Mises distribution) using a spectral projected gradient method (Birgin, Martinez, & Raydan, 2000) and Barzilai-Borwein step lengths (Barzilai & Borwein, 1988). The implementation in depmixS4 allows to include time-varying covariates for each parameter in the HMM. In gazeHMM, no such covariates were included and thus, only intercepts were estimated for each parameter.

1.2.3 Postprocessing

After classifying gaze samples into states, gazeHMM applies a postprocessing routine to the estimated state sequence. We implemented this routine because in a few cases, gazeHMM would classify samples that were not following saccades as PSOs. Constraining the probabilities for nonsaccade events to turn into PSOs to zero often caused PSOs not to appear in the state sequence at all. Moreover, gazeHMM does not explicitly control the duration of events in the HMM which occasionally led to unreasonably short events. Thus, the postprocessing routine heuristically compensates for such violations. This routine relabels one-sample fixations and smooth pursuits, saccades with a duration below a minimum threshold (here: 10 ms), and PSOs that follow nonsaccade events. Samples are relabeled as the state of the previous event. Finally, samples initially indicated as missing are labeled as noise (including blinks) and event descriptives are computed (e.g., fixation duration).

The algorithm is implemented in R (version: 3.6.3 R Core Team, 2020) and uses the packages `signal` (Ligges, Short, & Kienzle, 2015) to compute velocity and acceleration signals, `depmixS4` (Visser, 2011) for the HMM, `CircStats` (Lund & Agostinelli, 2018) for the von Mises distribution, and `BB` (Varadhan & Gilbert, 2009) for Barzilai-Borwein spectral projected gradient optimization. The algorithm is available on GitHub (`github.com/maltelueken/gazeHMM`). We conducted a parameter recovery study that is also available on GitHub (`github.com/maltelueken/gazeHMM_validation`) showing that the model recovers parameters well when the noise level is not too high.

1.3 Simulation Study

As a first step to validate the model, we need to ensure that fitting the model to the data results in recovering the properties of the underlying data generating process. The standard procedure in computational modeling is conducting parameter recovery study (Heathcote, Brown, & Wagenmakers, 2015). Although this step is crucial when developing new models, it is often not done or goes unreported in eye-tracking literature. To counter this trend, we report a simulation study we conducted to assess the recovery of parameter values and state sequences. The design and analysis of the study were preregistered on the Open Science Framework (`osf.io/vdjgp`). The majority of this section is copied from the preregistration (with adapted tenses). The study was divided in four parts. Here, we only report the first two parts, which investigate the influence of parameter variation and adding noise to generated data on recovery. The other two parts, which address starting values and missing data, can be found in the supplementary material `github.com/maltelueken/gazeHMM_validation`). The HMM repeatedly generated data with a set of parameters (henceforth: true parameter values). An example of the simulated data is shown in Figure 1.2.The same model was applied to estimate the parameters from the generated data (henceforth: estimated parameter values). We compared the true with the estimated parameter values to assess whether a parameter was recovered by the model. Additionally, we contrasted the true states of the HMM with the estimated states to judge how accurately the model recovered the states that generated the data.

1.3.1 Starting Values

The HMM always started with a uniform distribution for the initial state and state transition probabilities. Random starting values for the estimation of shape, scale, and concentration parameters were generated by gamma distributions with a shape parameter of $\alpha_{start} = 3$ and $\beta_{start;i} = \theta_i/2$, with θ_i being the true value of the parameter to be estimated in simulation $i \in (1, \ldots, I)$. This setup ensured that the starting values were positive, their distributions were moderately skewed, and the modes of their distributions equaled the true parameter values. The mean parameters of the von Mises distribution always

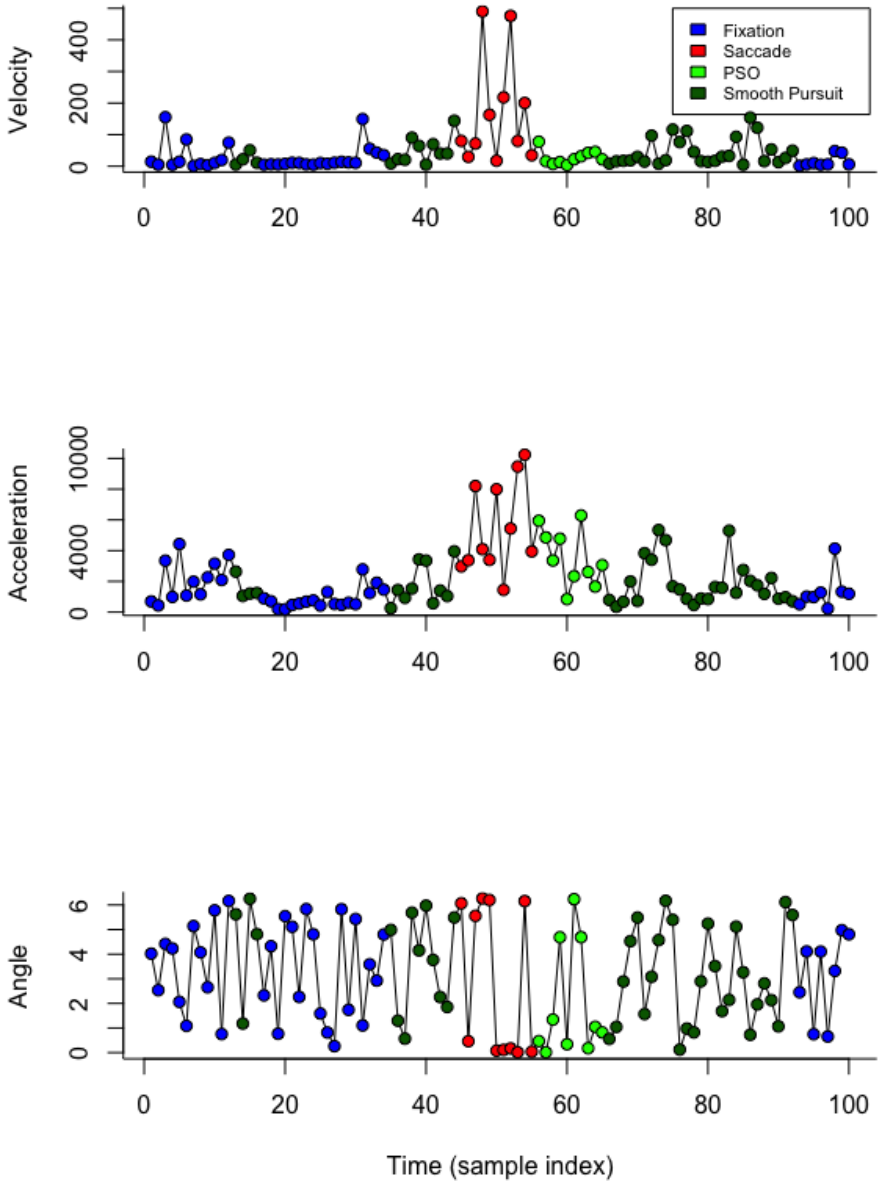


Figure 1.2: Example of data simulated from gazeHMM.

	ρ_i	$a_{i=j}$	$a_{i\neq j}$
Interval	-	[0.01,0.99]	$(1 - a_{i=j})/(k-1)$
Default	1/k	0.9	0.1/(k-1)

Table 1.2: Intervals and Default Parameter Values for the Transition Model in the Simulation Study. The initial state probability is denoted by ρ_i. The transition probability for staying in the same state is denoted by $a_{i=j}$ and the probability for switching to a different state by $a_{i\neq j}$. The number of states in the model is denoted by k.

started at their true values.

1.3.2 Design

In the first part, we varied the parameters of the HMM. For models with $k \in \{2, 3, 4\}$ states, $q \in \{10, 15, 20\}$ parameters were manipulated, respectively. For each parameter, the HMM generated 100 data sets with $N = 2500$ samples, and the parameter varied in a specified interval in equidistant steps. This resulted in $100 \times (10 + 15 + 20) = 4500$ recoveries. Only one parameter alternated at once, the other parameters were set to their default values. All parameters of the HMM were estimated freely (i.e., there were no fixed parameters in the model). We did not manipulate the initial state probabilities because these are usually irrelevant in the context of eye movement classification. For the transition probabilities, we only simultaneously changed the probabilities for staying in the same state (diagonals of the transition matrix) to reduce the complexity of the simulation. The leftover probability mass was split evenly between the probabilities for switching to a different state (per row of the transition matrix). Moreover, we did not modify the mean parameters of the von Mises distributions: As location parameters, they do not alter the shape of the distribution and they are necessary features for the HMM to distinguish between different states.

We defined approximate ranges for each response variable (see supplementary material) and chose true parameter intervals and default values so that they produced samples that roughly corresponded to these ranges. Tables 1.2 and

1.3 show the intervals and default values for each parameter in the simulation. Parameters were scaled down by factor 10 (compared to the reported ranges) to improve fitting of the gamma distributions. We set the intervals for shape parameters of the gamma distributions for all events to [1,5] to examine how skewness influenced the recovery (shape values above five approach a symmetric distribution). The scale parameters were set so that the respective distribution approximately matched the assumed ranges. Since the concentration parameters of the von Mises distribution are the inverse of standard deviations, they were varied on the inverse scale.

In the second part, we manipulated the sample size of the generated data and the amount of noise added to it. The model parameters were set to their default values. For models with $k \in \{2, 3, 4\}$ states and sample sizes of $N \in \{500, 2500, 10000\}$, we generated 100 data sets ($100 \times 3 \times 3 = 900$ recoveries). These sample sizes roughly match small, medium, and large eye-tracking data sets for a single participant and trial (e.g., with a frequency of 500 Hz, the sample sizes would correspond to recorded data with lengths of 1 s, 5 s, and 20 s, respectively). To simulate noise, we replaced velocity and acceleration values y with draws from a gamma distribution with $\alpha_{noise} = 3$ and $\beta_{noise} = (y/2)\tau_{noise}$ with $\tau_{noise} \in [1, 5]$ varying between data sets. This procedure ensured that velocity and acceleration values remained positive and were taken from moderately skewed distributions with modes equal to the original values. To angle, we added white noise from a von Mises distribution with $\mu_{noise} = 0$ and $\kappa_{noise} \in 1/[0.1, 10]$ varying between data sets. τ_{noise} and κ_{noise} were increased simultaneously in equidistant steps in their intervals.

	Velocity		Acceleration		Rel. angle	
	α	β	α	β	μ	κ
State 1						
Interval	[1,5]	[0.1,0.6]	[1,5]	[0.05,0.25]	-	-
Default	3	0.35	3	0.25	-	-
State 2						
Interval	[1,5]	[5,15]	[1,5]	[1,5]	-	1/[0.1,10]
Default	3	10	3	3	0	1
State 3						
Interval	[1,5]	[0.5,1.5]	[1,5]	[1,5]	-	1/[0.1,10]
Default	3	1	3	3	π	1
State 4						
Interval	[1,5]	[0.5,1.5]	[1,5]	[0.05,0.25]	-	1/[0.1,10]
Default	3	1	3	0.15	0	1

Table 1.3: Intervals and Default Parameter Values for the Response Model in the Simulation Study. Shape parameters are denoted by α, scale parameters by β, mean parameters by μ, and concentration parameters by κ. The default values for the uniform distribution in state one were min = 0 and max = 2π.

1.3.3 Data Analysis

For each parameter separately, we calculated the root median square proportion deviation (RMdSPD; analogous to root median square percentage errors, see Hyndman & Koehler, 2006) between the true and estimated parameter values:

$$\text{RMdSPD} = \sqrt{\text{Median}(\epsilon_1^2, \ldots, \epsilon_I^2)} \tag{1.4}$$

$$\epsilon_i^2 = \left(\frac{\hat{\theta}_i - \theta_i}{\theta_i} \right)^2, \tag{1.5}$$

where θ_i is the true parameter value and $\hat{\theta}_i$ is the estimated parameter value for simulation $i \in (1, \ldots, I)$, respectively. Even though it was not explicitly mentioned in the preregistration, this measure is only appropriate when $\theta_i \neq 0$.

This was not the case for some mean parameters of the von Mises distributions. In those cases, we used $\theta_i = 2\pi$ instead. We treated RMdSPD < 0.1 as good, $0.1 \leq$ RMdSPD < 0.5 as moderate, and RMdSPD ≥ 0.5 as bad recovery of a parameter. By taking the median, we reduced the influence of potential outliers in the estimation and using proportions enabled us to compare RMdSPD values across parameters and data sets.

Additionally, we applied a bivariate linear regression with the estimated parameter values as the dependent and the true parameter values as the independent variable to each parameter that has been varied on an interval in part one. Regression slopes closer to one indicated that the model better captured parameter change. Regression intercepts different from zero reflected a bias in parameter estimation.

To assess state recovery, we computed Cohen's kappa (for all events taken together, not for each event separately) as a measure of agreement between true and estimated states for each generated data set. Cohen's kappa estimates the agreement between two classifiers accounting for the agreement due to chance. Higher kappa values were interpreted as better model accuracy. We adopted the ranges proposed by Landis and Koch (1977) to interpret kappa values. Models that could not be fitted were excluded from the recovery.

1.3.4 Results

Parameter Variation

In the first part of the simulation, we examined how varying the parameters[4] in the HMM affected the deviation of estimated parameters and the accuracy of estimated state sequences. For the two-state HMM, the recovery of parameters and states was nearly perfect (all RMdSPDs < 0.1, intercepts and slopes of regression lines almost zero and one, respectively, and Cohen's kappa close to 1). Therefore, we chose to include the respective figures in the supplementary material.

For the HMM with three states, the RMdSPD is shown in Figure 1.3. When

[4]Note that the initial state probability ρ_i has RMdSPD $= 1$. Since the HMM only simulated one state sequence, this parameter is always either zero or one (leading to RMdSPD $= 1$). Therefore, we decided to exclude it from the analysis.

response parameters (other than $a_{i=j}$) were manipulated, the RMdSPDs for a_{12} and a_{31} were consistently between 0.1 and 0.5. Varying κ in states two and three led to RMdSPDs between 0.1 and 0.5 in the respective states, which we interpreted as moderate recovery. Otherwise, RMdSPDs were consistently lower than 0.1, indicating good recovery. Inspecting the regression lines between true and estimated parameters (see Figures 1.4 and 1.5) revealed strong and unbiased linear relationships (intercepts close to zero and slopes close to one). In contrast to the two-state HMM, larger deviations and more outliers were observed. Cohen's kappa values are presented in Figure 1.6. For most estimated models, the kappa values between true and estimated state sequences were above 0.95, meaning almost perfect agreement. However, for some models, we observed kappas clustered around zero or -0.33, which is far from the majority of model accuracies. An exploratory examination of these clusters suggests that state labels were switched (see supplementary material).

The RMdSPDs for the four-state HMM is shown in Figure 1.7. For estimated transition probabilities and α_{vel} and β_{vel} parameters in states one and four, RMdSPDs were between 0.1 and 0.5, suggesting moderate recovery. Also, estimated kappa parameters in state four were often moderately recovered when parameters in states two, three, and four were varied. Otherwise, RMdSPDs were below 0.1, indicating good recovery. Looking at Figures 1.8 and 1.9, the regression lines between true and estimated parameters exhibit strong and unbiased relationships. However, there were larger deviations and more outliers than in the previous models, especially for states one and four. Cohen's kappa ranged mostly between 0.6 and 0.9, meaning moderate to almost perfect agreement between true and estimated state sequences (see Figure 1.10). Here, some outlying kappa values clustered around 0.25 and zero.

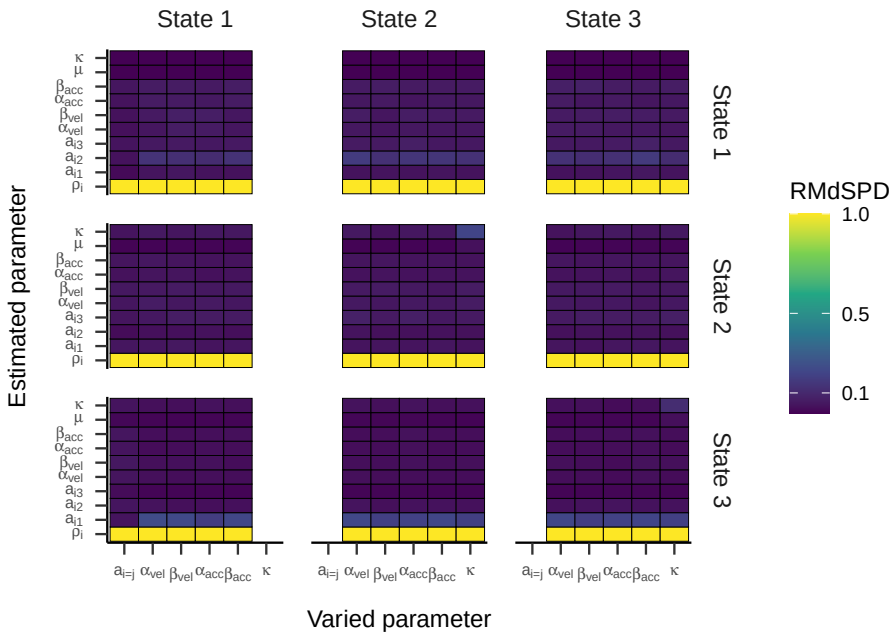


Figure 1.3: RMdSPD Between True and Estimated Parameters of the Three-State HMM in Part One of the Simulation. Labels on the x-axis indicate which true parameters have been manipulated and labels on the y-axis show for which estimated parameter the RMdSPD is displayed. Top facet labels specify in which state the parameters have been varied and right facet labels denote to which state estimated parameters belong. ρ_i is the initial probability for state i (indicated by the right facet label), $a_{i=j}$ is the probability to transition from state i to state j, α and β are the shape and scale parameters of the gamma distributions, and μ and κ are the mean and concentration parameter of the von Mises distribution.

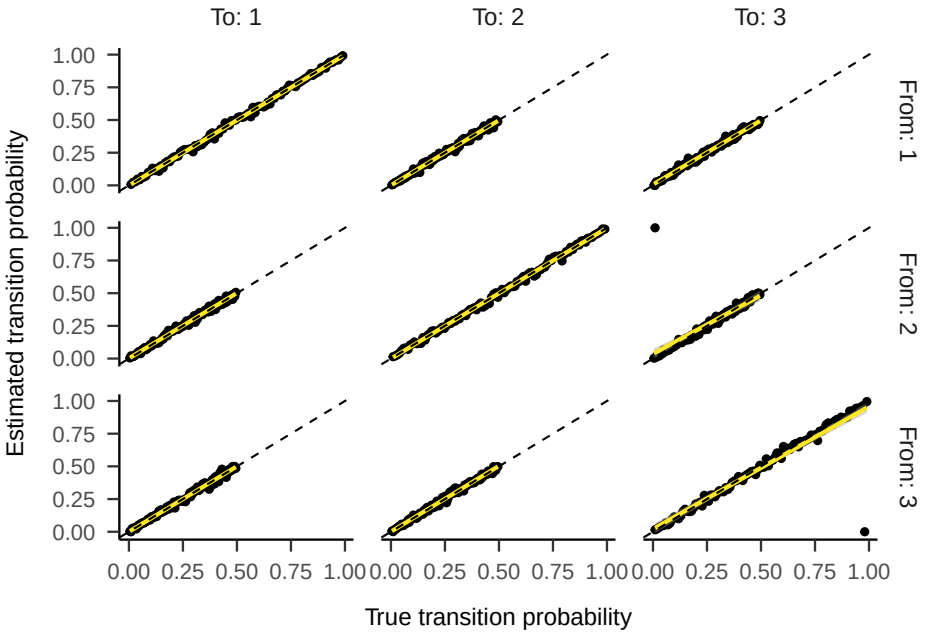


Figure 1.4: Regression Lines Between True and Estimated Transition Probabilities for the Three-State HMM in Part One of the Simulation. Right facet labels show *from* and top facet labels show *to* which state the HMM is moving. Dashed lines refer to perfect recovery.

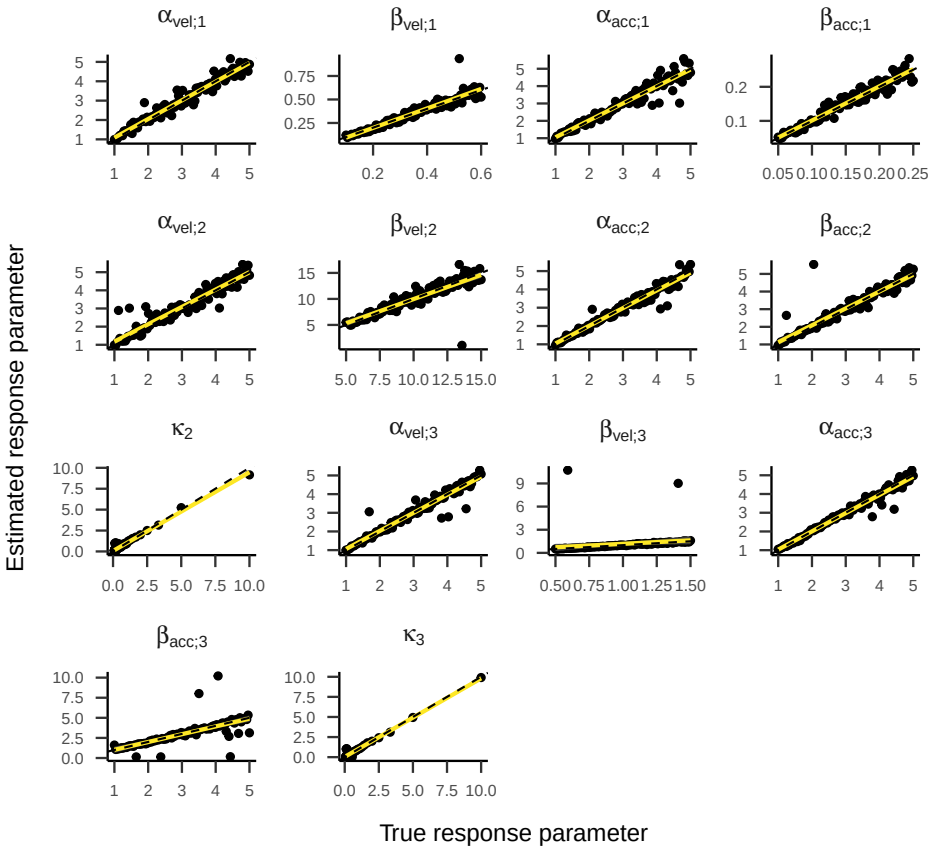


Figure 1.5: Regression Lines Between True and Estimated Response Parameters of the Three-State HMM in Part One of the Simulation. Top facet labels indicate response parameters. Dashed lines refer to perfect recovery. α and β are the shape and scale parameters of the gamma distributions, and κ is the concentration parameter of the von Mises distribution. Parameter subscripts indicate states and eye movement features.

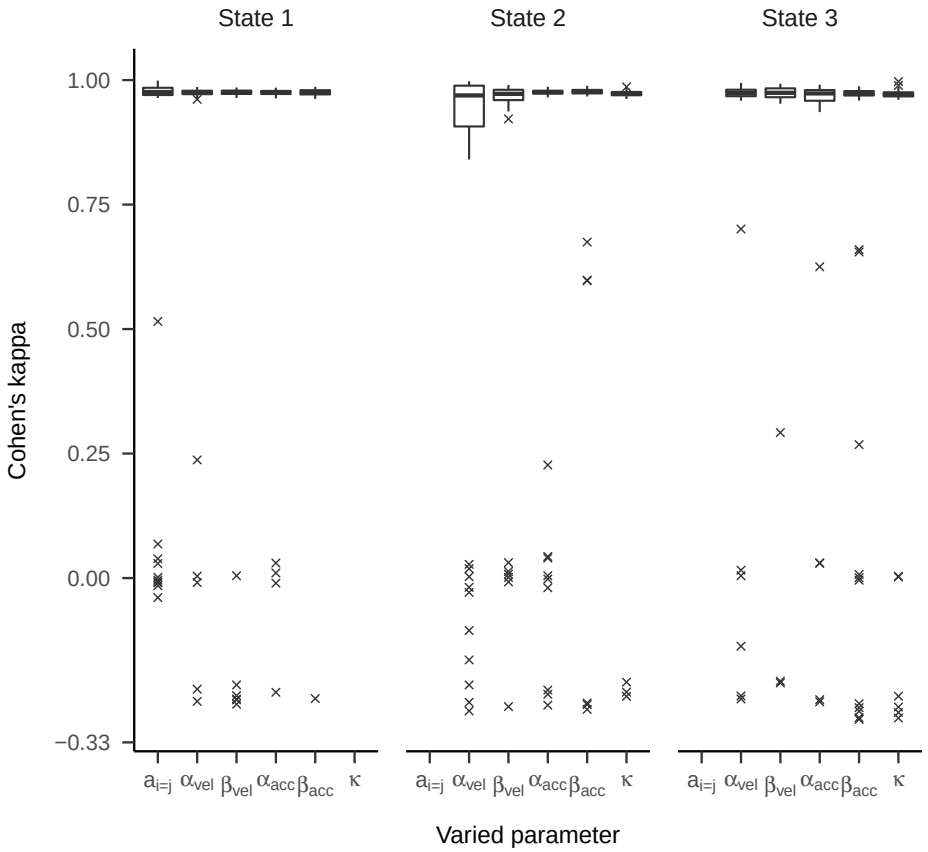


Figure 1.6: Cohen's Kappa Depending on Which Parameter of the Three-State HMM Has Been Manipulated in Part One of the Simulation. Top facet labels indicate for which state parameters have been manipulated. Black solid lines symbolize medians and hinges the first and third quartile. Whiskers range from hinges to lowest/highest value within 1.5 times the IQR. Crosses represent outliers. $a_{i=j}$ is the probability to stay in the same state, α and β are the shape and scale parameters of the gamma distributions, and μ and κ are the mean and concentration parameter of the von Mises distribution.

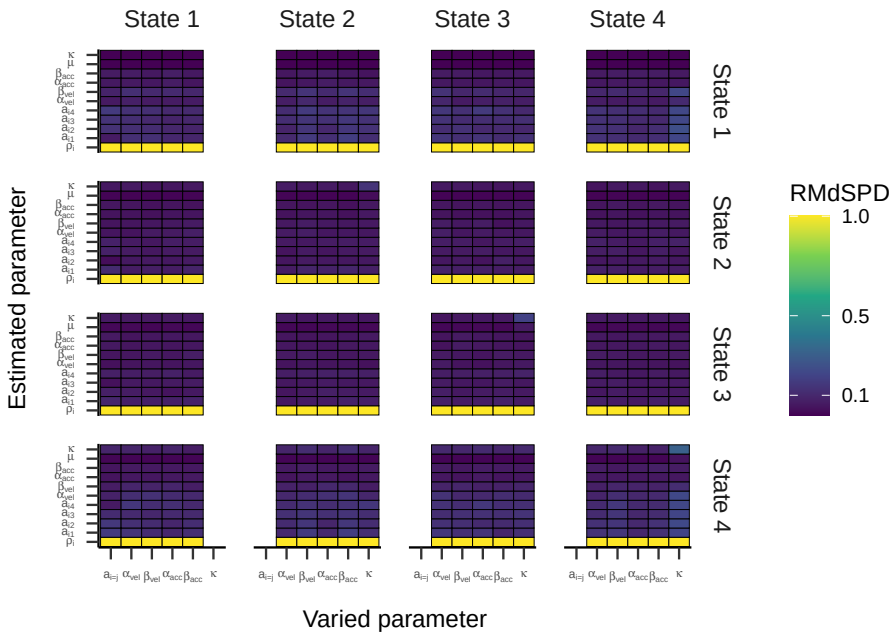


Figure 1.7: RMdSPD Between True and Estimated Parameters of the Four-State HMM in Part One of the Simulation. Labels on the x-axis indicate which true parameters have been manipulated and labels on the y-axis show for which estimated parameter the RMdSPD is displayed. Top facet labels specify in which state the parameters have been varied and right facet labels denote to which state estimated parameters belong. ρ_i is the initial probability for state i (indicated by the right facet label), $a_{i=j}$ is the probability to transition from state i to state j, α and β are the shape and scale parameters of the gamma distribu- tions, and μ and κ are the mean and concentration parameter of the von Mises distribution.

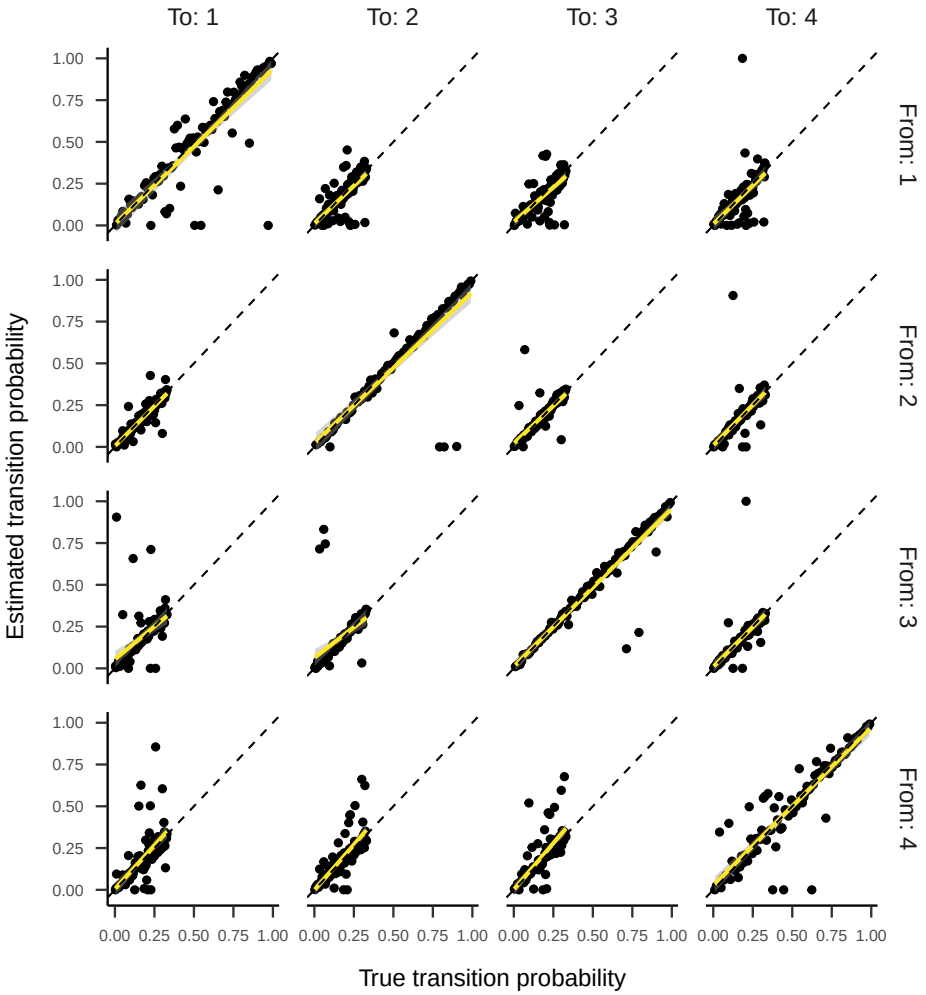


Figure 1.8: RMdSPD Between True and Estimated Parameters of the Four-State HMM in Part One of the Simulation. Right facet labels show *from* and top facet labels show *to* which state the HMM is moving. Dashed lines refer to perfect recovery.

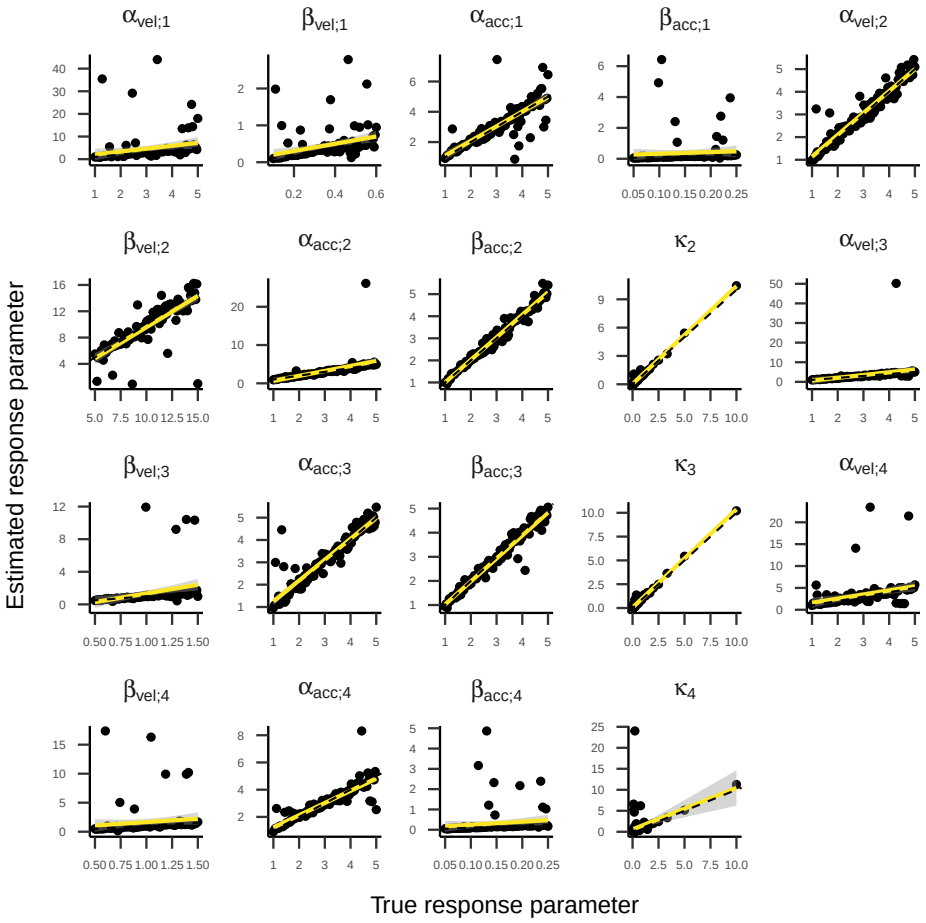


Figure 1.9: Regression Lines Between True and Estimated Response Parameters of the Four-State HMM in Part One of the Simulation. Top facet labels indicate response parameters. Dashed lines refer to perfect recovery. α and β are the shape and scale parameters of the gamma distributions, and μ and κ are the mean and concentration parameter of the von Mises distribution. Parameter subscripts indicate states and eye movement features.

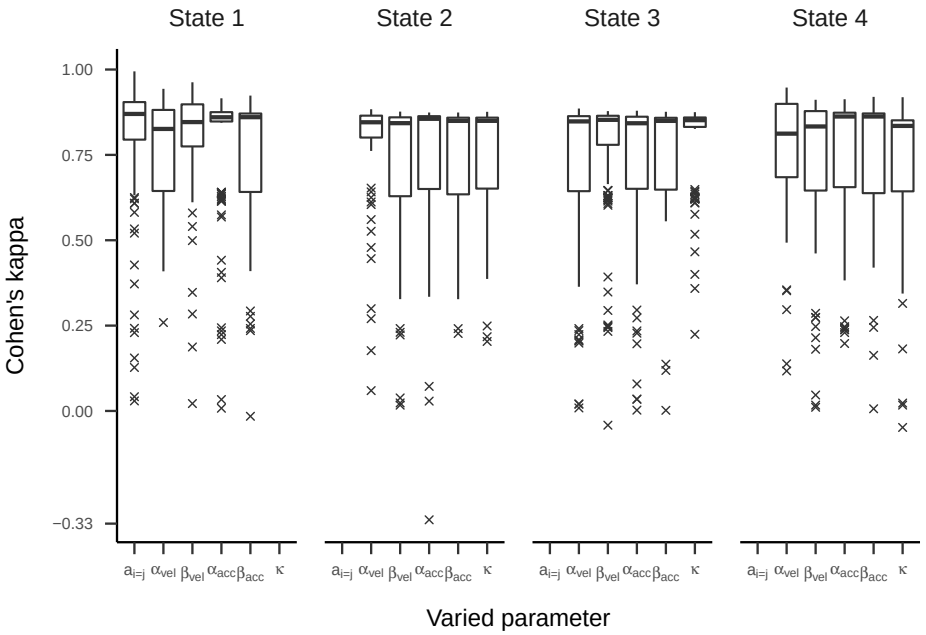


Figure 1.10: Cohen's Kappa Depending on Which Parameter of the Four-State HMM Has Been Manipulated in Part One of the Simulation. Top facet labels indicate for which state parameters have been manipulated. Black solid lines symbolize medians and hinges the first and third quartile. Whiskers range from hinges to lowest/highest value within 1.5 times the IQR. Crosses represent outliers. $a_{i=j}$ is the probability to stay in the same state, α and β are the shape and scale parameters of the gamma distributions, and μ and κ are the mean and concentration parameter of the von Mises distribution.

Sample Size and Noise Variation

In the second part, we varied the sample size of the HMM and added noise to the generated data. For the two-state HMM, the RMdSPDs were above 0.5 for β_{vel} and β_{acc} in both states (see Figure 1.11), suggesting bad recovery. The other estimated parameters showed RMdSPDs close to or below 0.1, which means they were recovered well. Increasing the sample size seemed to improve RMdSPDs for most parameters slightly. For β_{vel} and β_{acc} in both states, models with 2500 samples had the lowest RMdSPDs. Accuracy measured by Cohen's kappa was almost perfect with kappa values very close to one (see Figure 1.12, left plot).

For three states, the RMdSPDs for the β_{vel} and β_{acc} were above 0.5 in all three states (see Figure 1.13), indicating bad recovery. Again, the other estimated parameters were below or close to 0.1, only a_{12} and a_{31} with 500 samples were closer to 0.5. For most parameters across all three states, models with higher sample sizes had lower RMdSPDs. The state recovery of the estimated models was almost perfect with most kappa values above 0.95 (see Figure 1.12, middle plot). Several outliers clustered around kappa values of zero and -0.33.

RMdSPDs regarding the four-state HMM are displayed in Figure 1.14. For states one and four, values for most parameters (including all transition probabilities) were above 0.5, suggesting bad recovery. Similarly, β_{vel} and β_{acc} in states two and three showed bad recovery. For states two and three, higher sample sizes showed slightly lower RMdSPDs. As in the previous part, most Cohen's kappa values ranged between 0.6 and 0.9, meaning substantial to almost perfect agreement between true and estimated states (Figure 1.12, right plot). Multiple outliers clustered around 0.25 or zero.

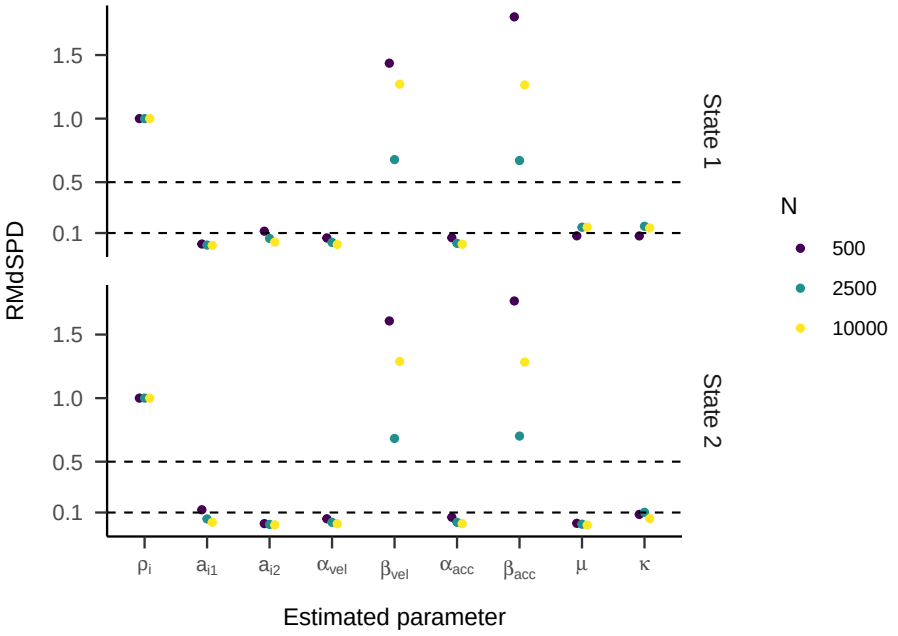


Figure 1.11: RMdSPD Between True and Estimated Parameters of the Two-State HMM in Part Two of the Simulation. Colours indicate different sizes of generated data. Labels on the y-axis indicate for which estimated parameter the RMdSPD is displayed. Right facet labels denote to which state estimated parameters belong. ρ_i is the initial probability for state i (indicated by the right facet label), $a_{i=j}$ is the probability to transition from state i to state j, α and β are the shape and scale parameters of the gamma distributions, and μ and κ are the mean and concentration parameter of the von Mises distribution.

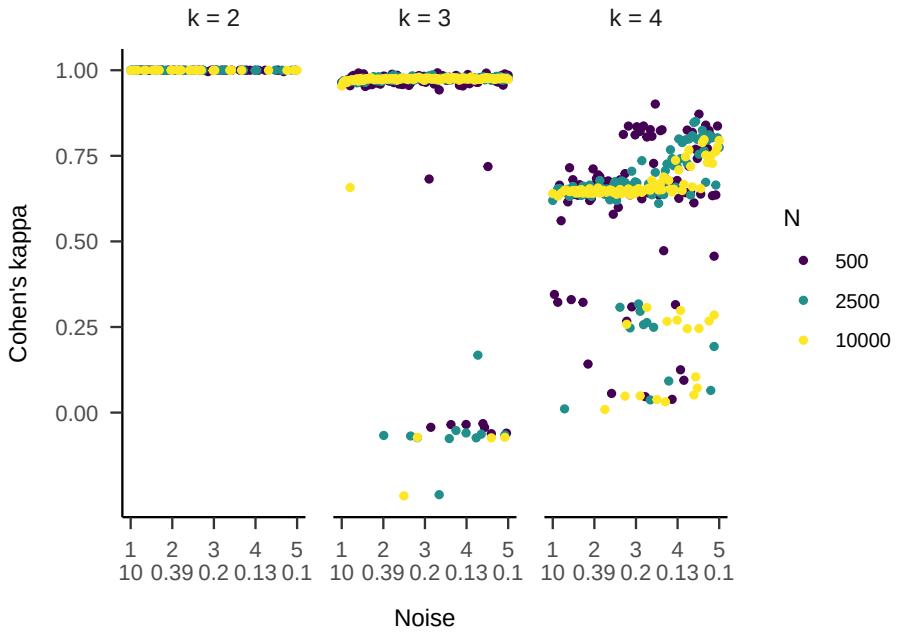


Figure 1.12: Cohen's Kappa Depending on the Variation of Noise Added to the Simulated Data. The upper labels on the x-axis indicate τ_{noise} and the lower labels κ_{noise}. Colours indicate different sizes of generated data. Top facet labels indicate the number of states in the HMM. ρ_i is the initial probability for state i (indicated by the right facet label), $a_{i=j}$ is the probability to transition from state i to state j, α and β are the shape and scale parameters of the gamma distributions, and μ and κ are the mean and concentration parameter of the von Mises distribution.

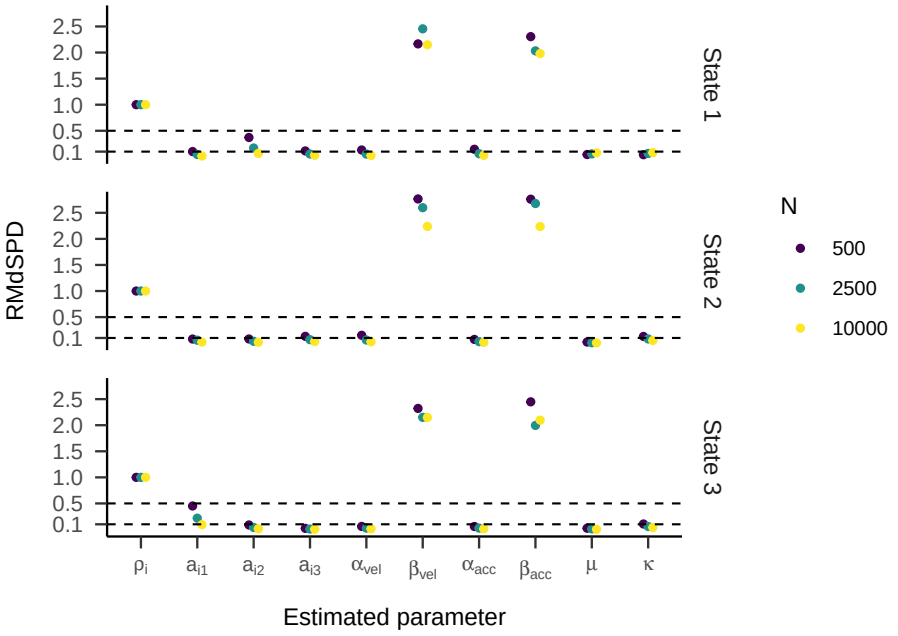


Figure 1.13: RMdSPD Between True and Estimated Parameters of the Three-State HMM in Part Two of the Simulation. Labels on the y-axis indicate for which estimated parameter the RMdSPD is displayed. Right facet labels denote to which state estimated parameters belong. ρ_i is the initial probability for state i (indicated by the right facet label), $a_{i=j}$ is the probability to transition from state i to state j, α and β are the shape and scale parameters of the gamma distributions, and μ and κ are the mean and concentration parameter of the von Mises distribution.

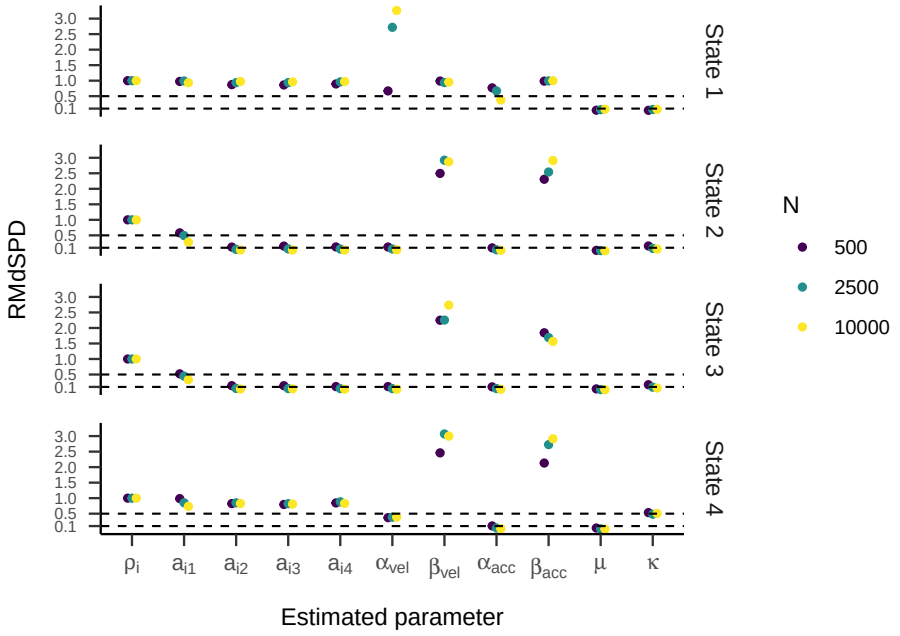


Figure 1.14: RMdSPD Between True and Estimated Parameters of the Four-State HMM in Part Two of the Simulation. Labels on the y-axis indicate for which estimated parameter the RMdSPD is displayed. Right facet labels denote to which state estimated parameters belong. ρ_i is the initial probability for state i (indicated by the right facet label), $a_{i=j}$ is the probability to transition from state i to state j, α and β are the shape and scale parameters of the gamma distributions, and μ and κ are the mean and concentration parameter of the von Mises distribution.

1.3.5 Discussion

In the simulation study, we assessed the recovery of parameters and hidden states in the generative model of gazeHMM. Simulations in part one demonstrated that the HMM recovered parameters very well when they were manipulated. Deviations from true parameters were mostly small. In the four-state model, estimated transition probabilities for state one and four deviated moderately. Moreover, the HMM estimated state sequences very accurately with decreasing accuracy for the four-state model. In the second part, noise was added to the generated data and the sample size was varied. Despite noise, the generative model was still able to recover most parameters well. However, in the four-state model, the parameter recovery for states one and four substantially decreased (even for low amounts of noise, see supplementary material). In the three- and four-state models, scale parameters of gamma distributions were poorly recovered (also even for low noise levels, see supplementary material). Increasing the sample size in the HMM slightly improved the recovery of most parameters. The state recovery of the model was slightly lowered when more states were included, but it was neither affected by the noise level nor the sample size. In the third part (included in the supplementary material), we showed that the variation in starting values used to fit the HMM did not influence parameter and state recovery. Missing data (in part four, also in the supplementary material) did not affect the parameter recovery but linearly decreased the recovery of hidden states. In all four parts, we observed clusters of outlying accuracy values. In part three, we exploratorily examined these clusters and reasoned that they can be attributed to label switching (i.e., flipping one or two state labels resolved the outlying clusters).

In general, the generative model recovers parameters and hidden states well and, thus, we conclude that it can be used in our classification algorithm. However, the recovery decreases when a fourth state (i.e., smooth pursuit) is added to the model and, especially with four states, many parameters in the HMM are vulnerable to noise. In the next sections, we will see how noise that is present in real eye movement data affects the performance of gazeHMM.

A limitation of this simulation study is that it only concerns the statistical part of the model, and investigates the ability of the model to recover the parameter values and state sequences. As such, the simulation study is an im-

plementation as well as feasibility check of the method. It does not, however, test accuracy of the final event labels, which are determined using the modeling output and postprocessing steps. Thus, the simulation might not be entirely realistic: for example, the generative statistical model is not constrained to allow PSO events follow only saccade events, and so this feature of the process would not be accounted for in the simulation results.

1.4 Validation Study

To validate gazeHMM, we applied the algorithm on two benchmark data sets. As starting values, we used $\rho = 1/k$ for the initial state model as well as $a_{i=j} = 0.9$ and $a_{i \neq j} = 0.1/k$ for the transition model. The values for the response model are displayed in Table 1.4. For a fifth eye movement event, we chose starting values that would enable the HMM to split any other event into two subevents (e.g., fixations into drift and microsaccades). In contrast to the simulation study, generating random starting values often led to bad model fits and label switching between states. To improve the fitting of the gamma distributions, velocity and acceleration signals were scaled down by factor 100, and so were the starting values for their gamma distributions[5].

1.4.1 Data Sets

We chose two data sets for validation: One was published in a study by Andersson et al. (2017) and has been widely used for validation purposes (Pekkanen & Lappi, 2017, e.g.,). It contains eye-tracking data from three conditions: A static condition, where subjects had to look freely at images, and two dynamic conditions, where they had to follow a constantly moving dot or objects in a video. The data were sampled with 500 Hz and two human coders (MN and RA) labeled them as belonging to six different eye movement events: Fixation, saccade, PSO, smooth pursuit, blink, or other. Andersson et al. (2017) used the data to compare 10 different classification algorithms. We adopted their results to compare these 10 algorithms and the two human coders with gazeHMM.

[5]Scaling down by factor 100 differs from the simulation study (scaling down by 10). The algorithm allows the user to manually specify this factor, and in this case, factor 100 led to better model fits than factor 10.

	Velocity		Acceleration		Rel. angle	
	α	β	α	β	μ	κ
Fixation	10	10	10	10	-	-
Saccade	50	50	50	50	0	10
PSO	50	50	50	50	π	10
Pursuit	20	20	20	20	0	10
Event 5	20	20	50	50	0	10

Table 1.4: Starting Values for the Response Model in the Validation Study. Starting values for velocity and acceleration signals are shown before scaling down by factor 100. Shape parameters are denoted by α, scale parameters by β, mean parameters by μ, and concentration parameters by κ.

We used the original data from the study but removed two recordings from the moving dots condition because they majorly contained samples labeled as "other" or blinks. Moreover, the recordings could not be matched to the results obtained by Andersson et al. (2017).[6]

The second data set was published in Ehinger, Groß, Ibs, and Peter (2019) and has to our knowledge not yet been used for validation. Here, we only took tasks four and five out of 10 tasks because these are qualitatively different from the first data set. In task four, subjects were instructed to fixate a central target for 20 s. Task 5 was set up similarly, but subjects had to blink when they heard one out of seven beeps (with a beep duration of 100 ms and 1.5 s intervals in between). Eye movements were recorded with 500 Hz for 10 participants and 250 Hz for 5 participants due to a technical mistake (Ehinger et al., 2019). We used only data obtained by the EyeLink (SR Research Ltd., Ontario, Canada) eye-tracker and excluded recording using PupilLabs glasses, as wearable eye-tracker violates our methods' definition of frame of reference (R. S. Hessels et al., 2018).

[6] Two recordings from the moving dots conditions were substantially longer than the other recordings in the condition and contained more samples than were classified by the algorithms in the study by Andersson et al. (2017). Since no sample indices were available in the data set, we could not match samples from the two recordings to the labels assigned by the algorithms and therefore decided to remove them from the analysis. We do not expect the conclusions of our analyses to depend on these two data sets.

1.4.2 Data Analysis

Successful validation of gazeHMM was determined by two approaches: First, we applied gazeHMM with generative models containing 1-5 states to both data sets. The fits of the generative models were compared using Schwarz weights (Wagenmakers & Farrell, 2004), a conversion of the BIC (Schwarz, 1978) into model weights. They can be interpreted as the probability of a model having generated the data compared to the competing models. For the static condition in the Andersson et al. (2017) data set, we expected the generative model with three states (fixation, saccade, and PSO), and for the dynamic conditions the model with four states (incl. smooth pursuit) to display the highest Schwarz weight. Regarding the Ehinger et al. (2019) data set, we assumed that the one-state model (only fixation) would show the highest weights for both tasks.

The algorithm was applied separately to every subject, condition/task. For the Andersson et al. (2017) data set, all generative models were successfully fitted, whereas, for the Ehinger et al. (2019) data set, it was only 780 out of 900 models (87%, 60 models per task). The erroneous model fits in the Ehinger et al. (2019) data occurred when applying HMMs with three states or more. We attribute them to low variance in the data (i.e., it is difficult to fit data where subjects only fixate the same location with an HMM that assumes three or more states/events).

Second, we compared gazeHMM to other algorithms and human coders. We applied our algorithm with a three-state generative model to the static condition in the Andersson et al. (2017) data set, and with a four-state model to the dynamic conditions. For comparison criteria, we followed Andersson et al. (2017): We calculated the RMSD of event durations and counts between all algorithms and the average of the two human coders. Our results differ slightly from the original study because we excluded two recordings (leading to fewer events) and calculated the event durations as $Dur(e) = max(\mathbf{t}_e) - max(\mathbf{t}_{e-1})$, where $\mathbf{t_e}$ is the vector of sample time stamps for the event e. Cohen's kappa was calculated for each event as the binary agreement between the algorithms and the average of the human coders. Lastly, the overall disagreement indicated which samples were classified differently by the algorithms compared to the average of the human coders across all events. The human coders were compared directly to each other.

1.4.3 Results

Model Comparison

Examining the Schwarz weights displayed in Figure 1.15, we observed that the five-state generative model showed the highest weights in all three conditions. Only in the moving dots condition, two subjects displayed the highest weights for the four-, and one subject for the three-state model. In sum, we concluded that the five-state generative model has most likely generated the Andersson et al. (2017) data, opposing our expectations. Because the Ehinger et al. (2019) data set showed a similar pattern, we included the results for this data in the supplementary material.

A recent model recovery study showed that the BIC tended to prefer overly complex HMMs when they were misspecified (e.g., the conditional independence assumption was violated; Pohle, Langrock, van Beest, & Schmidt, 2017). Instead, the integrated completed likelihood (ICL) criterion (Biernacki, Celeux, & Govaert, 2000) performed better at choosing the correct data-generating model. Therefore, we post hoc computed the weighted ICL criterion (analogous to Schwarz weights) for the models fitted to the Andersson et al. (2017) data set. Using the ICL as the model selection criterion yielded very similar results to the BIC (see supplementary material). The preference for the five-state generative model was even more consistent across conditions and subjects.

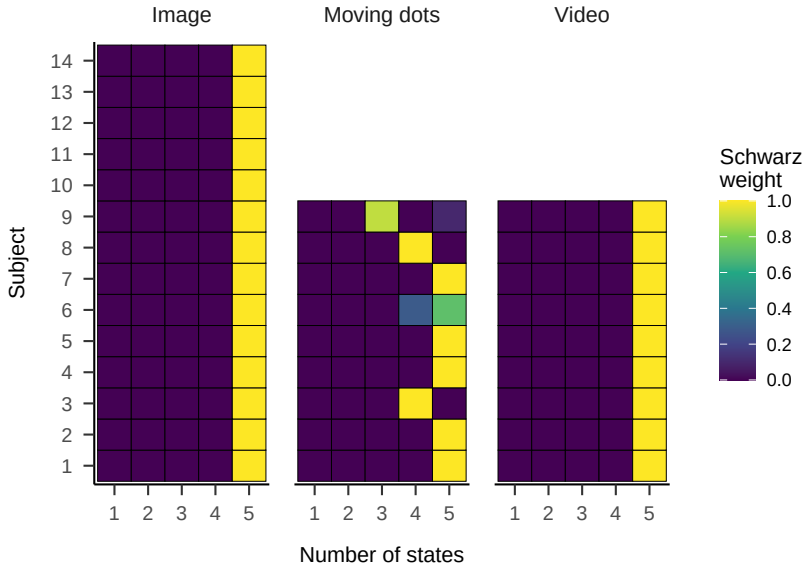


Figure 1.15: Schwarz Weights Displayed for Each Subject and HMMs With Different Numbers of States. Top facet labels indicate the condition in the Andersson et al. (2017) data set. Higher weights indicate a better model fit.

Comparison to Other Algorithms

As displayed in Table 1.5, gazeHMM showed a relatively low RMSD for fixations in the static condition compared to the other algorithms that were applied to the Andersson et al. (2017) data set. The lower RMSD for fixations indicated more similar classification to the human coders in terms of their mean and SD duration as well as the number of classified fixations. Oppositely, for fixations in the dynamic conditions, the RMSD of gazeHMM was one of the highest among the compared algorithms, suggesting substantial differences to the human coders. It can be seen that gazeHMM classified a much larger number of fixations with very short durations. For saccades, gazeHMM had a relatively high RMSD for the static condition but the lowest RMSD for the moving dots condition, and a moderate value for the video condition (see Table 1.6). The deviation was mostly because gazeHMM classified a higher number of saccades than the human coders. Only two other algorithms classified PSOs (NH and LNS; Larsson et al., 2013; Nyström & Holmqvist, 2010). Here, gazeHMM showed a consistently higher RMSD than LNS and lower RMSD than NH

(see Table 1.7). Our algorithm classified shorter and more PSOs than the human coders. No other algorithm parsed smooth pursuits, but the RMSD for gazeHMM was higher than among human coders (see Table 1.8). Again, it classified a much larger number of smooth pursuits with short durations.

Table 1.9 contains the sample-to-sample agreement between the algorithms and human coders measured by Cohen's kappa. For fixations, gazeHMM showed one of the highest agreements for static and *the* highest agreements for dynamic data. The absolute agreement was substantial for the static and slight to fair for the dynamic conditions (Landis & Koch, 1977). For saccades, gazeHMM had the lowest agreement for the static condition and moderate agreement for the dynamic conditions. In absolute terms, the agreement was fair to moderate. Concerning PSOs, gazeHMM showed higher agreement than NH in the image and video conditions but consistently lower agreement compared to LNS. The absolute agreement was slight (image) to moderate (video). Lastly, the agreement for smooth pursuit was lower compared to the human coders and fair in absolute values.

Concerning overall disagreement, Figure 1.16 shows that gazeHMM had less disagreement to the human coders across all events for the dynamic conditions. For the static condition, we interpreted the difference to most other algorithms as slight ($\text{Med}(\Delta) = 2.65\%$), but for the dynamic conditions, as substantial (video: $\text{Med}(\Delta) = 17.19\%$) and large (dots: $\text{Med}(\Delta) = 50.04\%$).

To explore which events gazeHMM classified differently than the average human coder, we looked at the confusion matrix between the two (see Table 1.10). It can be seen that gazeHMM classified many fixation samples as smooth pursuit samples and vice versa. Moreover, it confused many PSOs with saccade samples. The heuristic to detect blinks seemed to work successfully since gazeHMM classified most blink samples in agreement with human coding and only a minor part was mistaken for saccades. Inspecting an example of gaze data classified by gazeHMM compared to human coding leads to a similar notion: Figure 1.17 illustrates that gazeHMM is rapidly switching between classifying fixations and smooth pursuits, whereas the human coder identified one large smooth pursuit event. In the example, gazeHMM also disagrees with the human coder regarding the start of the PSO.

	Image				Moving dots				Video			
Algorithm	Mean	SD	Events	RMSD	Mean	SD	Events	RMSD	Mean	SD	Events	RMSD
coderMN	275	285	403	0.02	186	93	11	0.02	338	303	82	0.16
coderRA	271	287	391	0.02	174	94	12	0.02	255	185	81	0.16
gazeHMM-3	165	256	578	0.71	-	-	-	-	-	-	-	-
gazeHMM-4	-	-	-	-	16	17	381	1.12	22	25	1,243	1.18
CDT	465	643	276	1.44	60	115	177	0.44	244	354	226	0.18
IDTk	488	579	263	1.39	968	1,171	17	1.87	984	1,596	62	1.82
IKF	190	258	518	0.68	270	203	51	0.2	286	315	173	0.18
IMST	360	471	351	0.9	912	1,159	18	1.79	695	1,059	87	1.16
IHMM	148	236	717	0.72	316	298	47	0.34	259	348	207	0.18
IVT	127	223	843	0.73	292	298	51	0.31	225	334	240	0.18
NH	282	318	297	0.58	392	336	31	0.5	462	381	89	0.48
BIT	230	162	439	0.79	194	108	66	0.13	264	225	183	0.22

Table 1.5: Fixation Duration Descriptives and RMSD Between Algorithms and Human Coders. Durations are displayed in milliseconds. gazeHMM-3 classified three and gazeHMM-4 classified four events. RMSD = root mean square deviation. Table design adapted from Andersson et al. (2017).

	Image				Moving dots				Video			
Algorithm	Mean	SD	Events	RMSD	Mean	SD	Events	RMSD	Mean	SD	Events	RMSD
coderMN	32	16	377	0.12	23	11	38	0.06	27	12	117	0.1
coderRA	34	14	374	0.12	22	12	38	0.06	27	11	127	0.1
gazeHMM-3	39	29	657	0.84	-	-	-	-	-	-	-	-
gazeHMM-4	-	-	-	-	27	12	46	0.22	30	20	153	0.5
EK	27	24	787	0.98	18	13	59	0.49	22	18	252	1.08
IDTk	28	19	258	0.49	34	11	6	0.58	26	19	53	0.63
IKF	70	40	356	1.52	62	29	21	1.45	62	25	107	1.22
IMST	19	12	336	0.77	13	5	13	1.1	20	10	76	0.77
IHMM	54	28	370	0.75	41	19	19	0.63	46	18	109	0.46
IVT	46	25	375	0.48	35	14	20	0.33	40	17	112	0.25
NH	56	21	344	0.59	44	14	33	0.39	47	17	104	0.46
LNS	33	13	390	0.36	27	11	42	0.28	30	10	122	0.4

Table 1.6: Saccade Duration Descriptives and RMSD Between Algorithms and Human Coders. Durations are displayed in milliseconds. gazeHMM-3 classified three and gazeHMM-4 classified four events. RMSD = root mean square deviation. Table design adapted from Andersson et al. (2017).

	Image				Moving dots				Video			
Algorithm	Mean	SD	Events	RMSD	Mean	SD	Events	RMSD	Mean	SD	Events	RMSD
coderMN	25	14	313	0.68	14	5	24	0.63	23	13	97	0.43
coderRA	25	12	310	0.68	14	8	19	0.63	21	12	89	0.43
gazeHMM-3	14	15	518	1.48	-	-	-	-	-	-	-	-
gazeHMM-4	-	-	-	-	6	8	21	0.95	18	14	101	1.04
NH	31	15	237	1.54	23	13	11	1.25	31	20	78	1.19
LNS	30	15	319	1.19	20	8	21	0.54	30	19	87	0.82

Table 1.7: PSO Duration Descriptives and RMSD Between Algorithms and Human Coders. Durations are displayed in milliseconds. gazeHMM-3 classified three and gazeHMM-4 classified four events. RMSD = root mean square deviation. Table design adapted from Andersson et al. (2017).

	Image				Moving dots				Video			
Algorithm	Mean	SD	Events	RMSD	Mean	SD	Events	RMSD	Mean	SD	Events	RMSD
coderMN	363	187	3	0.23	370	238	36	0.4	559	391	51	0.14
coderRA	299	180	17	0.23	345	338	39	0.4	516	376	70	0.14
gazeHMM-4	-	-	-	-	23	22	400	1.97	21	23	1,281	1.95

Table 1.8: Smooth Pursuit Duration Descriptives and RMSD Between gazeHMM and Human Coders. Durations are displayed in milliseconds. gazeHMM-3 classified three and gazeHMM-4 classified four events. RMSD = root mean square deviation. Table design adapted from Andersson et al. (2017).

	Fixations			Saccades			PSOs			Smooth pursuits		
Algorithm	Image	Dots	Video	Image	Dots	Video	Image	Dots	Video	Image	Dots	Video
coderMN	0.84	0.84	0.65	0.91	0.79	0.87	0.76	0.57	0.65	0.34	0.81	0.66
coderRA	0.84	0.84	0.65	0.91	0.79	0.87	0.76	0.57	0.65	0.34	0.81	0.66
gazeHMM-3	0.67	-	-	0.36	-	-	0.19	-	-	0	-	-
gazeHMM-4	-	0.16	0.17	-	0.62	0.51	-	0.24	0.46	-	0.28	0.2
CDT	0.38	0.07	0.11	0	0	0	0	0	0	0	0	0
EK	0	0	0	0.64	0.73	0.67	0	0	0	0	0	0
IDTk	0.36	0	0.03	0.45	0.25	0.38	0	0	0	0	0	0
IKF	0.63	0.04	0.14	0.58	0.43	0.59	0	0	0	0	0	0
IMST	0.38	0	0.03	0.54	0.31	0.52	0	0	0	0	0	0
IHMM	0.67	0.03	0.13	0.69	0.58	0.71	0	0	0	0	0	0
IVT	0.67	0.03	0.13	0.75	0.59	0.76	0	0	0	0	0	0
NH	0.52	-0.23	0.01	0.67	0.58	0.68	0.24	0.14	0.25	0	0	0
BIT	0.67	0.02	0.14	0	0	0	0	0	0	0	0	0
LNS	0	0	0	0.81	0.75	0.81	0.64	0.56	0.63	0	0	0

Table 1.9: Cohen's Kappa Between Human Coders and Algorithms for Different Conditions and Events. gazeHMM-3 classified three and gazeHMM-4 classified four events. Table design adapted from Andersson et al. (2017).

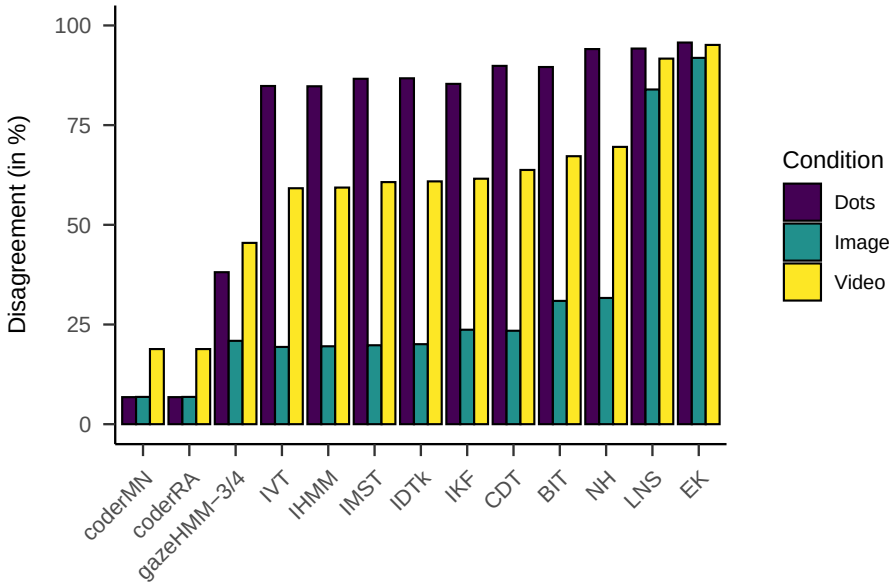


Figure 1.16: Disagreement Between Algorithms and Human Coders for Different Conditions. gazeHMM-3/4 classified three events for image data and four events for moving dots/video data. Algorithms are displayed in order according to mean disagreement over conditions (least/left to highest/right).

Event	Fixation	Saccade	PSO	Pursuit	Blink	Other
Image						
Fixations	0.88	0.08	0.08	0.72	0.07	0.03
Saccades	0.07	0.59	0.63	0.22	0.14	0.20
PSOs	0.04	0.08	0.23	0.03	0.05	0.16
Pursuits	0.00	0.00	0.00	0.00	0.00	0.00
Blinks	0.01	0.26	0.06	0.03	0.74	0.61
Moving dots						
Fixations	0.61	0.01	0.02	0.35	0.00	0.30
Saccades	0.01	0.78	0.60	0.03	0.00	0.48
PSOs	0.01	0.04	0.18	0.00	0.00	0.00
Pursuits	0.28	0.05	0.17	0.62	0.00	0.20
Blinks	0.10	0.12	0.04	0.00	0.00	0.02
Video						
Fixations	0.54	0.04	0.02	0.44	0.00	0.00
Saccades	0.02	0.57	0.35	0.02	0.14	0.82
PSOs	0.01	0.13	0.47	0.01	0.03	0.05
Pursuits	0.42	0.03	0.12	0.54	0.01	0.00
Blinks	0.00	0.23	0.04	0.00	0.82	0.14

Table 1.10: Confusion Matrix Between gazeHMM (Rows) and Human Coders (Columns) for Different Conditions. gazeHMM classified four events and blinks. Values indicate proportions of samples where gazeHMM and human coders agree divided by the total number of samples classified by the human coders for each event (i.e., columns sum to one).

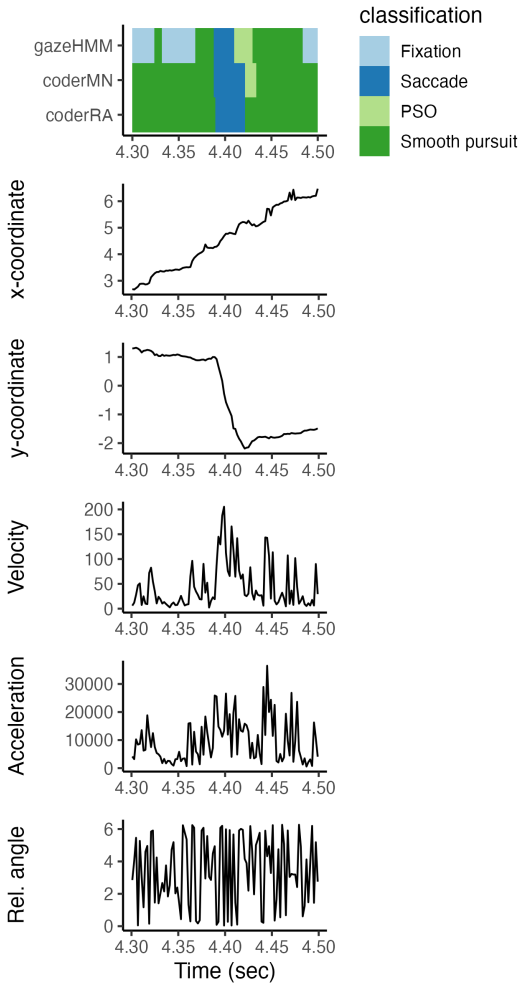


Figure 1.17: Classification of Example Data by Andersson et al. (2017). Data displayed as x-, and y-coordinates (in deg, upper two panels), velocity (in deg/s, middle panel), acceleration (in deg/s^2, fourth panel), and sample-to-sample (relative) angle (in radians, bottom panel). The top-most panel displays event classification by gazeHMM, coderMN, and coderRA, highlighted by color.

1.5 Conclusion & Discussion

In this report, we presented gazeHMM, a novel algorithm for classifying gaze data into eye movement events. The algorithm models velocity, acceleration, and sample-to-sample angle signals with gamma distributions and a mixture of von Mises and a uniform distribution. An HMM serves as the generative model of the algorithm and classifies the gaze samples into fixations, saccades, and optionally PSOs, and/or smooth pursuits. We showed in a simulation study that the generative model of gazeHMM recovered parameters and hidden state sequences well. However, adding a fourth event (i.e., smooth pursuit) to the model and introducing even small amounts of noise to the generated data led to decreased parameter recovery. Importantly, however, it did not lead to decreased hidden state recovery. Thus, the classification of the generative model should not be negatively affected by noise. Furthermore, we applied gazeHMM with different numbers of states to benchmark data by Andersson et al. (2017) and compared the model fit. The model comparison revealed that a five-state HMM had consistently most likely generated the data. This result opposed our expectation that a three-state model would be preferred for static and a four-state model for dynamic data. When comparing gazeHMM against other algorithms, gazeHMM showed mostly good agreement to human coding. On one hand, it outperformed the other algorithms in the overall disagreement with human coding for dynamic data. On the other hand, gazeHMM confused a lot of fixations with smooth pursuits, which led to rapid switching between the two events. It also tended to mistake PSO samples as belonging to saccades.

Considering the results of the simulation study, it seems reasonable that adding the smooth pursuit state to the HMM decreased parameter and state recovery: It is the event that is overlapping most closely with another event (fixations) in terms of velocity, acceleration, and sample-to-sample angle. The overlap can cause the HMM to confuse parameters and hidden states. The decrease in parameter recovery (especially for scale parameters) due to noise shows that the overlap is enhanced by more dispersion in the data. The scale parameters might be particularly vulnerable to extreme data points. Despite these drawbacks, the recovery of the generative model in gazeHMM seems very promising. The simulation study gives also an approximate reference for the maximum recovery of hidden states that can be achieved by the HMM (Co-

hen's kappa values of ~1 for two, ~0.95 for three, and ~0.8 for four events).

The model comparison on the benchmark data suggested that the generative model in gazeHMM is not yet optimally specified for eye movement data. There are several explanations for this result:

The model subdivided some events into multiple events, or found additional patterns in the data that do not fit the other four events the model was built for. Eye movement events can be divided into subevents. For example, fixations consist of drift and tremor movements (Duchowski, 2017) and PSOs encompass dynamic, static, and glissadic over- and undershoots (Larsson et al., 2013). A study on a recently developed HMM algorithm supports this explanation: Houpt et al. (2018) applied the unsupervised BP-AR-HMM algorithm to the Andersson et al. (2017) data set and classified more distinct states than the human coders. Some of the states classified by BP-AR-HMM matched the same event coded by humans. Since the subevents are usually not interesting for users of classification algorithms, the ability of HMMs to classify might limit their ability to generate eye movements.

Model selection criteria are generally not appropriate for comparing HMMs with different numbers of states. This argument has been discussed in the field of ecology (see Li & Bolker, 2017), where studies found that selection criteria preferred models with more states than expected (similar to the result of this study; e.g., Langrock, Kneib, Sohn, & Deruiter, 2015). Li and Bolker (2017) explain this bias with the simplicity of the submodels in HMMs: Initial state, transition, and response models for each state are usually relatively simple. When they do not describe the processes in the respective states accurately, the selection criteria compensate for that by preferring a model with more states. Thus, there are not more latent states present in the data, but the submodels of the HMM are misspecified or too simple, potentially leading to spurious, extra, states being identified in the model selection process, see discussion and potential solutions in Kuijpers, Visser, and Molenaar (2021). Correcting for model misspecifications led to a better model recovery in studies on animal movements (Langrock et al., 2015; Li & Bolker, 2017). However, Pohle et al. (2017) showed in simulations that the ICL identified the correct model despite several misspecifications. It has to be noted that the study by Pohle et al. (2017) only used data generating models with two states, so it needs to be veri-

fied whether this approach will work in the larger models that are being studied here.

The submodels of gazeHMM were misspecified. Pohle et al. (2017) identified two scenarios in which model recovery using the ICL did not give optimal results: Outliers in the data and inadequate distributions in the response models. Both situations could apply to gazeHMM and eye movement data: Outliers occur frequently in eye-tracking data due to measurement error. Choosing adequate response distributions in HMMs is usually difficult and can depend on the individual and task from which the data are obtained (Langrock et al., 2015). Moreover, gazeHMM only estimated intercepts for all parameters and thus, no time-varying covariates were included (cf. Li & Bolker, 2017). This aspect could indeed oversimplify the complex nature of eye movement data.

Comparing gazeHMM to other algorithms on benchmark data showed that gazeHMM showed good agreement with human coders. However, the evaluation criteria (RMSD of event durations, sample-to-sample agreement, and overall disagreement) yielded different results. The fact that gazeHMM outperformed all other algorithms regarding the overall disagreement can be because it is the only algorithm classifying all five events the human coders classified; algorithms that do not classify certain type of even will, by definition, disagree with human coders on samples that they classified as such. As the number of samples in different events depending on the stimuli (e.g., a lot of smooth pursuit in moving dots condition but virtually none in static images), different methods might be penalized differently depending on the condition and type of event they do not classify. Nevertheless, Cohen's kappa values of 0.67 (fixations - image) or 0.62 (saccades - moving dots) indicate substantial agreement to human coders, especially in light of the maximum references from the simulation study. At this point, it is important to mention that human coding should not be considered a gold standard in event classification: Hooge et al. (2018) observed substantial differences between coders and within coders over time. Despite these differences, they recommend comparisons to human coding to demonstrate the performance of new algorithms and to find errors in their design.

1.5.1 Advantages of gazeHMM

Given the four proposed goals that gazeHMM should fulfill, we can draw the following conclusions: Even though gazeHMM does require some parameter settings (in the pre- and postprocessing), it estimates many parameters adaptively from the data; as a result, compared to many other algorithms, it reduces the influence of human judgement and researcher decisions on the classification result. The parameters are merely included to compensate for the drawbacks of the generative model and their default values should be appropriate for most applications. A major advantage of gazeHMM is that it does not require human-labeled data as input. Instead, it estimates all parameters and hidden states from the data. Since human coding is quite laborious, difficult to reproduce, and by times inconsistent (as noted earlier, Hooge et al., 2018), this property makes gazeHMM a good alternative to other recently developed algorithms that require human coded input (Bellet et al., 2019; Pekkanen & Lappi, 2017; Zemblys et al., 2018). This could also explain why the agreement to human coding is lower for gazeHMM than for algorithms that learn from human-labeled data.

Another advantage of gazeHMM is its ability to classify four eye movement events, namely fixations, saccades, PSOs, and smooth pursuit. Whereas most algorithms only parse fixations and saccades (Andersson et al., 2017), few classify PSOs (e.g., Zemblys et al., 2018), and even less categorize smooth pursuits (e.g., Pekkanen & Lappi, 2017). However, including smooth pursuits in gazeHMM led to some undesirable classifications on benchmark data, resulting in rapid switching between fixation and smooth pursuit events. Therefore, we recommend using gazeHMM with four events only for exploratory purposes. Without smooth pursuits, we consider gazeHMM's classification as appropriate for application. Lastly, its implementation in R using depmixS4 (Visser & Speekenbrink, 2010) should make gazeHMM a tool that is easy to use and customize for individual needs.

To conclude, our methods shows promising results in terms of ability to classify various eye movement events, does not require previously labeled data, and reduces the number of arbitrary settings determined by the researcher. As such, in case the ultimate goal is event classification, the method is a good candidate for initial rough estimate of the event classification, which can be further

inspected and refined, if necessary. Compared to other approaches, the method is also easily extensible and modifiable, allows for model comparison, and as such offers applications where broadening our understanding of eye movement is of primary interest instead of the event classification itself.

1.5.2 Future Directions

Despite its advantages, there are several aspects in which gazeHMM can be improved: First, a multivariate distribution could be used to account for the correlation between velocity and acceleration signals (for examples, see Balakrishnan & Lai, 2009). Potential problems of this approach might be choosing the right distribution and convergence issues (due to a large number of parameters). Another option to model the correlation could be to include one of the response variables as a covariate of the other.

Second, instead of the gamma being the generic (and potentially inappropriate) response distribution, a non-parametric approach could be used: Langrock et al. (2015) use a linear combination of standardized B-splines to approximate response densities, which led to HMMs with fewer states being preferred. This approach could potentially combat the problem of unexpectedly high-state HMMs being preferred for eye movement data but would also undermine the advantages of using a parametric model.

Third, one solution to diverging results when comparing gazeHMM with different events could be model averaging: Instead of using the maximum posterior state probability of each sample from the preferred model, the probabilities could be weighted according to a model selection criterion (e.g., Schwarz weight) and averaged. Then, the maximum averaged probability could be used to classify the samples into events. This approach could lead to a more robust classification because it reduces the overconfidence of each competing model and easily adapts to new data (analogous to Bayesian model averaging; Hinne, Gronau, van den Bergh, & Wagenmakers, 2020). However, the model comparison for gazeHMM often showed extreme weights for a five-state model, which would lead to a very limited influence of the other models in the averaged probabilities.

Fourth, including covariates of the transition probabilities and response parameters could improve the fit of gazeHMM on eye movement data. As

pointed out earlier, just estimating intercepts of parameters could be too simple to model the complexity of eye movements. A candidate for such a covariate might be a periodic function of time (Li & Bolker, 2017) which could, for instance, capture the specific characteristics of saccades, e.g., the tendency of increasing velocity at the start of the saccade and decreasing velocity at the end of the saccade. Whether covariates are improving the fit of submodels to eye movement data could in turn be assessed by inspecting pseudo-residuals and autocorrelation functions (Zucchini et al., 2016).

Fifth, to avoid rapid switching between fixations and smooth pursuits as well as unreasonably short saccades, gazeHMM could explicitly model the duration of events. This can be achieved by setting the diagonal transition probabilities to zero and assign a distribution of state durations to each state (Bishop, 2006). Consequently, the duration distributions of fixations and smooth pursuits could differ from saccades and PSOs. This extension of the HMM is also called the hidden semi-Markov model and has been successfully used by Mihali et al. (2017) to classify microsaccades. Drawbacks of this extension are higher computational costs and difficulties with including covariates (Zucchini et al., 2016).

Lastly, allowing constrained parameters in the HMM could replace some of the postprocessing steps in gazeHMM. This could potentially be achieved by using different response distributions or parameter optimization methods. Moreover, switching from the maximum likelihood to the Markov chain Monte Carlo (Bayesian) framework could help to avoid convergence problems with constrained parameters, but would also open new research questions about suitable priors for HMM parameters in the eye movement domain, efficient sampling plans, accounting for label switching, and computational efficiency, naming only a few.

1.5.3 Conclusion

In the previous sections, we developed and tested a generative, HMM-based algorithm called gazeHMM. Both a simulation and validation study showed that gazeHMM is a suitable algorithm for simulating, understanding and classifying eye movement events. For smooth pursuits, the classification is not optimal and thus not yet recommended. On one side, the algorithm has some advan-

tages over concurrent event classification algorithms, not relying on human-labeled training data being the most important one. On the other side, it is not able to identify expected events in model comparisons. The current model constitutes a proof-of-principle that a generative, maximum-likelihood based approach can provide interpretable and reliable results that are at least as good as other approaches under some circumstances. The largest advantage of this approach is however that it provides the possibility to rigorously test progress in developing extensions and improvements.

Open Practices Statement

The simulation study was preregistered; the preregistration is available at `osf.io/vdjgp`. The gazeHMM implementation is available as an `R` package at `github.com/maltelueken/gazeHMM`. The analysis code is available at `github.com/maltelueken/gazeHMM_validation`.

Well, Clive, it's all about the two M's—movement and positioning.

–Ron Atkinson

Chapter 2

WALD-EM: Wald Accumulation of Locations and Durations of Eye Movements

This chapter is published as Kucharský, Š., van Renswoude, D., Raijmakers, M., and Visser, I. (2021). WALD-EM: Wald accumulation for locations and durations of eye movements. *Psychological Review*, *128*(4), 667-689. doi: 10.1037/rev0000292

Abstract

Describing, analyzing and explaining patterns in eye movement behavior is crucial for understanding visual perception. Further, eye movements are increasingly used in informing cognitive process models. In this article, we start by reviewing basic characteristics and desiderata for models of eye movements. Specifically, we argue that there is a need for models combining spatial and temporal aspects of eye-tracking data (i.e., fixation durations and fixation locations), that formal models derived from concrete theoretical assumptions are needed to inform our empirical research, and custom statistical models are useful for detecting specific empirical phenomena that are to be explained by said theory.

In this article, we develop a conceptual model of eye movements, or specifically, fixation durations and fixation locations, and from it derive a formal statistical model — meeting our goal of crafting a model useful in both the theoretical and empirical research cycle. We demonstrate the use of the model on an example of infant natural scene viewing, to show that the model is able to explain different features of the eye movement data, and to showcase how to identify that the model needs to be adapted if it does not agree with the data. We conclude with discussion of potential future avenues for formal eye movement models.

2.1 Introduction

As only a relatively small region on the retina provides the highest detail of the visual input, the human visual system heavily relies on the ability to control the gaze and movement of the eye over a stimulus (Duchowski, 2017). Much of the current research intends to determine the mechanisms and factors[1] that guide visual attention through fixations and saccades, i.e., periods of fixing the visual input relatively steady on the retina and periods of abrupt movements, respectively, as understanding these mechanisms provides insights into visual and attentional control and their impact on perception. Additionally, studying eye movements is not only essential for understanding perception and attentional control but can also inform variety of other topics, such as the study of higher cognitive processes like decision rules in economic games (Polonio, Di Guida, & Coricelli, 2015), strategic differences in analogical reasoning tasks (Hayes, Petrov, & Sederberg, 2015; Kucharský et

[1] Throughout the article, we use the term "factor" as "a circumstance, fact, or influence that contributes to a result" without having a specific functional form of the relationship in mind.

al., 2020), or individual assessment (Chen et al., 2014), to name a few.

Previous research distinguishes the mechanisms and factors that guide visual attention into three groups (Itti & Borji, 2014; Schütt et al., 2017; Tatler & Vincent, 2008). These groups can be roughly described as bottom-up, top-down, and systematic tendencies. The bottom-up factors include features of the visual environment, such as distribution of colors and contrast across the visual field, etc. Many of the so called saliency models aim to determine and detect these features (Itti & Koch, 2001; Tatler, Hayhoe, Land, & Ballard, 2011; J. Xu, Jiang, Wang, Kankanhalli, & Zhao, 2014). The top-down factors and mechanisms include characteristics and states of the observer, such as their motivation, purpose, task, (background) knowledge or individual differences (De Haas, Iakovidis, Schwarzkopf, & Gegenfurtner, 2019). The third group includes factors that are neither purely bottom-up (i.e., not necessarily tied to features in the environment) nor top-down (i.e., not necessarily unique to states or characteristics of observers), but rather experimentally observed phenomena (Tatler & Vincent, 2008). Systematic tendencies are believed to be relatively stable across stimuli, participants and tasks, such as fixation biases (e.g., central bias; Tatler, 2007; Tseng, Carmi, Cameron, Munoz, & Itti, 2009; van Renswoude, van den Berg, Raijmakers, & Visser, 2019) or saccadic biases (e.g., horizontal and leftward bias; Foulsham, Frost, & Sage, 2018; Foulsham, Gray, Nasiopoulos, & Kingstone, 2013; Le Meur & Liu, 2015; van Renswoude, Johnson, Raijmakers, & Visser, 2016).

Apart from experimental work establishing individual factors that influence gaze behavior, important aspect of understanding the mechanism behind the observed behavior is proposing theoretical and statistical models that are able to describe, explain, or predict empirical data and observed phenomena. There are many models with varying levels of abstraction, theoretical substance, the phenomena they aim to explain, and the type and level of data they are able to explain (Le Meur & Liu, 2015; Malem-Shinitski et al., 2020; Nuthmann, 2017; Reichle & Sheridan, 2015; Schwetlick, Rothkegel, Trukenbrod, & Engbert, 2020; Schütt et al., 2017; Tatler, Brockmole, & Carpenter, 2017; Trukenbrod & Engbert, 2014; Zelinsky, Adeli, Peng, & Samaras, 2013). In this article, we develop a new conceptual model of eye movements, and flesh it out in the form of a statistical model.

2.1.1 Model requirements

Two prominent questions regarding eye movement behavior that require explanation are *when* and *where* (Findlay & Walker, 1999; Tatler et al., 2017), i.e., what is the mechanism behind the *timing of saccades and fixation durations*, and what is the mechanism behind selecting *fixation locations*. Predominantly, these questions are asked separately by building models explaining either fixation durations or fixation locations (Nuthmann, Smith, Engbert, & Henderson, 2010; Schütt et al., 2017; Tatler et al., 2017). However, better understanding of visual behavior is perhaps only possible when considering *where* and *when* people look simultaneously (Tatler et al., 2017). It is of interest to consider spatial and temporal phenomena in one model, as these are likely not independent of each other (e.g., J. M. Henderson, Nuthmann, & Luke, 2013; Nuthmann, 2017). In this article, we propose a new account of how to model eye movements both spatially and temporally in a joint framework.

One of the critical features of theory driven models of any behavior is the ability to generate data, given its set of assumptions. This enables to assess whether a model is successful in generating phenomena that are putatively explained by said theory (Borsboom, van der Maas, Dalege, Kievit, & Haig, 2020; Robinaugh, Haslbeck, Ryan, Fried, & Waldorp, 2020), and also makes it possible to make counterfactual investigations. That is, we might use it to answer the question "according to the model, what would have happened if something would have occurred, but it did not?", which is useful for hypothesis generation and essentially more precise testing of theories underlying the models (e.g., Nuthmann et al., 2010). This is generally a useful approach that enables to check the explanatory adequacy of the underlying theory, inform us about where to look for crucial piece of evidence, and as such serving a crucial part of the theoretical cycle (Borsboom et al., 2020). Building data generative models of eye movements has a long tradition in the eye-tracking literature. In fact, the traditional approach to evaluate eye movement models typically involves simulating eye movement data from a model and comparing the synthetic data to experimentally established phenomena (Schütt et al., 2017).

Additionally to being used as generative models, formal modelling approaches are widely used in the empirical cycle as well in form of statistical models, where they play a crucial role in detecting and establishing new phenomena

from the collected data (Wagenmakers, Dutilh, & Sarafoglou, 2018). Thus, as dynamic models of eye movements gain importance in theoretical and experimental research, parameter estimation and model comparison are also gaining importance. This requires being able to specify a model as a statistical model (i.e., a probability distribution of the data given a set of parameters) that can be used to estimate the parameters (either using maximum likelihood or Bayesian approaches), and use the statistical machinery for assessing the uncertainty in parameter estimates and to conduct model comparisons (e.g., Malem-Shinitski et al., 2020; Schütt et al., 2017).

Detecting new phenomena is of great interest for eye movement researchers. For example, in studying phenomena such as the central bias (i.e., relative preference to focus on the center of the screen compared to other areas), there is an ongoing debate whether it can be explained away as a manifestation of bottom-up effects (such as distributions of objects on the screen) or whether it is a real systematic bias somehow ingrained in our visual system, and how to disentangle these explanations (Tatler, 2007; Tseng et al., 2009; van Renswoude et al., 2019). Having a possibility to modify the model such that it includes or excludes the central bias, would enable us to pit these explanations against each other. Through model comparison and parameter estimation, we can then assess whether and quantify to what extent these different factors come into play. Thus, it is important that a model can be modified to include, exclude or modify the functional form of the effect of different factors or mechanisms influencing the eye movement behavior.

Furthermore, it is highly likely that eye movement characteristics will depend on individual differences, differences between different populations, or within-person differences due to development (De Haas et al., 2019). It is thus important to be able to model these differences in one coherent modeling framework by allowing to specify parameters in the model to e.g., differ between populations or as random terms in a hierarchical fashion.

Models that are possible to use both in the theoretical cycle (i.e., as formal manifestations of a theory to check that the theory explains phenomena that it set out to explain) and empirical cycle (i.e., to assess the evidence for new phenomena that need explanation) are generally difficult to develop and rare, so rare that these two purposes of scientific models are often discussed as com-

pletely separate entities (Smaldino, 2017). However, having a model that is both statistical and informed by the underlying theory often offers deeper insights into the underlying mechanisms (Borsboom et al., 2020; Rodgers, 2010), and provides additional opportunities to learn both about the model and the natural phenomena (McElreath, 2020, pp. 525–552). In cognitive psychology, such models are sometimes referred to as cognitive process models, as they describe the cognitive processes that underlie the data, and possess parameters that often have clear interpretations (Forstmann & Wagenmakers, 2015).

2.1.2 Outline

In this article, we propose a model that explains fixation locations and fixation durations simultaneously, and is 1) generative (i.e., can make predictions about the locations of fixations at a particular time), 2) statistical (i.e., has a proper likelihood function), 3) modifiable (i.e., can be expanded to include different factors, including random factors), and 4) can be interpreted as a cognitive process model.

The structure of this article is as follows. In the next section, we introduce the model in conceptual terms, i.e., describe the architecture of the model to highlight the core assumptions which yield the model interpretable as a cognitive process model, while abstracting from particular analytic choices. Then, we show how to derive a particular realisation of the model. This will involve laying out concretely what analytic choices we made to make the model tractable. We lay out several factors that can optionally be included in the model, and apply different versions of the model to real data to answer substantive questions, thereby illustrating the model flexibility and usefulness. In the following sections, we limit ourselves mostly to the domain of free scene viewing. We believe that the proposed model could be extended or adapted further to other contexts or paradigms, but that is not the focus of the current article.

We will refer to the new model as WALD-EM, standing for "Wald accumulation of locations and durations of eye movements". Reasons for this name will become apparent in the following description of the model.

2.2 Conceptual WALD-EM model

Our model describes eye movement data as x and y coordinates and durations of fixations, and aims to provide answers to the questions about *when* and *where* simultaneously. As such, it consists of two parts: One that corresponds to the question *when*, and one that corresponds to the question *where*. These two parts are then intertwined together to capture potential dependencies between these two.

2.2.1 Model for *when*

A typical human (adult) in typical situations makes on average one saccade in 200-400 msec. The distribution of fixation durations is characteristically positively skewed with typically positive relationship between the mean and a variance, much like typical distributions of response times in decision tasks (Palmer, Horowitz, Torralba, & Wolfe, 2011). Hence, it is reasonable to borrow from the response time modelling literature, i.e., evidence accumulation models, such as LATER (R. H. S. Carpenter & Williams, 1995), Linear Ballistic Accumulation (S. D. Brown & Heathcote, 2008), or Diffusion Decision (Ratcliff & McKoon, 2008).

In our model, we represent the fixation duration as the time it takes the observer to make a decision to make a saccade. The decision process represents information uptake from a current location up to a point where the currently fixated location does not bring additional information compared to potential information sources at other locations. We assume that information uptake is a continuous-time stochastic process that rises to a threshold with some drift and noise level. The time to make the decision to make a saccade is the first passage time of this process. The simplest model for such a time is the Wald distribution with three parameters: drift (ν), decision boundary (α), and standard deviation of the noise (σ), one of which needs to be fixed for identifiability purposes (Chhikara & Folks, 1988). Apart from that the Wald distribution is a reasonable candidate as it reflects the noisy evidence accumulation process (a process that has been deemed as a neurally plausible mechanism for decision processes, Anders, Alario, & van Maanen, 2016), it has previously also been shown to fit fixation durations well (Palmer et al., 2011). Figure 2.1 shows the

mechanism that gives rise to the Wald distribution.

Other models contain similar data generating processes for fixation durations. For example, the LATEST model (Tatler et al., 2017) assumes that the fixation duration is also the time to make a decision to make a saccade to a new location. Our model assumes stochastic random walk accumulation, whereas LATEST assumes a linear ballistic process. Further our model assumes only one decision process at a time, whereas LATEST assumes many accumulators running in parallel. CRISP (Nuthmann et al., 2010) and ICAT (Trukenbrod & Engbert, 2014) models also rely on a stochastic random walk underlying the fixation durations. In CRISP and ICAT, however, decisions to make a saccade can be cancelled by additional processes, whereas our proposal is simpler in that passing the threshold immediately triggers a saccade. Further, in ICAT and CRISP, the stochastic rise to threshold is thought of as an autonomous timer, suggesting an inherent (but stochastic) rhythmicallity to saccades, whereas our accumulator depends not only on internal characteristics of the observer, but their surroundings as well.

2.2.2 Model for *where*

After the observer concludes that there is an advantage to move to another location, it is time to make a saccade.

Each location of the stimulus provides some amount of attraction to the observer. We call a function that maps the stimulus coordinates to that attraction an *intensity* function and denote it as: $\lambda : \mathbb{R}^2 \to \mathbb{R}_+$, and will write it as $\lambda(x, y|s)$, where s stands for the current fixation. The total amount of intensity of the whole stimulus is the integral (sum) of all the points of the stimulus: $\Lambda = \int\int \lambda(x, y|s)\, dxdy$. In essence, we assume that when observers decide *where* to go next, they pick a random location from a distribution proportional to this function. The function may or may not depend on the current or previous fixations, depending on whether we assume a homogeneous (static over time) or heterogeneous (evolving over time) process, and can be adjusted depending on the researcher's questions and desires.

In general, we will represent the intensity function as a combination of different factors that influence the intensity of different locations. These factors may represent different features and can be combined in different ways (see

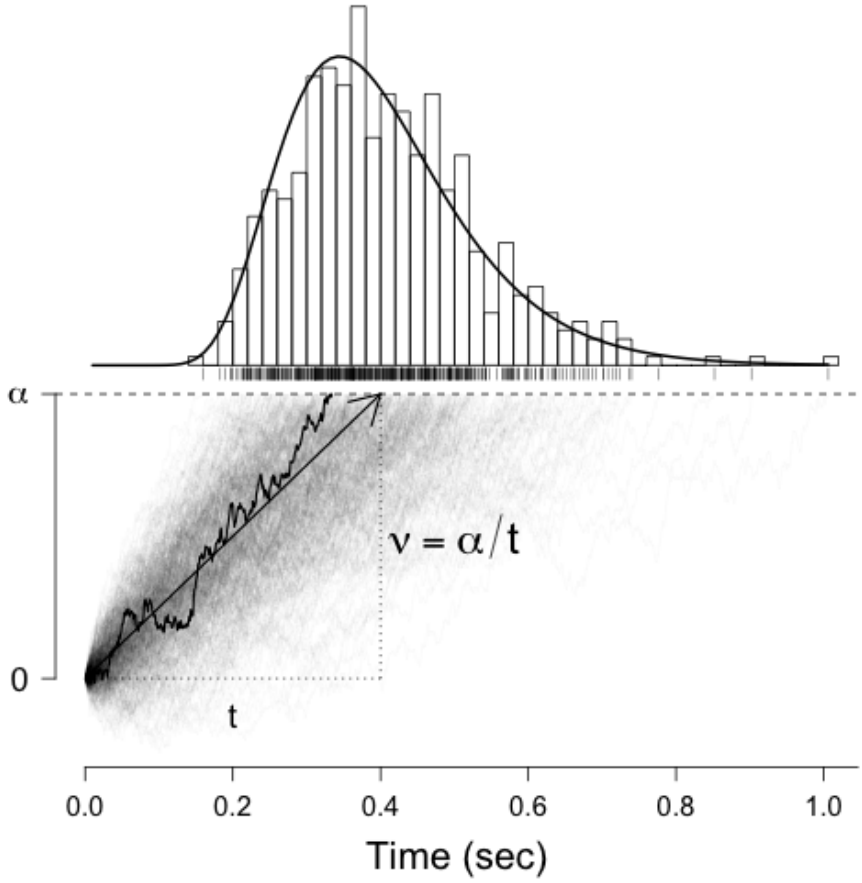


Figure 2.1: Illustration of the process that results in a Wald distribution. Evidence starts at 0 and accumulates as a Wiener process with a drift ν (displayed as arrow) until it reaches a threshold α. The process is inherently noisy as shown by 500 different traces generated with the same parameters (grey lines). The first passage time (the time it takes to trespass the threshold α for the first time) results in a Wald distribution (displayed on top).

Barthelmé, Trukenbrod, Engbert, & Wichmann, 2013). For example, we can build the intensity function such that it combines bottom up features of the stimulus (e.g., saliency) with systematic tendencies (e.g., central bias or horizontal bias), and so forth. Some of the factors can be thought of as representing information provided by the stimulus, assuming that locations that are rich in the information they provide will be attractive to fixate — and so will have a high intensity. However, people not always fixate on locations with a lot of information. Later, it will be important to make a distinction between two types of factors that combine in the intensity. The first group of factors will encompass those that in some sense represent, or encode, information provided by the stimulus, such as objects, shapes, colors, edges, faces, etc.[2] We will denote the combination of these factors as $\lambda_1(x, y|s)$ and the integral

$$\Lambda_1 = \int\int \lambda_1(x, y|s)\, dxdy \tag{2.1}$$

will represent the total amount of information provided by the stimulus. The second group comprises of factors that do not represent information of the stimulus but influence the attractiveness of the potential locations by another way, for example heightening the intensity near the center of the stimulus would represent a central bias.

2.2.3 Combining models for *when* and *where*

The crucial part of the WALD-EM is how it relates the model for *when* and the model for *where* to each other. Recall that we conceptualize fixation duration as a period of evidence accumulation from a stimulus, and that information that provides this evidence is a part of the intensity maps. However, not all information is accessible at any single fixation (which is why we make saccades in the first place). Indeed, human vision is limited by the fact that only at the fovea, the place of the retina where the light falls from roughly around the center of gaze, great detail is available. This provides a key insight that the fixation durations should be dependent on how much information there is available at the

[2] We use the term *information* for a lack of a better word, and do not use it in a strict sense associated with the work of Shannon (1948).

particular location the observer currently fixates. The physiological aspects of foveal, parafoveal and extrafoveal vision are out of the scope of this article, but similarly to other attempts for modeling of eye movements (Schwetlick et al., 2020; Schütt et al., 2017; Trukenbrod & Engbert, 2014), we represent the fact that vision is sharpest inside the fovea by implementing a so called "attentional window". This window suppresses intensity of locations relatively farther from the center of gaze, and effectively limits the total information that is accessible to the observer given the current fixation location.

In essence, we define an attentional window as a function $a : \mathbb{R}^2 \rightarrow \mathbb{R} \in [0, 1]$ and denote it as $a(x, y|s)$, where s stands for the x and y coordinates of the current fixation. The value of a corresponds to the proportion of the intensity of locations at (x, y) given the current fixation location s. To get a representation of the actual intensity of different locations, given a particular fixation location s, we can multiply the intensity function by this attentional window:

$$\omega(x, y|s) = a(x, y|s) \times \lambda(x, y|s), \tag{2.2}$$

and the total amount of accessible intensity during a particular fixation s is $\Omega = \int\int \omega(x, y|s)\, dxdy$; the total amount of information accessible to the observer at a particular location will be denoted as

$$\Omega_1 = \int\int \omega_1(x, y|s)\, dxdy = \int\int a(x, y|s) \times \lambda_1(x, y|s)\, dxdy. \tag{2.3}$$

Figure 2.2 illustrates this concept with examples in one dimension.

The concept of attentional window is important in our model as it provides a link between the model for *when* and model for *where* to enable dependencies between the two. Specifically, we make the model of *when* depend on the model of *where*, and the attention window specifies how does that happen. In the following, we make this link explicit. In the model for *when*, the time it takes the observer to make a decision (to make a saccade) can be modelled as a Wald distribution with parameters drift ν and decision boundary α. However, it is likely that fixation durations vary depending on the surroundings of the current fixation location (Einhäuser, Atzert, & Nuthmann, 2020; Nuthmann, 2017; Nuthmann et al., 2010).

To link the model for *when* and *where*, we also make a distinction between

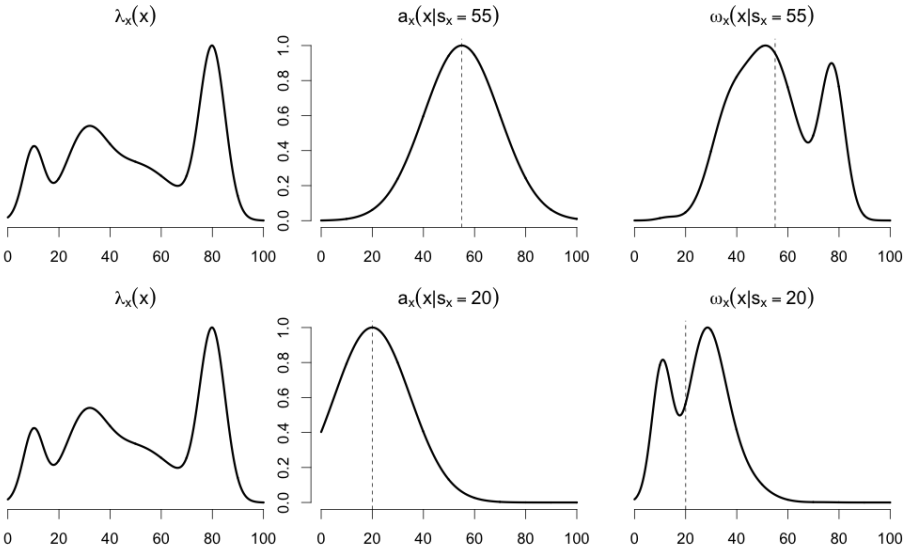


Figure 2.2: The left panel shows an example intensity function $\lambda(x)$ as a function of location along the x-coordinate. The middle panel shows the attention window given that the current fixation is at $s_x = 55$ (top) or $s_x = 20$ (bottom). The right panel shows the intensity accessible through the attention window.

factors that do and do not represent information provided by the stimulus, as we assume that only information has a potential to influence the fixation duration (e.g., fixating on a location particularly rich on detail will take longer on average than on a location with only a uniform background) and not other factors that do not provide information (e.g., central bias can attract people to make a saccade towards the center of the screen, but there is no immediately plausible mechanism for having longer fixation durations in the center of the screen compared to the edges just because it is in the center). Generally, the dependency of the fixation durations on fixation locations can be created in two ways, and the two approaches are discussed here.

In the first approach, we can assume that upon arriving to a location s, the observer harvests information from around that location with a drift rate ν, and once the information available from that location is depleted, the decision to *Go* is activated. In this framework, the total amount of information available through the attention window Ω_1 would replace the decision boundary α in

the Wald model.

In the second approach, we can adopt the idea from LATEST (Tatler et al., 2017) that the decision to make a saccade is based on continual comparison of two hypotheses (*Stay* vs *Go*), where the "evidence" is based on the information provided if one or another decision is adopted. The evidence supporting the decision to stay is the total amount of information accessible through the attention window (Ω_1), whereas the evidence supporting the decision to Go is the total amount of information provided by the stimulus (Λ_1). In this framework, the drift rate of the Wald model equals the log of the ratio of the two evidences:

$$\nu = \ln\left(\frac{Go}{Stay}\right) = \ln\left(\frac{\Lambda_1}{\Omega_1}\right), \tag{2.4}$$

and the evidence accumulation continues until the decision threshold α is reached. The second approach is consistent with the increasing evidence that fixation durations are depending on a competition between the current and potential future fixation locations (Einhäuser et al., 2020). Crucially, both approaches share two main predictions: 1) increasing the width of the attention window increases the amount of information accessible through a single fixation, which has the effect of prolonging (on average) fixation durations, and 2) fixations in areas with a lot of information will have (on average) longer durations than fixations in areas with low information.

2.3 Concrete WALD-EM model

In the previous section, we described the model in conceptual terms. However, in order to implement the model, there are several choices to be made about how to model the contribution of different factors, including their functional forms. Some of these choices will be purely pragmatic and statistical rather than theoretical, and are mostly motivated by the requirement to have a computationally tractable and modifiable model.

The model can be difficult to implement due to the two-dimensional integrals that are used to obtain the values of Λ_1 (total information on the stimulus) and Ω_1 (total information available through the attention window). The ana-

lytic tractability of these integrals relies on the functional form of the functions $\lambda(x, y)$ and $a(x, y)$, and consequently $\omega(x, y)$. This obstacle can be solved in two ways. However, these two approaches are not necessarily exclusive — later we apply a model combining both approaches. The two approaches we present here are not the only possible solutions, but are perhaps the most straightforward. Examples of other possible approaches are discussed in Gelman and Meng (1998), X.-S. Wang and Wong (2007), and Azevedo-Filho and Shachter (1994).

First, it is possible to divide the stimulus into a grid of discrete locations, leading to an approximation of the continuous space, which leads to tractability regardless of the functional forms (i.e., integrals become sums) at the expense of loosing precision due to the discrete approximation. The degree of precision is arbitrary as it can be increased or decreased by changing the size of the cells in the grid, but could quickly lead to a computational bottleneck for fine grained approximations due to the explosion of the number of terms to be summed.

Second, the construction of the functions at play can be carefully selected such that the integrals are analytically tractable. This avoids the need to specify the arbitrary precision of the discrete approach, and potentially leads to less computational burden. However, it may limit the flexibility of the model, as analytic solutions are possible only for a limited number of functional forms.

2.3.1 Modeling λ

The model for the function λ that converts the coordinates of the stimulus to intensity can be achieved in different ways. We generally desire to include different factors in the model, for example central and directional biases, information about locations of objects on the scene, etc. This can be achieved by following (Barthelmé et al., 2013):

$$\lambda(x, y) = \Phi\left(\sum \beta_k f_k(x, y)\right), \tag{2.5}$$

where β_k is a weight of a factor k, f_k is a function that maps factor k to the locations (x, y), and Φ is analogous to a link function in GLMs. Particularly suitable candidates for this function are $\Phi(x) = \exp(x)$, $\Phi(x) = x$, $\Phi(x) =$

$\ln(\exp(x)+1)$ or their combinations (see Barthelmé et al., 2013, for the discussion of the differences between them).

In our application, we use Φ to be an identity function, which by using appropriate restrictions (specified below Equation 2.6) results in a mixture model:

$$\lambda(x,y) = \sum \pi_k f_k(x,y), \tag{2.6}$$

where $\pi_k \in [0,1]$ and $\sum \pi_k = 1$, $f_k(x,y) \geq 0 \, \forall x, y$, and $\int\int f_k(x,y)dxdy = 1$, making the $\lambda(x,y)$ a proper probability density over a plane. The value of π_k then correspond to the relative importance of a factor k, and $f_k(x,y)$ corresponds to a distribution of x and y locations under that factor. By definition, the value of $\Lambda = 1$ (total intensity of stimulus) for whatever setting of the parameters. A particularly attractive property of such definition is the fact that the separation between the factors that represent information on the stimulus from the factors that do not is straightforward. For example, if the first and second factors ($k = 1$ and $k = 2$) encode objects on the screen and saliency (which can plausibly play a role in fixation durations), whereas the third factor ($k = 3$) encodes a central bias (which arguably does not influence fixation durations) then we can simply drop the third factor from the calculations used in the model for fixation durations, and define $\lambda_1(x,y) = \pi_1 f_1(x,y) + \pi_2 f_2(x,y)$, and $\Lambda_1 = \pi_1 + \pi_2$.

Conceptually, a canonical interpretation of such formulation is that the mixture represents a generative model where the observer chooses the next fixation by first randomly selecting a factor k with probability π_k and then selects the location by randomly drawing from the density of the chosen factor f_k (Barthelmé et al., 2013). It is questionable whether this assumption is the most realistic — for example, taking $\Phi = \exp(x)$ (a log-additive model) corresponds to observers combining all factors into one meshed weighted map which determines the next fixation, an approach taken by Barthelmé et al. (2013) — the difference being that whereas the mixture model formulation allows to identify (with some probability) which particular factor was responsible for emitting a particular fixation, it is not the case for other models where all factors cause all fixations at the same time, but some have more influence than others. We believe that which approach is more realistic can be addressed by empirical comparison of different models that differ in these kinds of assumptions.

2.3.2 Calculating Ω

The crucial step is to determine the value of Ω (or Ω_1) — the total intensity available after filtering through the attention window $a(x, y|s)$. Recall that:

$$\Omega = \int\int \omega(x, y)dxdy = \int\int a(x, y)\lambda(x, y)dxdy. \tag{2.7}$$

Given the specification of λ introduced in Equation 2.6, we can rewrite it as:

$$\Omega = \int\int a(x, y) \sum \pi_k f_k(x, y)dxdy = \sum \pi_k \int\int a(x, y)f_k(x, y)dxdy, \tag{2.8}$$

from which it is clearly visible that choice of the functional form of the attention window $a(x, y)$ and the individual factors $f_k(x, y)$ determine whether the model will be tractable without approximation through discretization. One of the possibilities to satisfy this is to model each $f_k(x, y)$ as a bivariate normal distribution, and $a(x, y)$ as a kernel of a bivariate normal distribution. Further, we will assume that the dimensions are uncorrelated, thus $f_k(x, y) = f_k(x)f_k(y)$ and $a(x, y) = a(x)a(y)$, where $f_k(.)$ is a Normal distribution with parameters μ_k and σ_k for the appropriate dimensions, and $a(.)$ is similarly the gaussian kernel with center at the current fixation (s) and scale parameter σ_a in the appropriate dimension. This allows us to rewrite the double integral in Equation 2.8 into a product of two integrals:

$$\Omega = \sum \pi_k \int a(x)f_k(x)dx \int a(y)f_k(y)dy, \tag{2.9}$$

which has a simple analytic solution:

$$
\begin{aligned}
&\int a(x) f_k(x) dx = \\
&= \int \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right] \exp\left[-\frac{(x-s_x)^2}{2\sigma_a^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \int \exp\left[-\frac{(x-\mu_k)^2}{2\sigma_k^2} - \frac{(x-s_x)^2}{2\sigma_a^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \int \exp\left[-\frac{\sigma_a^2(x-\mu_k)^2 + \sigma_k^2(x-s_x)^2}{2\sigma_a^2\sigma_k^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \int \exp\left[-\frac{(\sigma_a^2+\sigma_k^2)\left(x - \frac{\sigma_a^2\mu_k+\sigma_k^2 s_x}{\sigma_a^2+\sigma_k^2}\right)^2 + \frac{\sigma_a^2\sigma_k^2}{\sigma_a^2+\sigma_k^2}(\mu_k - s_x)^2}{2\sigma_a^2\sigma_k^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(\mu_k-s_x)^2}{2(\sigma_a^2+\sigma_k^2)}\right] \int \exp\left[-\frac{(\sigma_a^2+\sigma_k^2)\left(x - \frac{\sigma_a^2\mu_k+\sigma_k^2 s_x}{\sigma_a^2+\sigma_k^2}\right)^2}{2\sigma_a^2\sigma_k^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(\mu_k-s_x)^2}{2(\sigma_a^2+\sigma_k^2)}\right] \sqrt{2\pi}\sqrt{\frac{\sigma_a^2\sigma_k^2}{\sigma_a^2+\sigma_k^2}} \\
&= \frac{\sigma_a}{\sqrt{\sigma_a^2+\sigma_k^2}} \exp\left[-\frac{(\mu_k-s_x)^2}{2(\sigma_a^2+\sigma_k^2)}\right],
\end{aligned}
\tag{2.10}
$$

and equivalently in the y dimension.

Finally, we use the parametrization where the drift rate ν varies with the location of the fixation (Section 2.2.3). Combining the previous two equations, we can write the drift rate as follows:

$$
\begin{aligned}
&\nu|s, \sigma_a, \lambda = \ln(\Lambda_1) - \ln(\Omega_1) \\
&= \ln \sum_{k=1}^{K} \pi_k - \ln \sum_{k=1}^{K} \exp\left[\ln \pi_k + \right. \\
&\left. \sum_{i=1}^{2} \left(\ln \sigma_{ai} - \frac{\ln(\sigma_{ai}^2 + \sigma_{ki}^2)}{2} - \frac{(\mu_{ki} - s_i)^2}{2(\sigma_{ai}^2 + \sigma_{ki}^2)} \right) \right],
\end{aligned}
\tag{2.11}
$$

where the iteration over i only makes explicit the integration over x and y dimensions. The above expression was purposely written in the log-sum-exp-log form explicitly to bring it in line with its computational implementation (which is more stable in this form). In case not all factors provide information, the only change in Equation 2.11 would be the term K in the first summation (e.g., if the first two factors belong to λ_1 but not the third, $K = 3$ would be replaced with $K = 2$).

2.3.3 Likelihood

Assuming data in the form of $d \in \mathbb{R}_+^T$ as the durations and $s = (s_x, s_y) \in \mathbb{R}^{T\times 2}$ as the x and y coordinates of T observed fixations, the likelihood of the model can be written as:

$$\mathcal{L}(\theta|d, s) = \prod_{t=1}^{T} \lambda^{(t)}(s_x^{(t)}, s_y^{(t)}|\pi, \mu, \sigma) \times f_W(d^{(t)}|\nu^{(t)}, \alpha), \tag{2.12}$$

where the superscript for λ means that the intensity function might change during the course of time (which we show later), and $\nu^{(t)}$ changes depending on the current location through Equation 2.11. f_W stands for the p.d.f. of the Wald distribution.

In general, assuming K factors included in the model, the model can have K parameters $\pi_1, \ldots, \pi_K$, $2 \times K$ parameters $\mu_1, \ldots, \mu_K$ (each a vector of 2 for x and y direction), $2 \times K$ parameters $\sigma_1, \ldots, \sigma_K$ (each a vector of 2 for the x and y directions), 2 parameters for σ_a (the width of the attention window in the x and y directions), and the decision boundary α, totalling $5K + 3 - 1$ free parameters. Depending on the actual factors included in the model, we will be able to fix or equate some parameters to reduce the number of parameters to be estimated, although it is not necessary to do so.

2.3.4 Including saliency

An important branch of models that describe and predict distributions of fixation locations are saliency models. In our model, saliency can play a role as one of the factors determining the eye movement behavior. We define a saliency

model (Itti & Koch, 2000; Itti, Koch, & Niebur, 1998) as an algorithm that takes the image (stimulus) as an input, and which produces an output, usually by assigning each pixel a value representing the local saliency of that pixel. Common features that these models consider important are local-global contrasts in color, intensity, and edges.

Saliency models enjoy a lot of success in predicting eye movement behavior and thus it seems reasonable to include some form of a saliency map as one of the factors in our model. Unfortunately, given the nature of the output of saliency models, it is not possible to implement the model fully analytically, and we will instead use discretization. To further reduce the computational complexity, we will reduce the resolution of the output of a saliency map.

Let's define a saliency map as $\mathbf{Sal}$, where each of its element assigns a saliency to a pixel (in this context, pixels can be resized to contain multiple physical pixels of the display). Having I pixels in one dimension and J pixels in the other dimension, we have a total number of $P = I \times J$ pixels. We standardize the output of a saliency algorithm to ensure that $\sum_{p=1}^{P} \mathbf{sal}_p = 1$.

To include saliency into the model for *where*, we obtain a representation of the saliency on a continuous space of the x and y coordinates by defining the intensity function of saliency as a two dimensional step function:

$$f(x, y) = \frac{\mathbf{sal}_{p(x,y)}}{h \times w}, \tag{2.13}$$

where $p(x, y)$ returns the index of a pixel which is a super set of the position x and y, and where h and w is the height and width of the pixel. Standardization by the area of the pixel ensures that after converting the saliency map $\mathbf{Sal}$ to the intensity function, the volume $\int \int f(x, y)\, dxdy$ amounts to 1.

To include saliency into the model for *when*, we need to adopt additional simplifications as to evaluate the integral $\int \int a(x, y) \times f(x, y)\, dxdy$. We define x_p and y_p as the x and y coordinates of the center of a pixel p, respectively, and approximate $f(x, y)$ as:

$$f(x, y) \approx \sum_{p=1}^{P} \mathbf{sal}_p \text{Normal}(x|x_p, \kappa)\text{Normal}(y|y_p, \kappa), \tag{2.14}$$

which leads to (using results in Equation 2.8):

$$\int\int a(x,y)f(x,y)\,dxdy \approx \sum_{p=1}^{P} \mathbf{sal}_p \frac{\sigma_a^2}{\sigma_a^2+\kappa} \exp\left(-\frac{(x_p-s_x)^2+(y_p-s_y)^2}{2(\sigma_a^2+\kappa)}\right), \tag{2.15}$$

which we can further simplify by letting $\kappa \to 0$:

$$\int\int a(x,y)f(x,y) \approx \sum_{p=1}^{P} \mathbf{sal}_p \exp\left(-\frac{(x_p-s_x)^2+(y_p-s_y)^2}{2\sigma_a^2}\right). \tag{2.16}$$

These steps enable us to approximately compute the drift rate by substituting the discrete saliency map with a continuous function.

However, this implementation still requires serious computational resources: for example, fitting a model that includes a saliency map of resolution of 800×600 pixels would mean summing up $P = 800 \times 600 = 480,000$ terms for every fixation in every iteration of the fitting procedure.

There are generally 3 ways to alleviate the problem of the computational complexity. First, it is possible not to include the discrete factor in the model for *when*, but only include it into the factor for *where*. However, leaving it out does not solve the problem, but rather avoids it altogether. Second, it is possible to downsample the output of the saliency map. Indeed, many saliency algorithms already output the saliency map that has a resolution smaller than the original image (e.g., by a factor of 16 in each of the dimensions, Itti et al., 1998). Having an input image of dimensions of 800×600 pixels then leads to quite substantial reduction: Instead of summing up 480,000 terms we need to sum up only about 2,000. Downsampling the saliency maps to have smaller resolution than the input image is also desirable from a measurement perspective as the eye-tracking devices likely have measurement error that translate to several pixels of the input image. Downsampled saliency maps then correspond better to the level of precision of the data. Third, it is possible to limit the summation only for the terms that lie in a relative proximity from the current fixation. For example, the attention window lets through only at most 1.1% of the total weights of the pixels that lie at a distance of $3\sigma_a$ or more, essentially meaning

that many of the terms in the sum are basically zero. Leaving out the pixels that are that far from the current fixation can reduce the number of terms to be summed by a great amount while not sacrificing much of the computational accuracy. Downsampling and limiting the summations are not mutually exclusive and can be used at the same time — an approach we take in the practical implementation of our model.

2.4 Application: Infant scene viewing

Here, we apply a particular realization of the model to data by van Renswoude, Voorvaart, van den Berg, Raijmakers, and Visser (in prep) to demonstrate its use in applied context. This data set was originally collected with the aim to investigate the role of bottom up versus top down factors in infants' eye movements. Specifically, the data set was collected to assess whether object familiarity is associated with specific patterns in eye movement behavior of infants when looking at pictures of real world scenes. In the following application, we are interested in the extent to which four different factors influence the distribution of fixation locations and the timing of saccades.

The four factors that we considered are the 1) locations (and sizes) of objects on the scene (van Renswoude et al., in prep; J. Xu et al., 2014), 2) saliency (Itti & Koch, 2000, 2001; Itti et al., 1998), 3) exploitation (i.e., tendency to make repeated fixations in a relative proximity to a previous fixation; Malem-Shinitski et al., 2020), and 4) central bias (van Renswoude et al., 2019).

The model was fitted on half of the data set (with the other half used for a following cross-validation). To accommodate individual differences between participants, we generalized the model using hierarchical modeling (the details are explained below). As such, we obtain assessment of the individual differences between the participants in terms of their tendency to dwell longer on current locations (captured by the decision boundary), and the width of their attention window.

2.4.1 Data Descriptives

The data contains recordings of 47 participants looking at 29 static pictures selected from the pool of 700 images created by J. Xu et al. (2014). 39 participants

looked at all 29 stimuli (min = 5, mean = 27.6, median = 29, max = 29 viewed images per participant). The mean number of fixations per trial was 11.4 (sd = 4.3); the total number of fixations in the data set is 14,807.

We split the data set in two parts, one of which we used to estimate the parameters of a model (number of fixations = 7,207), and one of which we used to validate the predictions of the model. We counter balanced the number of trials per participant in the two sets to ensure that both data sets contain some data from all participants and all items. The details of the procedure are available at `github.com/Kucharssim/WALD-EM/blob/master/scripts/prepare_data.md`.

The exact form of splitting the data in this way had the following reasons. First, the aim of this article is not to generalize findings, but rather as a conceptual proof of concept — that the model is applicable to eye tracking data and captures some interesting patterns in the data. Second, being able to generalize to a population is contingent on additional requirements besides cross-validation procedure, such that the participants and stimuli were randomly selected from the target population. We did not define which population of infants we would like to generalize to, and don't assess whether they indeed represent that population. Further, we *know* for certain that the 29 stimuli are not randomly selected from the pool of 700 images by J. Xu et al. (2014), making it difficult to generalize even to this pool, and even more problematic to some more general population of static images. Lastly, in terms of our goals, our procedure is more robust against differences between the train and test sets caused by randomly splitting small sets of data. Usually, for smaller data sets, procedures such as k-fold cross validation (or leave-one-out) is usually done for this purpose, compared to split half cross validation. However, k-fold cross validation was not an option due to the computational demands of the model. Thus, by giving up aspirations for generalizing our findings, our splitting procedure ensured that we could still perform cross-validation, but ensuring that potential problems of cross validation are not caused by randomly choosing an "outlying" participant or stimulus into the train or test sets.

2.4.2 Initial Model

The model contains four factors that determine the fixation locations, two of which are included in the model for fixation duration.

The model for *where* is composed of four factors, and so we can describe the distribution of fixation locations as follows:

$$(x, y) \sim \sum_{k=1}^{4} \pi_k f_k(x, y|\theta_k), \tag{2.17}$$

where π_k are the weights of different factors and f_k is the distribution of a factor k with parameters θ_k.

The first factor is the location and sizes of objects on the scene. We assume that each object on the scene can have different level of attractivity and that larger objects distribute their total attractivity over larger area. This idea can be expressed by another mixture:

$$\begin{aligned} f_1(x, y|\theta_1) = \sum_j \omega_j \text{Normal}(x|\text{center}_{xj}, \gamma \times \text{width}_j) \times \\ \text{Normal}(y|\text{center}_{yj}, \gamma \times \text{height}_j), \end{aligned} \tag{2.18}$$

where ω_j are the individual attractivities of different objects on a particular image. Parameter γ is a scaling factor that stretches or compresses the attractivity of objects proportionally to their sizes.

The second factor is the saliency, which we treated as described in Equation 2.13.

The third factor can be described as an exploitation factor, and captures the phenomenon that people tend to linger close to the current fixation location: we model it as a bi variate normal distribution centered at the fixation location at time t to predict the fixation location at time $t + 1$:

$$f_3(x, y|\theta_3) = \text{Normal}(x|s_x^t, \sigma_e)\text{Normal}(y|s_y^t, \sigma_e) \tag{2.19}$$

The fourth factor represents the central bias, and is modelled as a bi variate normal distribution centered at the center of the screen ($x_c = 400$, $y_c = 300$)

$$f_4(x, y|\theta_4) = \text{Normal}(x|400, \sigma_d)\text{Normal}(y|300, \sigma_d) \tag{2.20}$$

For the fixation durations, we only consider the first two factors influential: the latter two factors do not stand for *information* presented on the screen, but rather spatial biases, and therefore should not have any influence on saccade timing.

The model for fixation duration can be summarised as following:

$$d \sim \text{Wald}(\nu, \alpha)$$

$$\nu = \ln(\pi_1 + \pi_2) - \ln\left(\int\int a(x, y|\sigma_a)\left(\pi_1 f_1(x, y|\theta_1) + \pi_2 f_2(x, y|\theta_2)\right) dxdy\right).$$

We also modelled the individual differences between participants by modelling their decision boundary and width of the attention window as random terms. Because both of these parameters need to be positive, we modelled them on the log scale:

$$\ln(\alpha_i) = \mu_\alpha + z_i * \sigma_\alpha$$
$$z_i \sim \text{Normal}(0, 1),$$

where α_i stands for the decision boundary of participant i, and μ_α with σ_α are the estimated group mean and standard deviation of the parameter α on the log scale. The same approach was taken for the attention window σ_a.

We used weakly informative priors on the parameters that were based on prior predictive simulations done when building the model (Schad, Betancourt, & Vasishth, 2019). The priors are accesible at the model file: `git.io/JfjuJ`.

We implemented the model using the probabilistic programming language Stan (B. Carpenter et al., 2017) interfacing with R (R Core Team, 2020) using the package `rstan` (Guo, Gabry, & Goodrich, 2020). The following additional R packages were used to produce the output (in no particular order):

`OpenImageR` (Mouselimis, 2019), `Rcpp` (Eddelbuettel et al., 2020; Eddelbuettel & François, 2011), `ggforce` (Pedersen, 2019a), `gtools` (Warnes, Bolker, & Lumley, 2020), `here` (Müller, 2017), `imager` (Barthelme, 2020), `knitr` (Xie, 2014, 2020), `patchwork` (Pedersen, 2019b), `plotrix` (Lemon, 2006; Lemon et al., 2019), `pracma` (Borchers, 2019), `tidybayes` (Kay, 2020), and `tidyverse` (Wickham, 2019). Throughout the development of the model, we conducted simulation studies to validate our implementation. The results are documented in the following folder of our project repository `github.com/Kucharssim/WALD-EM/tree/master/documents`. The current (small) simulation results are encouraging in terms of parameter recovery, but we did not invest our resources into a full validation study due to the computational demands of the model.

2.4.3 Results — Initial Model

We ran 10 MCMC chains with random starting values and default tuning parameters set by Stan. Each chain ran for 1,000 warm up and 1,000 sampling iterations, resulting in a total of 10,000 samples used for inference. The model ran without any divergent transitions. We examined the potential scale reduction factor $\hat{R}$, trace plots, auto correlations, and the number of effective samples to identify potential problems with convergence. We did not find indications of poor convergence, and thus proceed with interpreting the model.

Posterior predictive checks. We generated posterior predictives for the data set used for estimating the parameters, and for the hold out data set, to assess whether the fitted model reproduces the observed patterns in the data, and whether the patterns that the model picks up from the data carry over to the hold out data set. This enables us to contrast features that are desirable to be captured by the model (i.e., patterns that are present both in the fitting and hold out data set), from features that are not so desirable to be captured by the model (i.e., patterns that are present in fitting but not hold out data set).

The model is able to capture the characteristic distribution of the fixation durations, as documented in Figure 2.3, although the model predicts a slightly fatter right tail than that of the data. We also inspected the model's predictions of the fixation durations for individual participants, to assess whether it captures their individual differences. Figure 2.4 shows that the model is well able

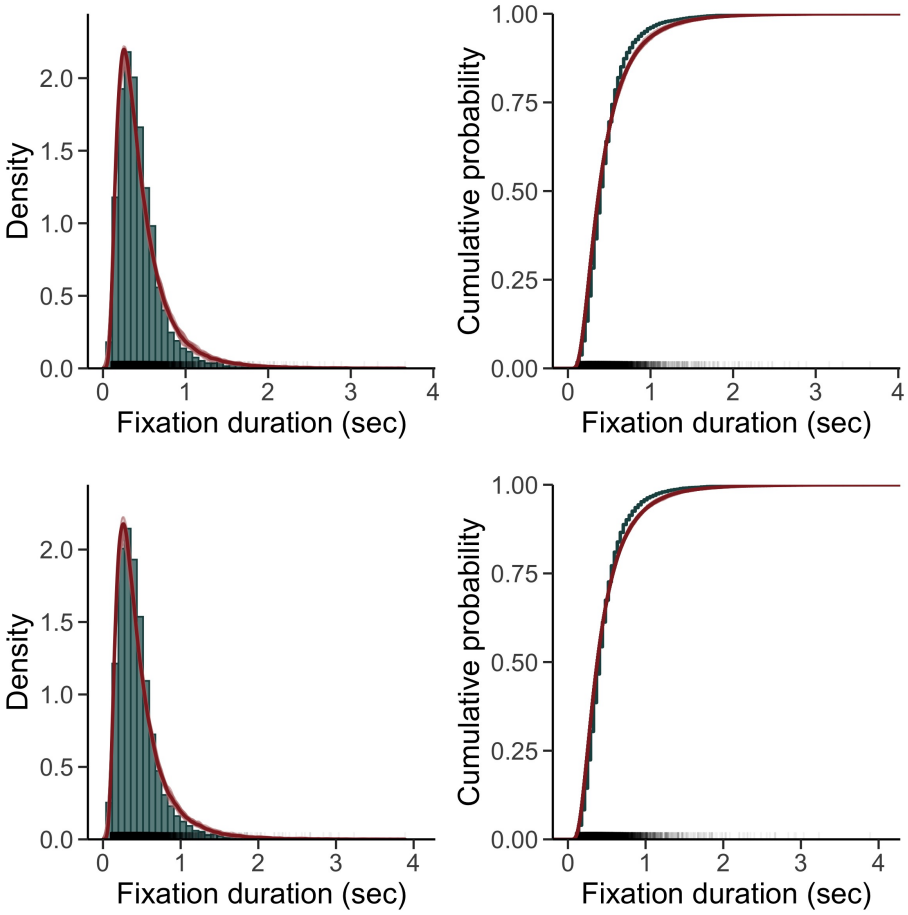


Figure 2.3: Predicted (red) versus observed (green) distribution of the fixation durations; predictions using the initial model. Left panel shows histogram of the empirical data versus the density estimate using gaussian kernel of the posterior predictives. Right panel shows empirical cumulative distribution functions. Top panel shows the data used for fitting the model, bottom panel the hold out data used for cross-validation.

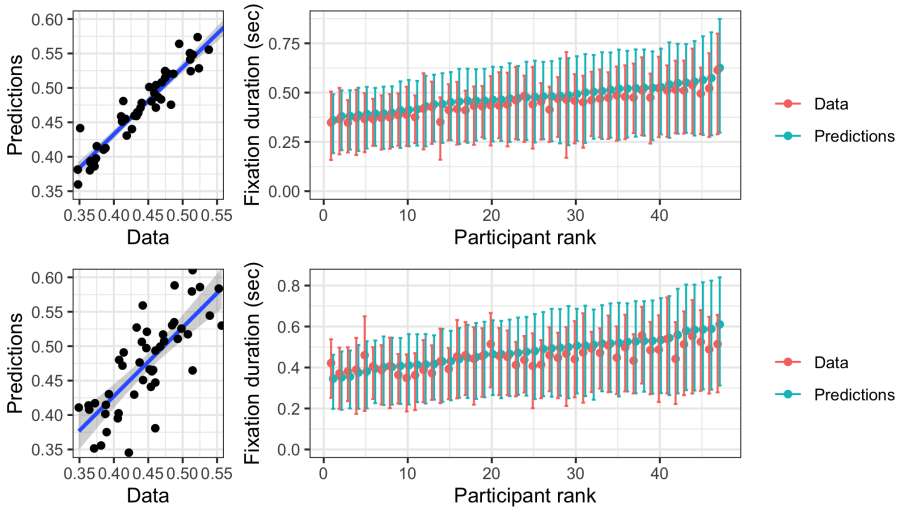


Figure 2.4: Posterior predictive checks for the individual differences in fixation durations. Left panel shows the observed (x-axis) and predicted (y-axis) mean fixation durations for each participant separately. Right panel shows the observed (red) versus predicted (blue) mean fixation durations, 20% and 80% quantiles (whiskers). The top panel shows the results in training set, bottom panel shows the results in the test set.

to capture individual differences between participants in respect to their fixation durations.

The model also reproduces the distributions of fixation locations. Figure 2.5 shows an example for one particular stimulus (image number 251 from J. Xu et al., 2014). The top-right in Figure 2.5 displays the four factors included in the model, which combine proportionally to their weights to the posterior predictive distribution (labeled as predicted fixations).

Next to the variables used to fit the data (fixation durations and fixation locations), we also checked on other quantities implied by the model. Specifically, we checked whether we can reproduce the distribution of saccade amplitudes and the distribution of saccade angles (as these derivative measures provide additional insights into the model performance, e.g., J. M. Henderson & Hayes, 2018). Saccade amplitude was measured as the Euclidean distance between two successive fixations in units of pixels. Saccade angle was calculated as an angle in

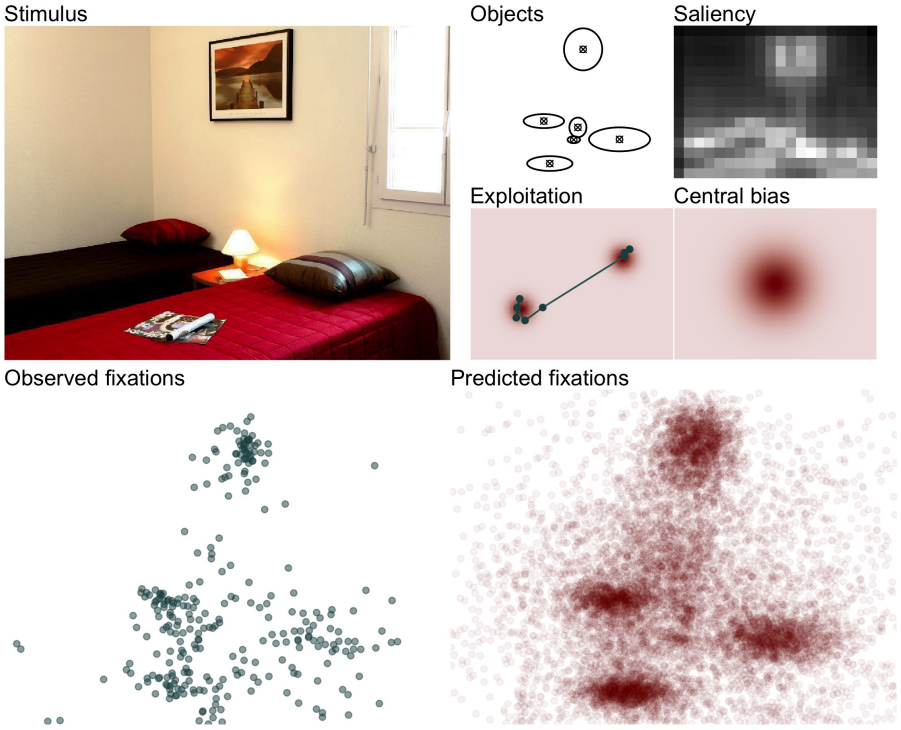


Figure 2.5: Example of the model for fixation locations on the stimulus created by J. Xu et al. (2014) (top left panel). The four factors influencing the fixation locations are depicted in the top right quadrant. The bottom left panel shows the observed fixation locations, and the bottom right the draws from the posterior predictive distribution of the fitted model.

radians between the horizontal axis of the screen and the vector that connects two successive fixations.

Figure 2.6 shows the observed versus predicted distributions of saccade amplitudes on one example stimulus (shown in Figure 2.5). The model usually captures the distribution of saccade amplitudes relatively well, exhibiting two modes. Figure 2.6 shows the observed versus predicted distributions of saccade amplitudes on the same stimulus. The prediction of saccade angles is relatively good, as the model picks up patterns of saccade directions specific to a specific stimulus. As an illustration, Figure 2.7 shows that the model captures saccade directions in the top-right and bottom-left directions on the stimulus shown in Figure 2.5. However, overall the model does not capture well an excess of saccades in the horizontal direction (see Figure 2.12), which could be an indication that the model needs to be expanded with a factor that represents a horizontal bias (van Renswoude et al., 2016).

Parameter estimates The results indicated that the most important factor driving fixations was the locations of objects on the scene (weight = 0.37, 95% CI[0.35, 0.40], followed by exploitation (weight = 0.30, 95% CI[0.28, 0.31]), saliency (weight = 0.18, 95% CI[0.16, 0.20]), and central bias (weight = 0.16, 95% CI[0.14, 0.17]).

The parameter that controls sizes of objects as identified by J. Xu et al. (2014) indicated that people fixate relatively close to the centroids of the objects (scale = 0.23, 95% CI[0.22, 0.24]). The exploitation region had a standard deviation $\sigma = 34.58$ (95% CI[33.15, 36.06]) pixels, whereas the central bias region had a standard deviation $\sigma = 93.84$ (95% CI[88.65, 98.91]) pixels. Relatively speaking, central bias is less focused than the exploitation factor.

2.4.4 Extended Model

The original model fared well capturing the distribution of fixation durations and the overall distribution of fixation locations, and was able to a small degree to capture an excess of horizontal saccades without this being explicitly built into the model. However the discrepancy between the model's predictions and data show that the tendency to make horizontal saccades is particularly noteworthy and possibly needs an extra explanation.

To explore whether we can improve the model's capability to reproduce the

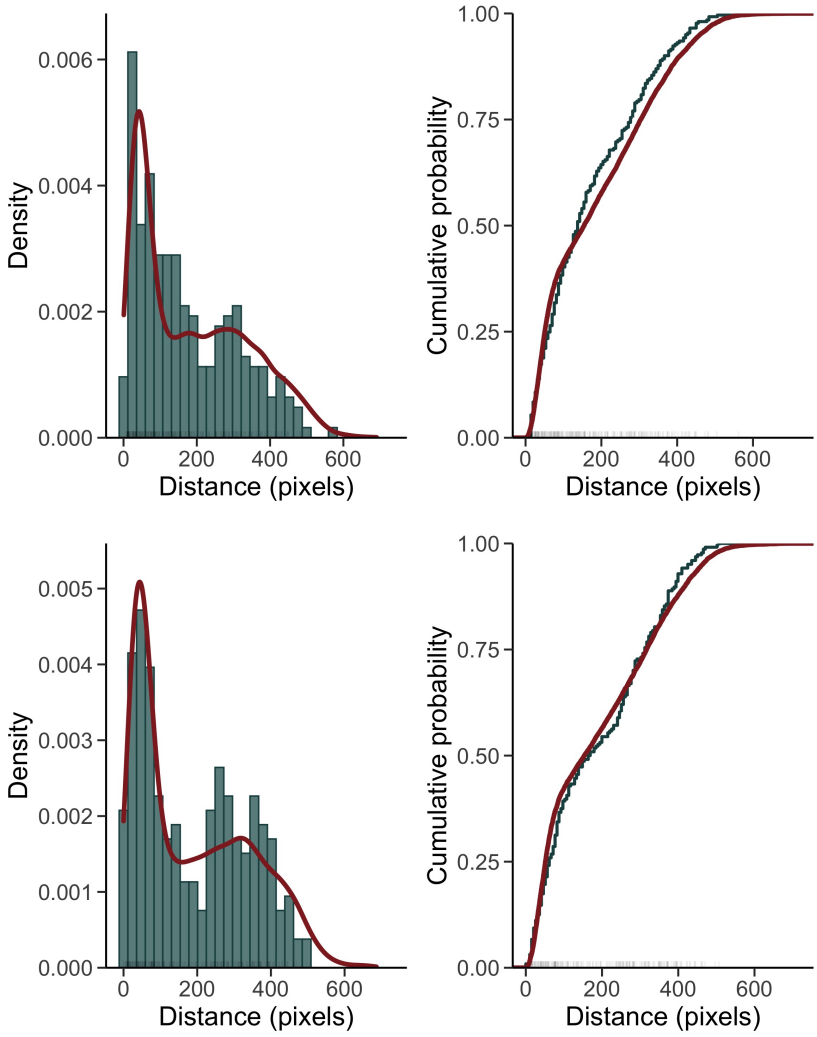


Figure 2.6: Observed (blue) versus predicted (red) saccade amplitude on one particular stimulus; predictions using the initial model. Top panel shows the data used for fitting the model, bottom panel the hold out data used for cross-validation.

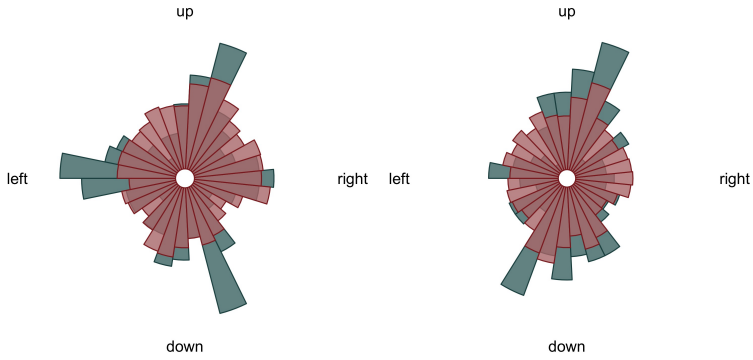


Figure 2.7: Observed (blue) versus predicted (red) saccade angle on one particular stimulus. Plot on the left shows the data used for fitting the model, plot on the right shows the hold out data used for cross-validation.

amount of saccades in the horizontal direction, we extended the model. Specifically, we added another factor into the model for fixation locations, representing the horizontal bias. To create a factor that represents a saccadic bias (such as horizontal bias), instead of location preferences, it is possible to transform fixation locations (x and y coordinates) into a saccade representation (angle θ and amplitude r of a saccade):

$$\begin{aligned} \theta &= \arctan\left(\frac{\Delta y}{\Delta x}\right) \\ r &= \sqrt{\Delta x^2 + \Delta y^2}, \end{aligned} \tag{2.21}$$

where $\Delta x = x^t - x^{t-1}$ and $\Delta y = y^t - y^{t-1}$ represent a fixation as the difference of the x and y coordinates compared to the previous fixations (we set $x^0 = 400$ and $y^0 = 300$ as that is the middle of the screen).

That way, we can substitute a factor of locations with a factor of saccade angles and amplitudes:

$$f(x, y) = \frac{f(\theta, r)}{r}, \tag{2.22}$$

where the denominator r is the Jacobian determinant representing the stretching of the space after the change of variables from cartesian to polar coordinates: $dxdy = rdrd\theta$.

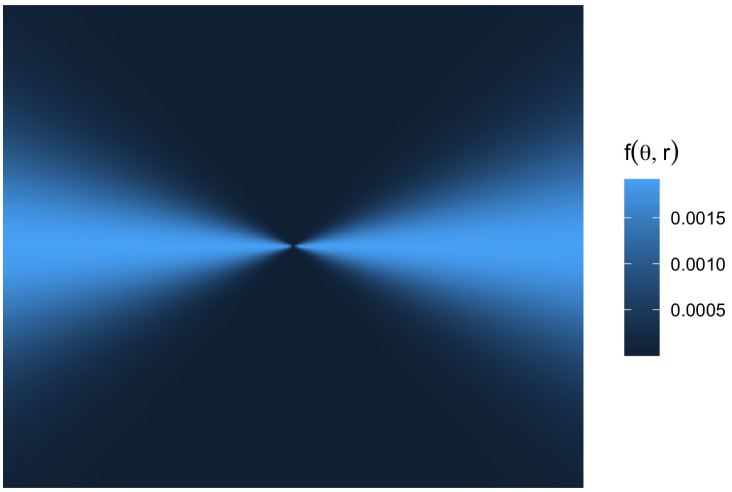


Figure 2.8: Example of the joint density of saccade angle and saccade amplitude plotted on the screen dimensions. The density highlights saccades in the left and right directions relative to the current fixation (in this figure, the center of the screen), representing the horizontal bias.

To create the joint density of the angle and amplitude, we express it using the chain rule of probability:

$$f(\theta, r) = f(\theta) \times f(r|\theta). \tag{2.23}$$

The important part of this factor is the distribution of saccade angles, for which we specify the following distribution:

$$f(\theta) = 0.5\text{vonMises}(0, \kappa) + 0.5\text{vonMises}(\pi, \kappa), \tag{2.24}$$

which specifies a mixture of von Mises distributions with centers fixed to 0 and π (i.e., right and left direction, respectively), and a concentration κ which is estimated from the data. The mixture weights are fixed to 0.5 as we assume that saccades in the left direction are equally attractive as saccades to the right direction.

The conditional density $f(r|\theta)$ is chosen to be a uniform stretched over the interval between 0 and the maximum saccade lenght that would not fall

outside of the screen if it was launched from the position (x^{t-1}, y^{t-1}) under the direction θ.

The generative mechanism for such a joint density is the following. First, the observer draws a saccade angle from the distribution $f(\theta)$. Then, the observer draws a point along a line under the sampled angle θ, that goes between location (x^{t-1}, y^{t-1}) and the edge of the screen. This point is the new fixation position.

Figure 2.8 shows and example of the function $f(\theta, r)$ on the screen coordinates, with (x^{t-1}, y^{t-1}) set to the center to the screen, and $\kappa = 15$.

The rest of the model stayed the same.

2.4.5 Results — Extended Model

We fitted the extended model in the same way as the initial model: We ran 10 MCMC chains with random starting values and default tuning parameters set by Stan. Each chain run for 1,000 warm up and 1,000 sampling iterations, resulting in a total of 10,000 samples used for inference. The model ran without any divergent transitions. We examined the convergence diagnostics, to find that we could not identify potential problems with convergence. Thus, we proceed with interpreting the model.

Posterior predictive checks. We conducted posterior predictive checks the same way as with the previous model: Comparing the predicted and observed distribution of fixation durations, fixation locations, saccade amplitudes, and saccade angles. The extended model performed similarly to the initial model in terms of the first three variables (see Figures 2.9 and 2.11). As Figure 2.12 demonstrates, the extended model did better in terms of reproducing the overall distribution of saccade angles - being able to reproduce the excess of saccades going in the horizontal direction better after we have explicitly added a factor that represents horizontal bias. However, it is still visible that there is still some potential to improve the model predictions.

Model comparison using cross-validation. To assess whether the extended model did better at predicting the data compared to the initial model, we computed the log-likelihood of the hold-out set under the two models, given the posterior distributions of the parameters. This way, we obtained a distributions of the log-likelihood for the two models based on their out of sample per-

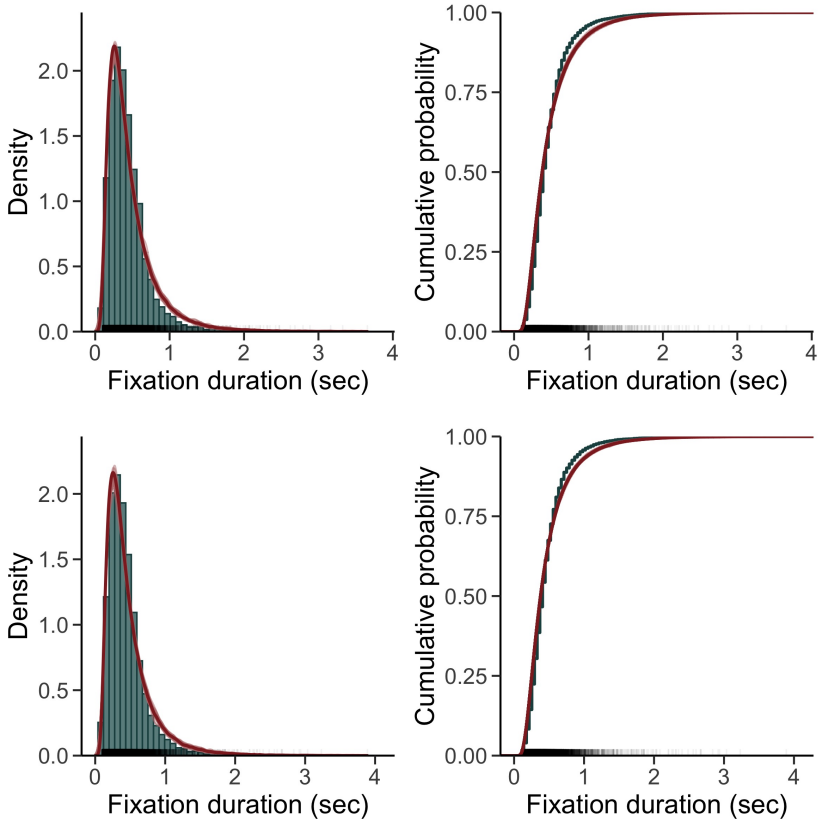


Figure 2.9: Predicted (red) versus observed (green) distribution of the fixation durations; predictions using the extended model. Left panel shows histogram of the empirical data versus the density estimate using gaussian kernel of the posterior predictives. Right panel shows empirical cumulative distribution functions. Top panel shows the data used for fitting the model, bottom panel the hold out data used for cross-validation.

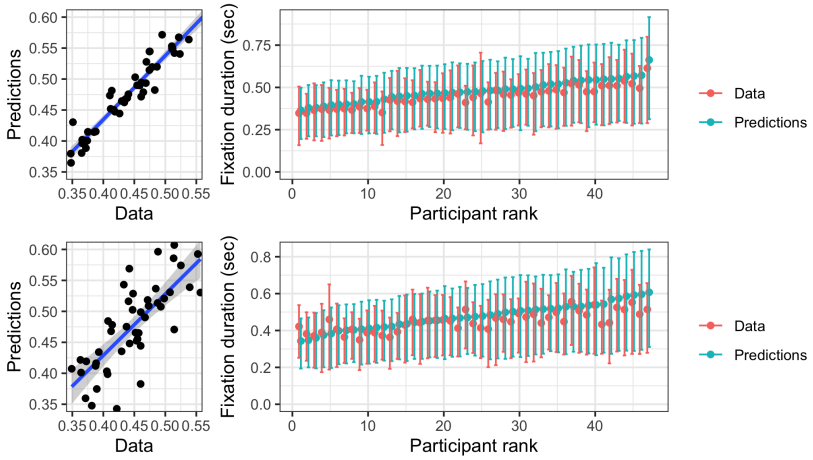


Figure 2.10: Posterior predictive checks for the individual differences in fixation durations. Left panel shows the observed (x-axis) and predicted (y-axis) mean fixation durations for each participant separately. Right panel shows the observed (red) versus predicted (blue) mean fixation durations, 20% and 80% quantiles (whiskers). The top panel shows the results in training set, bottom panel shows the results in the test set.

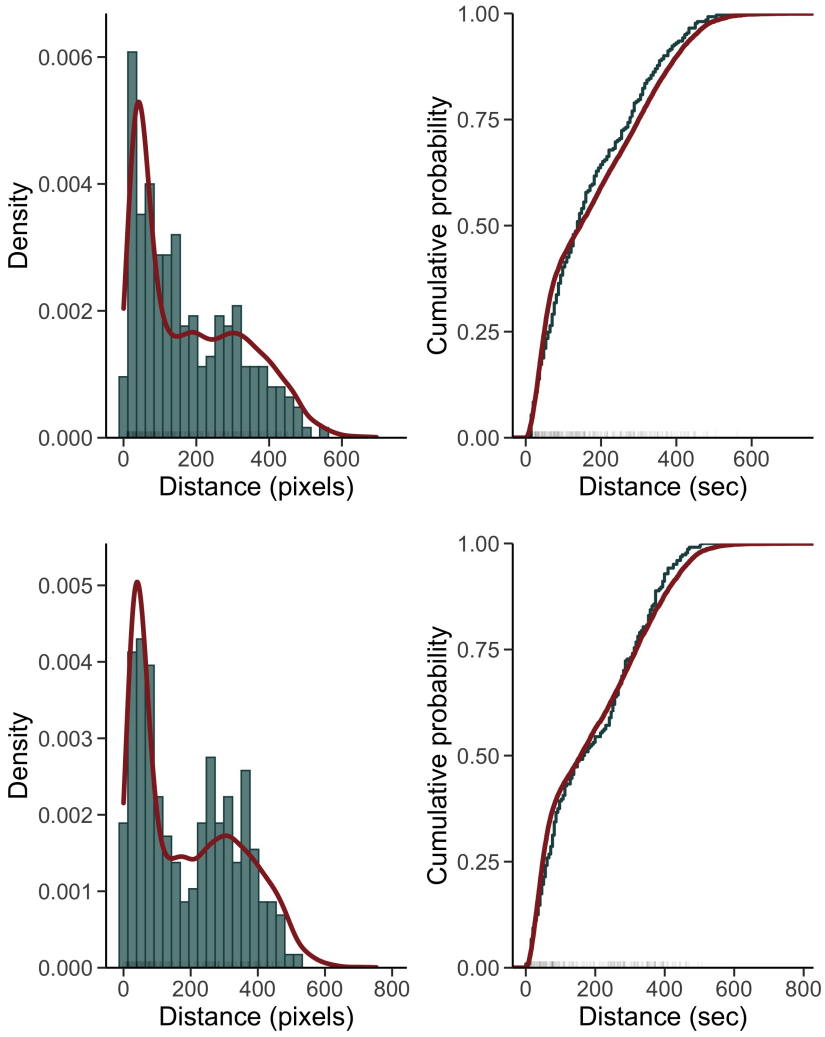


Figure 2.11: Observed (blue) versus predicted (red) saccade amplitude on one particular stimulus; predictions using the extended model. Top panel shows the data used for fitting the model, bottom panel the hold out data used for cross-validation.

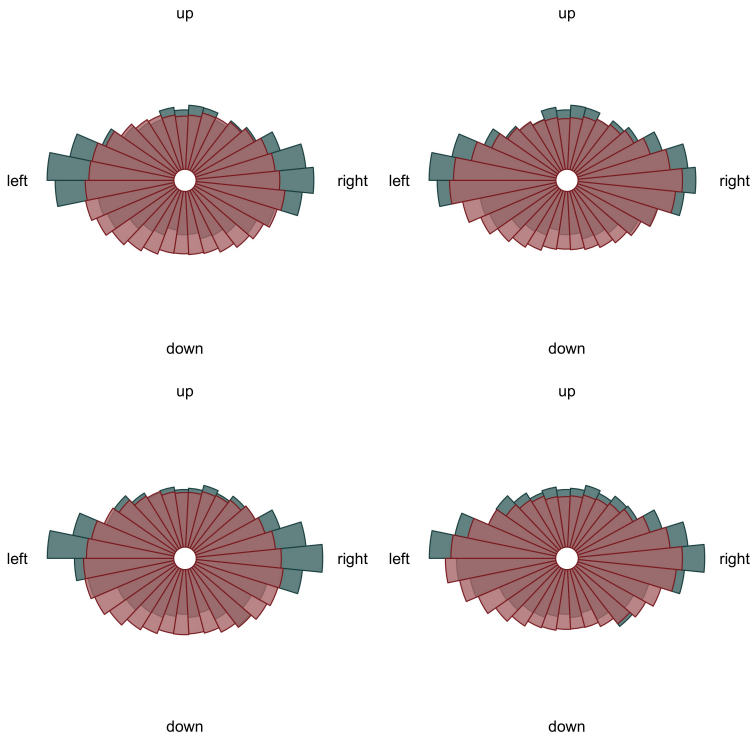


Figure 2.12: Predicted versus observed distribution of saccade angles under the initial (left) and extended (right) model, over all stimuli in the data set; the top panel displays the data set used to fit the model, bottom displays the hold out data set.

			Quantiles	
Factor	Mean	SD	2.5%	97.5%
Objects	0.36	0.01	0.34	0.39
Saliency	0.17	0.01	0.16	0.19
Exploitation	0.33	0.01	0.32	0.35
Central bias	0.13	0.01	0.11	0.15

Table 2.1: Descriptives of the posterior distribution of the factor weights under the initial model.

formance which we use for cross-validating the models. To compare the two distributions, we computed the distribution of the log-likelihood differences: $\Delta \log \mathcal{L} = \log \mathcal{L}(\text{Model 2}) - \log \mathcal{L}(\text{Model 1})$: Positive values mean that the extended model predicted the hold-out data better than the initial model, and negative values mean that the initial model predicted the hold-out data better than the initial model.

The log-likelihood difference distribution (median=45.77, IQR [15.18, 77.06]) indicated that the extended model was better at predicting the hold-out data than the intitial model: adding horizontal bias into the model increased the model's predictive success.

Parameter estimates. The estimates indicated that the most important factor were the objects on the scene (weight = 0.35, $95\%\text{CI}[0.33, 0.37]$, followed by exploitation (weight = 0.31, $95\%\text{CI}[0.29, 0.32]$, saliency (weight = 0.14, $95\%\text{CI}[0.13, 0.16]$, central bias (weight = 0.13, $95\%\text{CI}[0.11, 0.15]$, and lastly the horizontal bias (weight = 0.07, $95\%\text{CI}[0.06, 0.8]$).

The parameters that control the individual factors were very similar to those of the initial model. The parameter that controls sizes of objects indicated that people fixate relatively close to the centroids of the objects (scale = 0.23, $95\%\text{CI}[0.22, 0.24]$). The exploitation region had a scale ($\sigma = 34.28$, $95\%\text{CI}$ $[32.80, 35.78]$) of about a third of that of the central bias ($\sigma = 98.68$, $95\%\text{CI}$ $[93.42, 103.99]$). The additional parameter that controls the concentration of the horizontal bias was estimated to $\kappa = 18.36$, $95\%\text{CI}[13.4, 24.11]$.

			Quantiles	
Factor	Mean	SD	2.5%	97.5%
Objects	0.35	0.01	0.33	0.37
Saliency	0.14	0.01	0.13	0.16
Exploitation	0.31	0.01	0.29	0.32
Central bias	0.13	0.01	0.11	0.15
Horizontal bias	0.07	0.01	0.06	0.08

Table 2.2: Descriptives of the posterior distribution of the factor weights under the extended model.

2.5 Benefits of joint modeling of fixation locations and fixation durations

The application of the model presented in the article showed that the model is able to fit a particular data set relatively well. The strength of our approach is that it can model both fixation durations and fixation locations concomitantly. This has two general benefits compared to models that consider fixation durations and fixation locations separately. In this section, we discuss these benefits more explicitly.

First, having different sources of data gives us more information to estimate parameters of interest. For example, one could fit some variant of a mixture model to fixation locations to estimate the importance of various factors that influence eye movements. However, in our model, the weights of these factors not only play a role for the fixation locations, but come into play when calculating the drift rate of the Wald process, therefore they are informed by the fixation durations as well. This benefits parsimony of our models as well as introduces the potential to estimate parameters with greater precision, allowing to even estimate parameters that would have been otherwise hardly identifiable.

Second, modeling fixation locations and durations together enables us to capture some dependencies between the two. Potentially, this could lead to modeling phenomena that occur in both spatial and temporal dimensions. In the case of the current approach, the model has a built-in global dependency between locations and durations due to the way it represents their joint probabil-

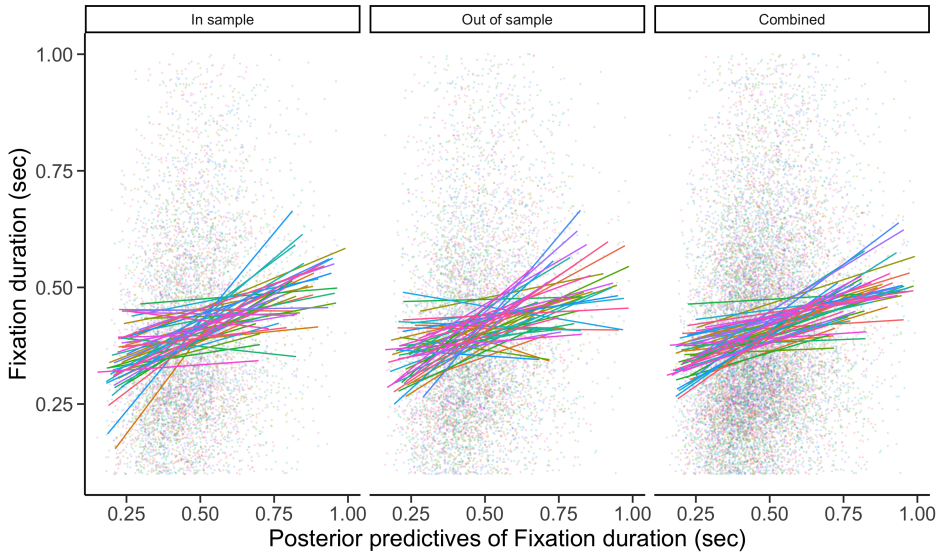


Figure 2.13: The mean of the posterior predictive for the fixation durations (x-axis) vs the observed fixation durations (y-axis) for each participant separately. The scatterplot suppresses outliers above 1 sec. Regression lines are superimposed to highlight the variablity between participants.

ity. Specifically, the distribution of fixation durations depends both on the individual characteristics of the observer (by having two parameters vary between participants), but also on the surroundings of the current fixation. The surroundings of the observers' fixation are taken into account when evaluating the drift rate of the Wald process, where the intensity function is passed through the attention window. The intensity function is concurrently the function that (stochastically) determines fixation locations. Thus, the model makes the prediction that fixations on locations that are more likely to be fixated than others will (on average) be longer than locations that are relatively less likely to be fixated. Conceptually, locations with a lot of information will have a lot of attraction, and lead to more fixations and longer durations.

To test that the model's predictions are accurate in this respect, we derive the following two hypotheses[3] that follow from the model's predictions. (1) The model is able to predict individual fixation durations. Because the model al-

[3]N.B. these hypotheses were formulated post hoc, based on the feedback from an anonymous reviewer to whom we are thankful for this idea.

ready explains individual differences in fixation durations, we will look into the correlation between predicted and observed fixation duration for each participant separately. A positive correlation within a participant suggests that the model is able to pick up information around the fixation location to adjust its prediction about the fixation duration. (2) Both in the data and in the model, there exists a positive correlation between how often (or how likely) a particular area on the image is fixated, and a mean fixation duration in that area. To test this hypothesis, we split images into a grid and calculate the correlations for each image separately (there might be differences between images that we left unmodelled). For both hypotheses we used the Bayesian inference scheme for correlations developed by Ly et al. (2018). We present the results both for the training set, the test set, and with the two sets combined. Here we show the results from the extended model. The results from the initial model are nearly identical.

Figures 2.13 and 2.14 show the results related to hypothesis (1) that the model's predictions positively correlate with the observed fixation durations. Figure 2.13 shows the scatterplot between the mean of the predictive distribution for each fixation duration and the observed fixation durations, superimposed by the regression line for each participant separately. Figure 2.14 shows the corresponding Pearson's correlation coefficients and their Bayes factors (testing null hypothesis of no correlation versus alternative hypothesis of positive correlation). We calculated the same for the train set ("In sample"), the test set ("Out of sample") and the two data set combined ("Combined"). Overall, most of the correlations were positive, suggesting that the model is able to pick up some information about the surroundings of a fixation to inform the fixation duration. However, the correlations were relatively low and some correlations remained in the region where the Bayes factor does not strongly prefer either the null or the alternative, suggesting that much of the variability of fixation durations is yet to be explained by additional mechanisms.

Figures 2.15 and 2.16 show the results related to hypothesis (2) that there is a relationship between fixation duration and fixation probability, i.e., that locations that are more frequently visited are also fixated with longer durations. To assess this hypothesis we first split each image into a grid of 50×50 pixels, leading to 16×12 cells in the grid in each image. For each cell, we calculated the

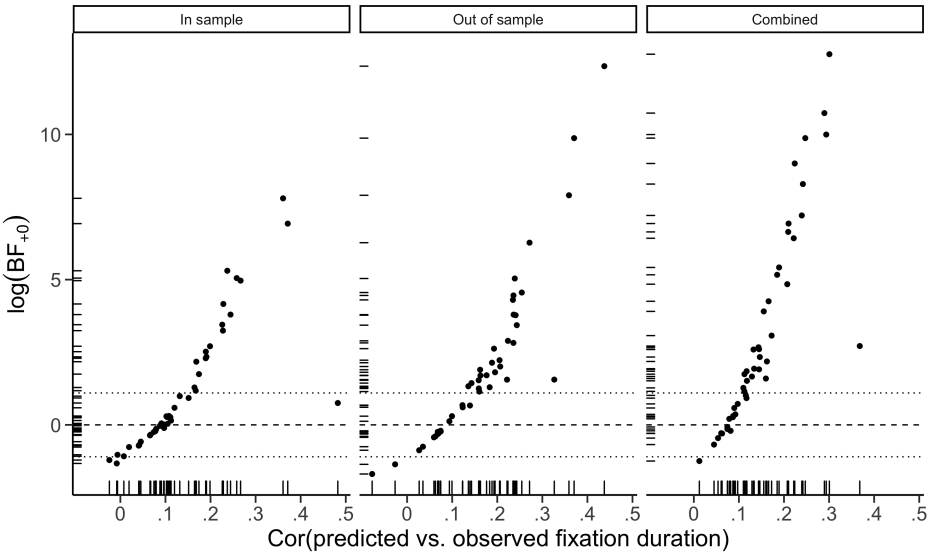


Figure 2.14: The correlations between predicted mean fixation duration and observed fixation duration for each participant separately and the log Bayes factors testing the null model (no correlation) vs the alternative (positive correlation, specified by stretched Beta(10,10) truncated at zero). The dotted lines show the region of "anectdotal" evidence (Bayes factor between 1/3 and 3), i.e., that there is not enough information to say anything meaningful about presence or absence of the correlation.

probability of fixating inside of it using the model's predictive distribution for the fixation locations, and the predicted and observed mean fixation durations. As before, we conducted this calculation for the train set, test set, and combined data.

Figure 2.15 shows the scatterplot of the log probability of fixating a particular cell of the grid and the log mean fixation duration, with superimposed regression line for each image separately. The top panel shows the relation that was found in the data, the bottom panel shows the relation that is reproduced by the model.

Figure 2.16 further shows the observed Pearson's correlation and the corresponding Bayes factors testing the null hypothesis of no correlation versus the alternative hypothesis of positive correlation. For the majority of images (except for three images that show correlations near zero), there appears to be a

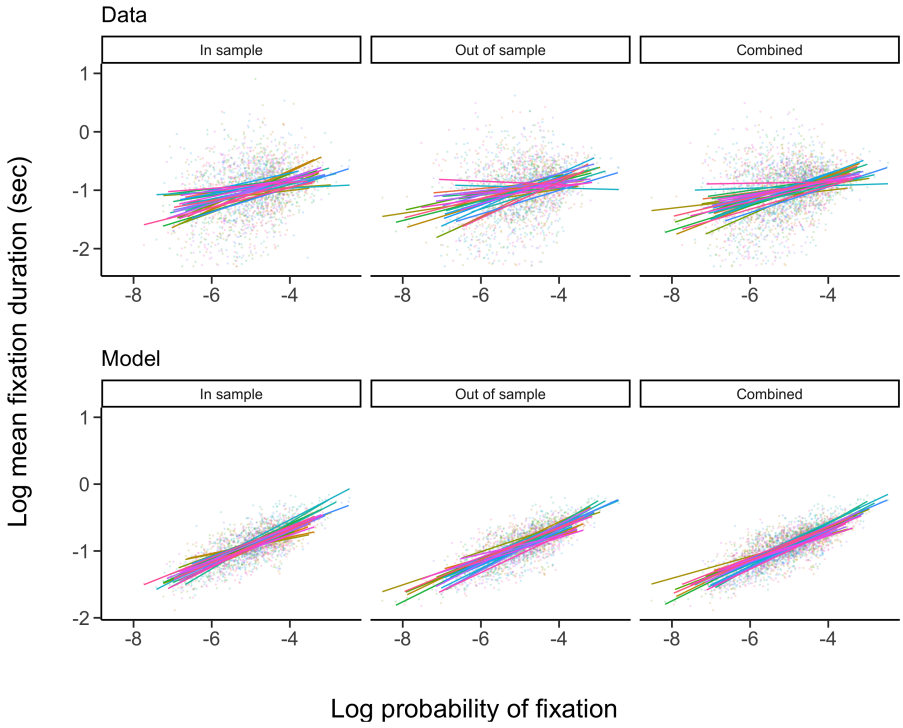


Figure 2.15: The correlations between the log of the predicted probability of fixation and the log of the mean fixation duration of cells in the grid. Superimposed are the regression lines per each image separately. Regression lines are superimposed to highlight the variability between items. The top panel shows the relationship in the data, the bottom panel shows the relationship reproduced by the model.

positive relation between probability of fixating a particular location and the mean fixation duration at that location. Arguably the relationship is stronger in the model than in the data (see Figure 2.15). However, one needs to keep in mind that the mean observed fixation durations are noisy because they are often calculated from only a couple of fixations inside a particular cell of the grid. Whether this is a sufficient explanation or there is additional model misspecification that causes this discrepancy is a potentially interesting avenue for future research. Taken together, these results demonstrate that the model is able to capture some global dependency of fixation durations on the attractivity of lo-

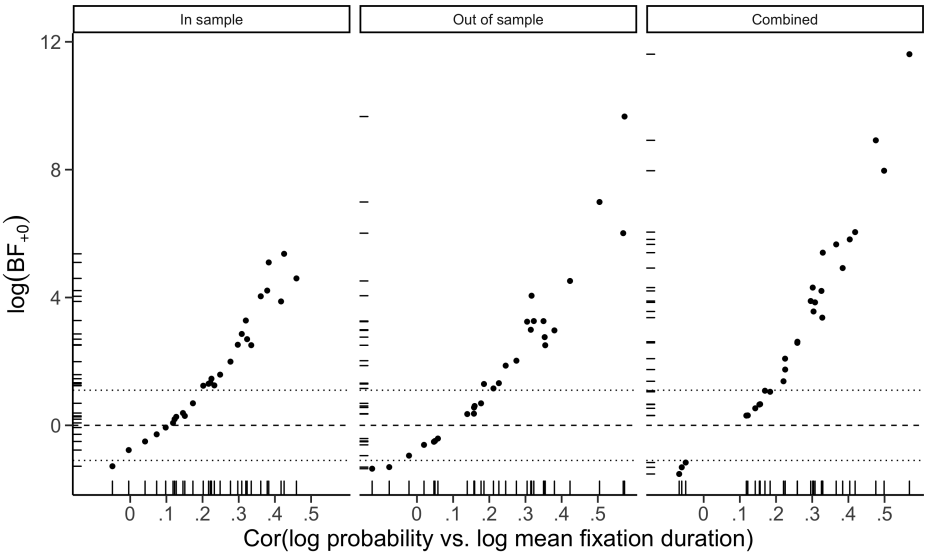


Figure 2.16: The correlations between the log of the predicted probability of fixation and the log of the mean fixation duration of cells in the grid, for each image separately on the x-axis and the corresponding log Bayes factors testing the null model (no correlation) vs the alternative (positive correlation, specified by stretched Beta(10,10) truncated at zero). The dotted lines shows the region of "anectdotal" evidence (Bayes factor between 1/3 and 3), i.e., that there is not enough information to say anything meaningful about presence or absence of the correlation.

cation of the image.

2.6 Conclusion & Discussion

This article presents arguments that theoretically grounded statistical models are important for validating predictions of the emerging theoretical framework against observed phenomena as well as detecting new empirical phenomena to be explained by said theory. Our model is grounded in the theoretical assumptions that can be verbally summarised as follows: 1) fixation durations depend on observers harvesting information from the stimulus, which is a noisy accumulation process, 2) saccades are launched when the observer concludes that staying at the current location is no longer advantageous over moving to an-

other location, 3) picking a new location depends on an internal "intensity" map over the stimulus, which is a combination of different "factors", such as information on the screen or for example systematic tendencies that highlight certain locations in contrast to others, and 4) observers harvest information from the relative proximity of the center of gaze, subjected to the limitations of their visual acuity — an assumption that provides the link between fixation durations and fixation locations. We consider these the core theoretical assumptions of the model. From this listing of assumptions, it should be evident that we are relatively more vague on the mechanism behind selection of the location of the next saccade. This is because the model offers flexibility by making use of "factors" that influence this selection, and because these "factors" on their own can represent different theoretical viewpoints. For example, original saliency models, such as that of Itti et al. (1998), can be considered a theory of fixation selection of itself, as it describes the rise of saliency map as neurons firing according to surround-background differences in image intensity, contrast of colors, and orientation of edges. We think this is a strength of our model as it allows to "plug-in" different explanations of the data without having to heavily modify the model.

In this article, we developed a model to analyse eye movement data by specifying a joint probability distribution of the fixation duration and fixation locations. To our knowledge, this is the first attempt to model fixation durations and fixation locations by defining a joint likelihood function of these two random variables. Using Bayesian inference, we were able to fit and extend the model such that the predicted patterns of the fixation durations and fixation locations align very closely with those of the observed data. Drawing upon the strengths of specifying models using likelihood functions (Schütt et al., 2017), we demonstrated how to diagnose, improve, and compare models so that they capture phenomena of interest present in real data. An example application showed that adding horizontal bias to the model improved the model's ability to capture the distribution of saccade angles.

The advantage of this approach is that it is possible for the model to be fitted to data (given that it is a statistical model with a likelihood function) and used to generate new data that can be contrasted with observed phenomena (such as distribution of saccade angles). In case the model does not perform

well in capturing these phenomena, it can be iteratively modified or improved until the model does so, or is ultimately rejected. Further, the model provides a relatively straightforward interpretation of the model parameters, facilitating the inference and possibly theorizing about the underlying mechanisms.

In our application, the results suggested that the most important factors determining eye movement behavior are the locations of objects on a scene, immediately followed by the tendency to make repeated fixations in a location nearby the current fixation. Salience and central bias had lower importance, and horizontal bias the least, although all factors made a significant contribution to fitting the data.

With respect to the central and horizontal biases, there is an ongoing debate on what is exactly the cause of these phenomena. Possible explanations range from being caused entirely by the image content (e.g., objects mostly aligned in the center of the images, or objects mostly aligned along horizontal axes or the horizon), to being some sort of interaction between image content and systematic bias towards centers or horizontal saccades, to being completely explained by systematic tendencies, caused by physiological, learned or strategic aspects (Foulsham et al., 2018, 2013; Le Meur & Liu, 2015; Tatler & Vincent, 2008; Tseng et al., 2009; van Renswoude et al., 2016, 2019). It is also possible that these three sources of the observed "biases" are not mutually exclusive. The model would be able to generate some central and horizontal bias with only the objects and saliency factors, representing the first category. In our model, we ended up using additional central bias and horizontal bias factors that were modeled as completely independent of the image content, hence representing the third category (independent of image content). Including these additional factors improved the fit of the model above factors that encode the image content, lending some credit to the third type of explanations. However, apart from the need of replicating this finding on other data sets, one needs to also implement central and horizontal biases that explicitly interact with the image content. Then, it will be possible to test all these explanations against each other.

In this article, our focus was mainly on free scene viewing, and so was our example. We hope and believe that the current model can be adapted to different contexts as well, as is it so easily modified that it can include different factors

or possibly various terms that accommodate various experimental designs or research questions. For example, it should be possible to use the model to compare demographic (e.g., adults versus infants) or experimental groups (e.g., free viewing instructions versus visual search instructions), providing alternatives to already established analytic methods (e.g., Coutrot, Hsiao, & Chan, 2018), or even adopt the model to specific purposes – such as strategic influences on eye movements in cognitive tasks (e.g., Kucharský et al., 2020) or economic games (e.g., Polonio et al., 2015).

This article followed an unusual strategy in model comparison, and that leads to some considerations regarding generalizability of our findings. The strategy of splitting the data set in two parts allows us to assess the adequacy of the models to describe patterns in the data that were not used for fitting the model. Thus, cross-validation procedures (for e.g., model comparison) are possible. However, the splitting procedure ensured that for each trial, whether or not included in the train or test sets, the model has some information about the participant (from other trials done by that participant) and the stimulus (from other participants on that stimulus) in that trial. Thus, the data in the test set cannot be considered completely *out of sample* in the traditional sense, which is one of the requirements for generalizability (next to additional assumptions). The cross-validation still gives us information about over-fitting, but does not aim for generalizing to a new population. This means that if we talk about one model fitting better than another model, we mean it is better at capturing patterns in the current sample of participants looking at the current sample of stimuli.

2.6.1 Extensions and Future Directions

Although the final model fits relatively well, there are plenty of ways to make it better in the future. For example, previous research suggested that the central bias is slightly more stretched in the horizontal compared to the vertical dimension (Clarke & Tatler, 2014; Tatler, 2007). In our application, we hold the width of the central bias in two dimensions equal. This could have created a slight misfit of the central bias factor, and could also underestimate the model's ability to produce saccade angles in the horizontal direction. Further, we hold the widths of factors in the model constant wherever possible to make

the simplest model we could apply do the data, and so this limitation can relate also to the exploitation factor and the attention window (both of which we assume is spherical). We also modelled all factors as independent normal distributions. In general, this assumption is not strictly necessary, and could be relaxed by specifying the components as bivariate normal, i.e., estimating their correlations. Luckily, these issues can be solved easily in case the data indicate to do so.

Potential model misspecification could also arise from modeling the horizontal bias. It has been shown that von Mises distribution is not necessarily optimal for describing the distribution of the saccade angles, due to the fact that the real distributions of saccade angles are typically more peaked than what von Mises distribution allows (K. Mulder, Klugkist, van Renswoude, & Visser, 2020). We used the von Mises distribution because it is relatively well known and can be fitted easily in Stan, whereas alternative distributions — such as the power Batchelet distribution as proposed by K. Mulder et al. (2020) — would make the implementation much more complicated. A second potential misspecification of the horizontal bias could be that the current implementation assumes that at any point in time, it is equally likely to make a saccade to the left direction as to right direction. However, this is likely not true, as intuitively we could think that having a fixation very close to a left border of the scene would lead to a rightward saccade with a very high probability (Clarke, Stainer, Tatler, & Hunt, 2017). This assumption could have underestimated the weight of the horizontal bias contribution compared to the other factors.

Additional model misspecifications could arise from modeling many parameters as fixed across participants and stimuli. In the current model, we only modelled the most obvious source of individual differences — the width of the attention window and the decision boundary — as random between participants. Importantly, these parameters only affect fixation durations, therefore the current model cannot capture individual differences in selecting fixation locations. However, it is probable that to better account for the patterns in the data (and to justify generalizibility to a population of observers and a population of stimuli; Yarkoni, 2019), we would need to model many of the currently fixed parameters as random. For example, it is desirable to assume that participants can differ in the weights of the different factors, or that the importance of

different factors are different even in different stimuli. Being able to generalize beyond the current data set was however not the focus of this article. However, the model is relatively flexible and including parameters as random should be possible in future applications. However, as explained above, the aim of this article is not a generalization per se, for this reason the current model is perhaps not unreasonably simplistic. Additionally, allowing the model to capture random effects does not automatically ensure generalizability, and additional effort and checks need to be put in place when designing the experiment.

The current implementation of the model does not capture differences between images. This means that one is not able to investigate the effect of the differences of total saliency or other factors, such as the number of objects on the screen, between different images on the eye movement behavior. This is due to the constraint that 1) the factor weights sum to 1 and are set equal between images, 2) observed factors (such as saliency) are normalized before entering the model. It is possible to either relax these constraints or add more parameters accounting for the differences between images, which is a candidate for future extensions. However, careful development and validation of different approaches should precede the application. As such, it is not presented in this article.

It is possible that the proposed mechanism underlying the model's architecture will need to be adapted in the future. For example, our assumption is that observers linger on a current fixation for the time it takes to decide to move to another location. It is possible that a different mechanism drives fixation durations. We also assume that the time it takes to select a new fixation location, and plan and execute the saccade is zero, that observers plan new target fixations only one step ahead (ignoring pre-planned saccades), or that once a decision to make a saccade is made, there is no stopping in launching it — assumptions that were relaxed in different modeling approaches (Nuthmann, 2017; Nuthmann et al., 2010; Trukenbrod & Engbert, 2014). We also assume that different factors combine in an additive manner (and can only increase the intensity), which may not be a realistic assumption (Barthelmé et al., 2013) – for example, a typical factor that is plausibly affecting fixation locations is the inhibition of return, which inhibits intensity of locations that were already visited (Klein, 2000). We believe that such alternative conceptual ideas could be

contrasted with the current model by developing new mathematical and statistical models that concretely implement these. Having specific models that are derived from concrete theoretical assumptions will hopefully facilitate our understanding of the real generative mechanisms (Borsboom et al., 2020; Schütt et al., 2017) that are relevant in eye-movement research.

We believe that similar attempts to modelling eye movements can influence both experimental practice as well as the theoretical advancements in the eye-tracking research. We made our code available online (`github.com/Kucharssim/WALD-EM`), along with additional materials that provide details about building and applying the model, so that other researchers can seek inspiration and help, if they wish to use our ideas for furthering their own work. Additional work should be done on the front of model validation through more extensive simulation studies. We hope that the current model will eventually be superseded by a better one — which would be a good sign of a healthy progress of our scientific understanding of visual perception. In the meantime, we hope that the proposed model will spark interest in applied and theoretical research of eye movements and provide valuable insights.

Open Practices Statement

The data, code and other materials are publicly available at the project's public repository: `github.com/Kucharssim/WALD-EM`.

The analyses in this manuscript are purely exploratory; all data were collected prior to the analysis. None of the analyses were preregistered. All modeling decisions were made after we have collected the data, but we did in fact split the data set in a training and test set before any modeling exercises.

Sometimes maybe good, sometimes maybe shit.

–Gennaro Gattuso

Chapter 3

Cognitive Strategies Revealed by Clustering Eye Movement Transitions

This chapter is published as Kucharský, Š., Visser, I., Truțescu, G.-O., Laurence, P. G., Zaharieva, M., and Raijmakers, M. E. (2020). Cognitive strategies revealed by clustering eye movement transitions. *Journal of Eye Movement Research*, *13*(1). doi: 10.16910/jemr.13.1.1

Abstract

In cognitive tasks, solvers can adopt different strategies to process information which may lead to different response behavior. These strategies might elicit different eye movement patterns which can thus provide substantial information about the strategy a person uses. However, these strategies are usually hidden and need to be inferred from the data. After an overview of existing techniques which use eye movement data for the identification of latent cognitive strategies, we present a relatively easy to apply unsupervised method to cluster eye movement recordings to detect groups of different solution processes that are applied in solving the task. We test the method's performance using simulations and demonstrate its use on two examples of empirical data. Our analyses are in line with presence of different solving strategies in a Mastermind game, and suggest new insights to strategic patterns in solving Progressive matrices tasks.

3.1 Introduction

Traditionally, response behavior such as accuracy and more recently response time are typically used to make inferences about participants' cognitive states, processes, or abilities to solve cognitive tasks (Groner & Groner, 1982; van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). Eye movements are a valuable source of information which extends our ability to make this kind of inference (e.g., Findlay & Gilchrist, 2003).

However, analyzing eye-tracking data is a challenging problem especially when cognitive strategies are to be inferred from the locations at which participants look and the order in which they look at them. Analyzing the information about the spatial and temporal dimensions of eye movements is commonly referred to as scanpath analysis, where the term scanpath concerns the spatio-temporal sequence of fixations and saccades, a term coined by Noton and Stark (1971).

Pioneering work of Yarbus (1967) showed (among other discoveries, Tatler, Wade, Kwan, Findlay, & Velichkovsky, 2010) that giving different instructions to observers changes their gaze behavior. This inspired the eye-tracking research community to devote its attention towards the so called inverse Yarbus problem (Greene, Liu, & Wolfe, 2012; Haji-Abolhassani & Clark, 2014): in this area of research, the question is whether it is possible to infer a task or a strategy

from eye movement patterns rather than whether a task or strategy invokes different gaze behavior. Most of the applications to investigate the inverse Yarbus problem deal with situations where we know what task or strategy a participant uses, e.g., using an experimental manipulation or recruitment based on diagnosis or a cognitive development stage. This enables researchers to use supervised techniques to show that some form of eye-tracking data representation can be used to describe the strategy of the observed groups. The representations range from similarity measures based on string edit and sequence methods (Cristino, Mathôt, Theeuwes, & Gilchrist, 2010; Glady, Thibaut, & French, 2013; Kübler, Rothe, Schiefer, Rosenstiel, & Kasneci, 2017; von der Malsburg & Vasishth, 2011), classifying raw eye tracking statistics (Boisvert & Bruce, 2016; Greene et al., 2012; J. M. Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013; Hild, Voit, Kühnle, & Beyerer, 2018; Kanan, Ray, Bseiso, Hsiao, & Cottrell, 2014), Markov models (Groner & Groner, 1982; Groner, Walder, & Groner, 1984) or hidden Markov models (Coutrot et al., 2018; Haji-Abolhassani & Clark, 2014; Kit & Sullivan, 2016; Y. Liu et al., 2009). For a review of different approaches to predict a task from eye movements see Boisvert and Bruce (2016). However, the question of how to infer a task or a strategy is a topical issue especially when it is unobserved (i.e., latent) and has to be inferred from the eye movements alone (i.e., a latent inverse Yarbus problem). This is precisely the issue of the current study.

3.1.1 Discovering latent groups

Latent groups are of interest whenever there is a reasonable expectation that observers might use a set of qualitatively different approaches to the task, and these differences would manifest through their gaze behavior, but it is unknown which observer uses which approach on which stimuli, other than what can be inferred from the eye movements themselves. This distinguishes the latent group problem from the prediction problem. In the prediction of a task, one has the information about the groups of observers which are supposed to qualitatively differ in the eye movement patterns and needs to learn which features of eye movements discriminate between these groups. In the latent group problem, one has to learn about the presence or absence of qualitatively distinct groups, and identify the features of the eye movements that are characteristic of

these groups. The discussion of latent groups manifesting through eye movements appears in the context of cognitive tasks (Glady et al., 2013; Hayes, Petrov, & Sederberg, 2011; Hayes et al., 2015; Loesche, Wiley, & Hasselhorn, 2015; Vigneau, Caissie, & Bors, 2006), decision making (Polonio & Coricelli, 2018; Stewart, Gächter, Noguchi, & Mullett, 2016), visual search tasks (Crosby & Peterson, 1991), face recognition and exploration (Chuk, Chan, & Hsiao, 2014, 2017; Chuk, Crookes, Hayward, Chan, & Hsiao, 2017; Coutrot, Binetti, Harrison, Mareschal, & Johnston, 2016), reading (Meseguer, Carreiras, & Clifton, 2002; von der Malsburg & Vasishth, 2011), and various other topics (Hayes & Henderson, 2017; Y. Liu et al., 2009; West, Haake, Rozanski, & Karn, 2006).

In the context of cognitive tasks, the detection of qualitatively distinct groups of eye movements can be especially informative, because these groups might be related to a cognitive strategy a person uses to solve the problem at hand (Bethell-Fox, Lohman, & Snow, 1984; P. A. Carpenter, Just, & Shell, 1990). Using the eye-tracking patterns to identify these strategies can bring additional insights as to how people solve these problems and can thus complement more conventional analyses of response behavior (Gierasimczuk, van der Maas, & Raijmakers, 2013; Steingroever, Jepma, Lee, Jansen, & Huizenga, 2019; van der Maas & Straatemeier, 2008).

Detecting latent groups from eye-movements can be viewed similarly as detecting latent groups from response behavior (Steingroever et al., 2019; van der Maas & Straatemeier, 2008), with the only difference being the type of data that are used as an input. Generally, the goal of detecting strategies can be achieved by unsupervised clustering methods or mixture modeling of the eye movement data. Unsupervised methods for clustering similar eye movement patters has already been used in context of face recognition (clustering hidden Markov models, e.g. Chuk et al., 2014; Chuk, Chan, & Hsiao, 2017; Chuk, Crookes, et al., 2017), reading (latent profile modelling based on scanpath similarity measure, von der Malsburg & Vasishth, 2011), free viewing (hierarchical clustering based on similarity measure, West et al., 2006), visual search (manual classification, Crosby & Peterson, 1991), or useability testing (Goldberg & Helfman, 2010a, 2010b), among others.

In the context of cognitive tasks hypotheses about latent solving strategies are currently not always tested using latent group analyses. In many cases, based

on theoretical expectations on how the latent strategies should manifest, researchers first define aggregate statistics from the eye movements (e.g., number of transitions, frequency of transitions between different areas of interest, etc.). Then they relate them to performance, thereby showing that different strategies result in different eye-movement statistics that are subsequently correlated with performance in the task at hand (Laurence, Mecca, Serpa, Martin, & Macedo, 2018; Loesche et al., 2015; Vakil & Lifshitz-Zehavi, 2012; Vigneau et al., 2006), although an alternative approach has been proposed to model the eye-tracking data (Successor Representation Scanpath Analysis (SRSA), Hayes et al., 2011, 2015). In short, SRSA builds successor representation matrices which contain information about the higher order transition dependencies in the data, which are then reduced in smaller number of dimensions and used as predictors of performance in the task. By adjusting parameters that control the specifications of these matrices, the method searches for a solution which maximizes the prediction of the task performance (i.e., a semi-supervised approach whereby task performance substitutes an indicator of the strategy, Hayes et al., 2011, 2015). It is often the case that the relationship between the latent strategy and the task performance (or other variable) is itself an empirical question. In this situation, conducting an unsupervised latent group analysis first will enable us to separate two questions from each other – first, whether we can detect qualitatively different eye movement patterns, and second, whether these patterns relate to performance (or other variables or interest). Crucially, this approach allows discovering groups that are not necessarily related to performance, and thus provides an opportunity to assess the latter question empirically. This distinction is important when the sole predictive performance is not of such an importance compared to assessing theories about qualitatively different cognitive processes, and to explain, rather than predict, individual differences (for in depth discussion of the trade-off between prediction and explanation, see Yarkoni & Westfall, 2017).

3.1.2 Eye movement representation

To conduct a latent group analysis, a choice needs to be made how to represent the eye movements data (in terms of its spatial and temporal features) to serve the purpose of finding the latent groups in the specific context. The need

to choose between different representations arises due to the fact that the raw eye-tracking data are, in their totality, too complicated (and perhaps noisy) to provide meaningful insights into the phenomenon under investigation. Thus, researchers usually need to define which features of the data are meaningful or discriminatory for the specific application and model them as such. For example, many authors emphasize individual differences in the processing of facial features, resulting in a distinction between holistic and analytic strategies in face recognition (Chuk et al., 2014; Chuk, Chan, & Hsiao, 2017; Chuk, Crookes, et al., 2017; Groner et al., 1984). Thus, hidden Markov models are suitable for this purpose as they allow to identify the important parts of the stimulus in a bottom-up manner. Furthermore, the transition patterns of the hidden Markov model between the facial features enables to discriminate between left-eye biased and right-eye biased analytic patterns. Based on a careful consideration of the specifics of eye movements in reading, von der Malsburg and Vasishth (2011) use their own similarity measure which does not require discretization of the stimulus into regions of interest and can take into account the fixation duration, which is important in the context of syntactic analysis of sentences. Another approach (West et al., 2006) relies on string edit distances (Levenshtein, 1966; Needleman & Wunsch, 1970; Waterman, 1981) to cluster sequences based on similarities between pairs of eye movement recordings.

In case the stimuli can be unambiguously divided into distinct meaningful areas of interest and the number, shape and position of these areas is assumed to be constant between the latent groups (as is the case in many cognitive tasks, e.g., Polonio & Coricelli, 2018; Trutescu & Raijmakers, 2016; Vigneau et al., 2006), a promising candidate for such representation is a transition matrix between pre-defined areas of interest, in which we quantify the probability of the next fixation on any area of interest conditionally on the position of the current fixation. Constructing or fitting transition matrices is relatively well established in the eye-tracking literature, either as descriptive statistic of the transition patterns (Althoff & Cohen, 1999; Ellis & Stark, 1986; Ponsoda, Scott, & Findlay, 1995) or as an integral set of parameters specifying (hidden) Markov models (see Coutrot et al., 2018; Visser, 2011, and references therein). Compared to the hidden Markov models, constructing transition matrices from the fixated areas of interest significantly reduces the complexity of the analysis at the expense of

binning fixations into pre-defined areas of interest instead of treating them as hidden states that need to be estimated from the data. This essentially simplifies the problem into a representation of categorical time-series (Pamminger & Frühwirth-Schnatter, 2010), without the need for a complicated evaluation of the likelihood of each eye movement sequence as a whole as is the case in hidden Markov models. It is important to note that if the areas of interest cannot be defined in advance (e.g., because their position is itself a topic of empirical investigation), the data are too noisy relative to the sizes of the areas of interest, or if there are too many bordeline fixations, this simplification may not be justified and other, more complex, approaches (e.g., hidden Markov Models) may be necessary.

Despite the fact that by using only the first-order transition matrices one potentially ignores informative features of the eye movements data (Hayes et al., 2011; von der Malsburg & Vasishth, 2011), they can still provide rich information about the transition patterns between the areas of interest, patterns which in many cases should be different between solution strategies that participants apply in cognitive tasks. Using transition matrices should be, in some cases (as we show later), sufficient to detect latent groups, while providing rich description of the characteristic features of the transition patterns that define these groups.

3.1.3 Goals & outline

Following the arguments in the previous sections, we believe that a method for detecting latent groups from eye movement data would be informative to investigate the existence of different solution strategies in cognitive tasks, and eventually also their relation to task performance. Such a method should generally meet the following desiderata. First, the eye movement patterns should be analyzed (summarised) such that the features of the hypothetical strategies can be detected. Second, the method should be unsupervised to allow detecting latent groups, even if they do not relate to external variables. Third, it should be possible to use some selection method for the number of such groups. Fourth, the classification of an eye-movement pattern into a latent group should be possible on an individual item basis to allow the possibility that participants switch between strategies during the task (for example, due to learning).

This article demonstrates the use of transition matrices as a representation of eye movements in order to detect latent groups of similar eye movement transition patterns. Specifically, we use a relatively easy to apply unsupervised method to discover latent groups, and present ways in which the classifications can be used in further analyses.

The structure of the article is as follows. In the next section, we provide details about how to construct transition matrices, and present the method we use for their clustering. Next, using simulations, we show that this method is able to retrieve the groups corresponding to strategies for solving a Mastermind Game and present an application of the method to real data. Then, we apply the method to a data set of the Wiener Matrizen Test (a test very similar to the Raven's Progressive Matrices; Laurence et al., 2018). We conclude with discussion of our findings, as well as with the limitations, alternatives, and extensions to our approach.

3.2 Clustering transition matrices

Our goal is to introduce a method that can be used for unsupervised clustering of eye movement sequences. Our approach is the following. First, we process the fixation coordinates into pre-specified areas of interest (AOIs). Such approach is typical in eye-tracking literature, at least in tasks with clearly distinct meaningful parts of the stimulus, although there is some discussion on how to optimally choose and delineate AOIs (e.g., R. S. Hessels, Kemner, van den Boomen, & Hooge, 2016).

From each individual sequence of the fixated AOIs, we create the transition probability matrix, where each row corresponds to a "sender" AOI, and each column to the "receiver" AOI. Each row of the matrix is computed by counting the number of transitions from the sender AOIs to all other AOIs and dividing the row by the sum of the total transitions from that AOI. The entries of the $d \times d$ transition probability matrix M can be interpreted as follows: Given that a fixation is on the i^{th} AOI, the probability that the next fixation is on the j^{th} AOI is equal to M_{ij}.

All transition matrices are then reshaped into vectors of length d^2 and stored in a data matrix where the rows correspond to the individual matrices (i.e., rep-

resentations of the individual AOI sequences), and the columns correspond to the cells in the transition probability matrices. The resulting data matrix is then subjected to the standard k-means clustering algorithm (Hartigan & Hartigan, 1975).

As an unsupervised method, k-means provides us with an opportunity to find distinct groups of eye movement patterns, patterns which differ in their transition matrices. Scree plots can be used to diagnose solutions for different numbers of clusters. Furthermore, each cluster is assigned a mean transition matrix which identifies the characteristic features of the transition patterns in each groups, which can be used to interpret the groups and assign a label according to hypothesised cognitive processing. We use standard k-means based on minimizing the within-cluster sum of squared Euclidean distances from the centroids to ease of interpretation of the cluster centroids. Relative Euclidian distances of individual matrices to the cluster centers can be used to assess the representativeness of each eye movement recording of that particular group. The cluster assignments indicators can be used for further analyses, for example examining the relationship of the clusters to performance. The demonstration of our approach follows in the next two examples and a simulation study.

3.3 Application: Deductive Mastermind

Here, we present an example of detecting cognitive strategies in a Deductive Mastermind Game (DMM). In the DMM, the player is supposed to deduct a sequence of flowers based on multiple "conjectures" composed of a sequence of flowers and their corresponding feedback presented as a collection of colors next to the conjecture. The green feedback means that a flower in the conjecture belongs to the solution, red feedback means that a flower does not belong to the solution, and orange feedback means that a flower belongs to the correct solution but is on a wrong place in the sequence (Gierasimczuk et al., 2013).

The DMM was implemented as a part of web-based math and logic training system in primary schools called Math Garden (`rekentuin.nl` or `mathsgarden.com`). Gierasimczuk et al. (2013) analyzed data collected with Math Garden and revealed that the player ratings and the item difficulty have a tri- and bi-modal distributions, respectively. Logical analysis of the game

showed that the items can be solved using different strategies, ones that vary in the number of steps a player needs to deduct the correct solution. One possible explanation for the multimodality of the player ratings might be that the population of players is a mixture of people using different strategies, strategies which relate to the efficiency in solving the game.

The logical analysis predicts at least two strategies to occur during solving the items. The first strategy is characterized by scanning the feedback in the order in which it is presented (i.e., from top to bottom) – this prediction relies on the assumption that it is a natural (i.e., learned) way of processing information before internalizing the differences in information value that different feedback holds. The prediction of the second strategy, in contrast, relies on the fact that each row of stimuli can have different information value. Thus, this strategy would be characterized by selectively scanning the feedback starting from the conjectures which contain the most information and proceeding to those which complement it. Figure 3.1 shows one of the items with superimposed eye-tracking patterns under the two strategies. Notice that the first order transition matrix differs between the strategies. Thus, it should be possible to discriminate between them using only the first order transition patterns.

3.3.1 Methods

We use subset of the data that was collected with adults outside the educational system as part of a larger project (Trutescu & Raijmakers, 2016). The data are available at `osf.io/he43s/`. Twenty-six university students with normal or corrected-to-normal vision participated in the study. Two participants were excluded from the analysis due to missing data in the eye-tracking measurements. The study comprised of one learning block (13 items) and two test blocks (16 items each) in randomised order of the items within each block. The 2-pin items were constructed as an adapted version of the DMM task suitable for eye-tracking by adjusting the layout of the displayed conjectures (see Appendix in Trutescu & Raijmakers, 2016). The items in the learning phase were designed such that they are easily solvable regardless of the scanning strategy; items with all combinations of feedback were presented to give participants the opportunity to establish the difference between feedback types. For a concise presentation of the current method, we further analyse only four items (two from each

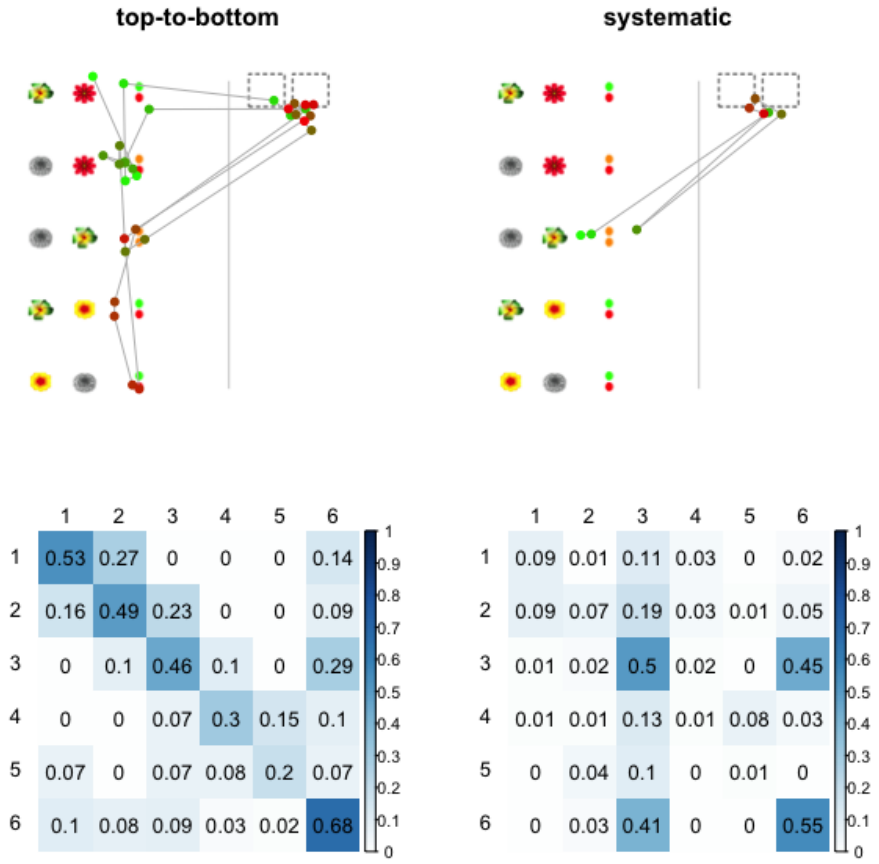


Figure 3.1: Synthetic data of two strategies in one of the selected Mastermind games. The left panel shows the top-to-bottom strategy, and the right panel shows the systematic strategy of selective processing. The top panel shows examples of the scanpaths, where dots correspond to fixations (the color gradually changes from bright green to dark red based on the order of the fixations) and lines connect the successive fixations (i.e., saccades). Bottom panel shows the transition matrices of the simulated strategies averaged over 1,000 simulations.

test block) containing orange-orange feedback at the third row. These items can be solved by focusing attention only on the third row, as the orange-orange feedback informs to swap the positions of the two presented flowers, see Figure 3.1.

The eye movements were recorded using EyeLink-1000 eye tracker with 500 Hz sample rate (SR Research Ltd., Ontario, Canada). Participants were seated at a desk with a chin rest about 55 cm in front of a 17-inch computer monitor, subtending an approximate $27° \times 34°$ visual angle. Before the data collection, a five-point calibration was used, which was repeated until the recorder point of gaze reached the best possible quality.

The raw data were parsed into fixations and saccades using Gazepath algorithm (van Renswoude et al., 2018). The identified fixations were classified into six rectangular areas of interest, one each for the row of the task in the left panel, and the sixth belonging to the response area on the right panel (see Figure 3.1). Thus, "scanpath" is in our case operationalized as a series of fixated AOIs. By doing so we artificially segregate the items into meaningful chunks of information: the units of information are the pairs of flowers and associated feedback, which corresponds to the definition of the conjectures in the logical analysis of the Mastermind game (Gierasimczuk et al., 2013). This approach imposes rather strong assumptions on the semantic connotation (i.e., the structure of the task as interpreted by individual participants) and could be prone to measurement noise (in applications where the AOIs are relatively small or close to each other); we nevertheless consider this a sensible approach in the current task as the rows of the stimuli were designed to be visually well separated and correspond to the semantic denotation of the task, which had been communicated through the experimenter's instructions and demonstrated during the learning block.

Transition matrix eye movement analysis

Before clustering the data, we conducted a simulation study in order to investigate the method's performance. We previously analyzed all scanpaths on the four selected items for which classification into the top-to-bottom and systematic strategy was possible by visual inspection. We wrote a simulation function which mimics the two strategies (top-to-bottom and systematic, shown in Fig-

ure 3.1). The simulated patterns were matched with real data with respect to several criteria (see osf.io/82fau/ and osf.io/bz3ny/). This enabled us to simulate an arbitrary number of participants using one or the other strategy with some variability in the patterns within the strategies (associated R code at osf.io/jxwrk/).

We tested the method's performance with respect to two varying features of the simulated data. For each of the features, we selected three values, resulting in a 3×3 simulation design. The features we varied and their values were the following:

1. Sample size: $n = (20, 60, 100)$. We varied the total number of the participants in the simulated studies.

2. Proportion of strategies: $p = (0.25, 0.5, 0.75)$. We varied the proportion of participants using one or another strategy. The value of p corresponds to the proportion of participants using the top-to-bottom strategy. This number was treated as a sample proportion (not population proportion) and thus there was no sampling variance between the simulations using the same value.

We simulated 600 data sets per each combination of parameters (totaling $3 \times 3 \times 600 = 5400$ simulated studies). In each simulation, each participant solved only one item. This allowed us to inspect the robustness of the method even for item-wise analysis (i.e., with relatively sparse data).

For each data set, the procedure was as follows. Each individual sequence of AOIs was converted to a 6×6 transition matrix. The individual matrices were reshaped into a vector of length $6 \times 6 = 36$ and stored into a $n \times 36$ matrix. The k-means clustering was applied to this matrix with solutions from 1 to 10 clusters to inspect whether the scree plot identifies the correct number of clusters (2).

Next, we assumed that the correct number groups was selected and investigated the classification accuracy of the two-cluster solution. We also investigated the stability and accuracy of the estimated cluster centers. To do this, we had to resolve an issue of label switching. In each simulation, we created a 2×2 confusion matrix of the true group membership against the estimated labels given by the k-means. If the sum of the diagonal entries in this matrix was

greater than the sum of the off-diagonal entries of the matrix, we kept the labels as they are. If this was not the case, the cluster indicators from k-means were relabeled.

After the simulations were conducted, we applied clustering of transition matrices to the real data. The subset of the whole data consists of $24 \times 4 = 96$ eye movement sequences for cluster analysis. For each sequence, we calculated the 6×6 transition probability matrix and reshaped it into a vector of length 36. The resulting 96×36 matrix was subjected to the k-means clustering.

3.3.2 Results

Simulations

Extracting the correct number of strategies

A rule of thumb for selecting the number of clusters is to inspect a scree plot to see at which point the amount of unexplained variance by the clusters stops decreasing rapidly. Given the subjective nature of this procedure, we cannot report exact numbers of the cases where the scree plot would identify the true number of latent groups (2) correctly. However, from a qualitative inspection of the scree plots, we saw that the classic "elbow" shape emerges mostly when 1) the sample size is large, and 2) when the sizes of groups are even. Decreasing the sample size results in mostly uninformative scree plots (i.e., the scree plot decreases gradually).

Classification accuracy

For all of the simulations, we inspected how accurate is the participant assignment using the solution with two latent groups.

Table 3.1 shows the median and interquantile range of the assignment accuracy for all combinations of p and n. Overall, the classification accuracy is high, in most scenarios higher than 90%. The total accuracy is the highest when the two strategies are equally represented in the sample, and slightly increases with the sample size. The accuracy of correctly classifying top-to-bottom pattern is slightly lower than the accuracy of classifying the systematic strategy, which might be due to the fact that the top-to-bottom pattern is more variable, and

p	n	Total	Systematic	Top-to-bottom
0.25	20	0.90 (0.75, 0.95)	0.93 (0.80, 1.00)	1.00 (0.60, 1.00)
	60	0.95 (0.90, 0.97)	0.96 (0.89, 0.98)	0.93 (0.87, 1.00)
	100	0.95 (0.92, 0.97)	0.96 (0.92, 0.99)	0.96 (0.92, 0.96)
0.50	20	0.95 (0.85, 1.00)	1.00 (0.90, 1.00)	0.90 (0.80, 1.00)
	60	0.95 (0.92, 0.97)	0.97 (0.93, 1.00)	0.93 (0.87, 0.97)
	100	0.94 (0.92, 0.96)	0.98 (0.96, 1.00)	0.92 (0.88, 0.96)
0.75	20	0.80 (0.70, 0.90)	1.00 (1.00, 1.00)	0.80 (0.67, 0.87)
	60	0.83 (0.73, 0.92)	1.00 (1.00, 1.00)	0.78 (0.64, 0.89)
	100	0.84 (0.75, 0.90)	1.00 (0.96, 1.00)	0.79 (0.68, 0.88)

Table 3.1: Median and interquantile range of classification accuracy based on the k-means clustering of transition matrices. Column Total denotes the proportion of correctly classified cases, Systematic denotes the proportion of correctly classified cases simulated under the systematic pattern, and Top-to-bottom denotes the proportion of correctly classified cases simulated under the top-to-bottom pattern. Column p denotes the proportion of the top-to-bottom patterns in the sample, n denotes the total sample size.

can also contain characteristic features of the systematic pattern (namely, transitions from the third row to the response).

The assignment accuracy means that if we wished to estimate the proportion of the strategies in the sample, we would estimate it correctly, except when the top-to-bottom pattern is dominant. In that case, a large portion of the true top-to-bottom patterns would be classified as systematic, leading to underestimation of the number of top-to-bottom patterns in the data, as can be seen in Table 3.2.

Stability of strategy representation

We also inspected whether the cluster centers are stable (i.e., show relatively similar transition matrices across simulations). Table 3.3 shows the median and interquantile range of the pairwise Pearson's correlations between the cluster centers. Overall, the correlations are quite high, suggesting that the representations of the transition matrices remains similar across the simulations. The

	p		
n	0.25	0.5	0.75
20	0.25 (0.20, 0.30)	0.45 (0.40, 0.50)	0.60 (0.50, 0.70)
60	0.27 (0.25, 0.30)	0.47 (0.45, 0.50)	0.60 (0.48, 0.68)
100	0.27 (0.25, 0.29)	0.48 (0.45, 0.50)	0.60 (0.51, 0.66)

Table 3.2: Median and interquantile range of the proportion of patterns classified as top-to-bottom. Columns p indicate the true proportion in the data, column n denotes the total sample size.

average cluster representations across all simulations is shown in Figure 3.2.

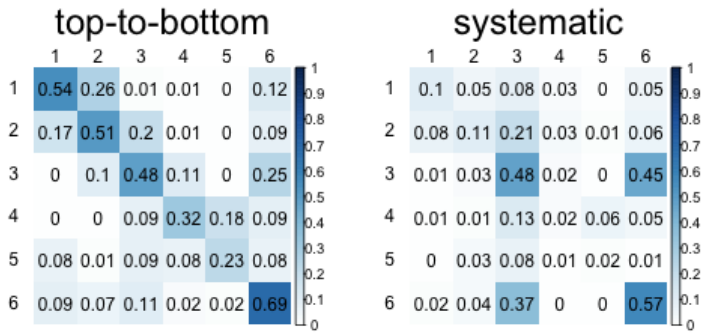


Figure 3.2: The average transition matrices identified by the k-means across all simulations.

Data

Here, we present the results of the k-means clustering applied to the real DMM data (Trutescu & Raijmakers, 2016) from four items. The R script is at `osf.io/g2yp4/`. The scree plot was uninformative as it did not show a clear "elbow" pattern. Thus, we inspected the agreement between the solutions spanning from two to four clusters.

Figure 3.3 shows the average transition matrices for the two, three and four

		cluster label	
p	n	systematic	top-to-bottom
0.25	20	0.91 (0.87, 0.94)	0.71 (0.44, 0.82)
	60	0.97 (0.96, 0.98)	0.91 (0.87, 0.94)
	100	0.98 (0.98, 0.99)	0.95 (0.93, 0.96)
0.50	20	0.88 (0.84, 0.91)	0.89 (0.84, 0.92)
	60	0.96 (0.94, 0.97)	0.96 (0.95, 0.97)
	100	0.97 (0.96, 0.98)	0.98 (0.97, 0.98)
0.75	20	0.77 (0.67, 0.84)	0.91 (0.86, 0.93)
	60	0.90 (0.85, 0.93)	0.96 (0.94, 0.98)
	100	0.93 (0.89, 0.96)	0.98 (0.96, 0.98)

Table 3.3: Median and interquantile range of the pair-wise Pearson's correlations of the cluster centers. Column p denotes the proportion of the top-to-bottom patterns in the sample, n denotes the total sample size.

cluster solutions, and the pair-wise confusion matrices of the cluster membership. Comparing the two and three clusters solution suggests that the cluster 1 from the two clusters model is almost perfectly separated in two clusters under the three clusters model (see confusion matrix in the first row and third column). In addition, the four clusters solution finds one additional sub-cluster which is characterized by transitions between conjectures 1-3, but does not proceed further (which could be explained by the the participant attempting to solve the item from top to bottom and terminating the process once the most informative feedback was found). Overall, these results suggest that the data are in line with the prediction of two general patterns – that of systematically searching for the most informative feedback, and that of attempting to solve the item in the order of conjectures as they are presented.

To check qualitatively whether the fixation sequences clustered in the groups correspond to the systematic and top-to-bottom patterns as described above, we also plot the most representative sequences for each of the clusters. We compute the "representativenness" of a sequence to a particular cluster as an Euclidian distance of the transition matrix of the sequence to that cluster center, relative to the sum of the Euclidian distances to all other clusters.

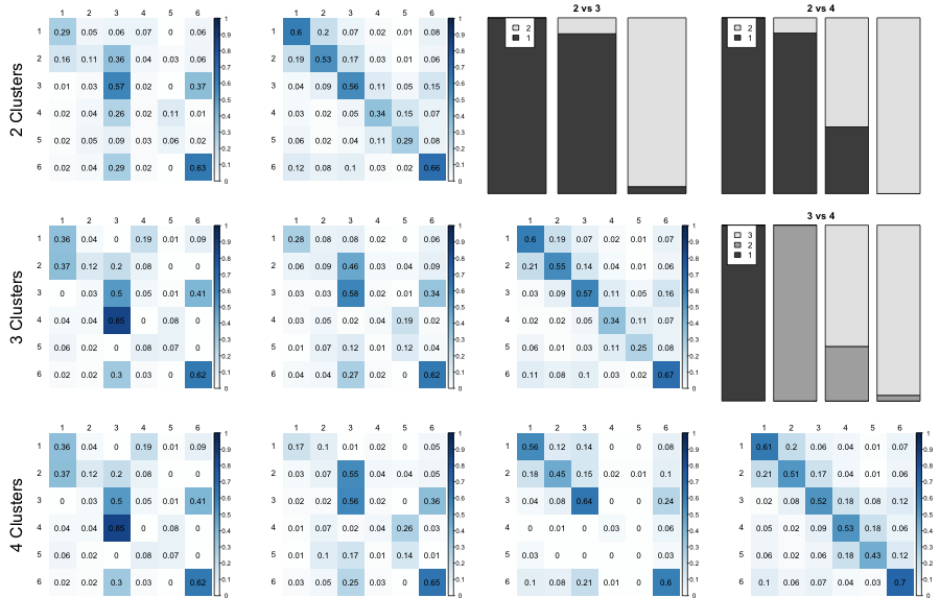


Figure 3.3: The cluster centers of the k-means solutions with two, three and four clusters, and the confusion counts of the different solutions. Each row corresponds to one solution (two, three, and four centers from top to bottom). The bars on top right of the Figure correspond to the overlap between cluster assignments. For example, panel 2 vs 3 indicates that most of the cases assigned to the first and second cluster in the three cluster solution were classified into the first cluster in the two cluster solution, whereas most of the cases classified into the third cluster with the three cluster solution were classified into the second cluster in the two cluster solution.

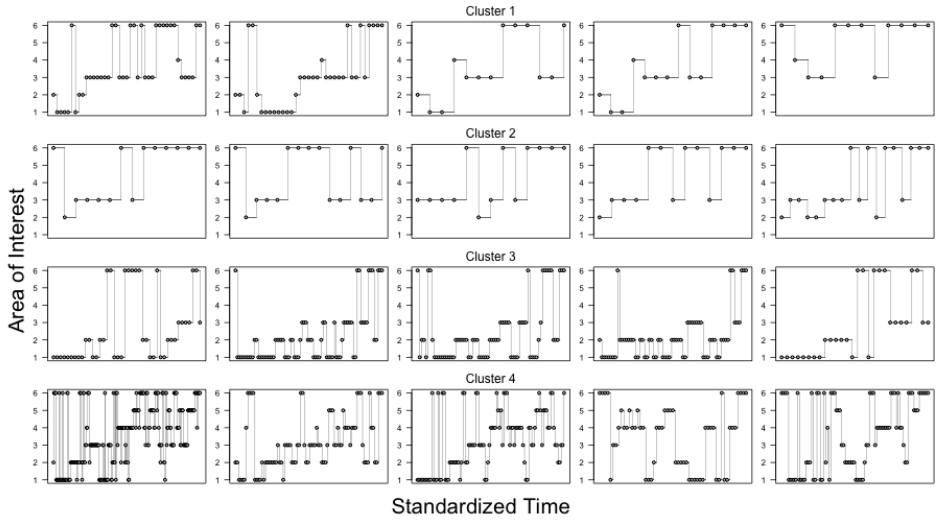


Figure 3.4: Each row shows five example scanpaths assigned to one of the clusters. Fixations to particular AOIs are shown as dots and transitions are connected with lines.

Figure 3.4 shows the fixation sequences, where the points show the individual fixations on particular AOIs (on the y-axis) as a function of time (time has been normalized to span between 0 and 1). Because there is a strong overlap between the cluster assignments between the 2-4 k-means solutions, we only show the representative fixation sequences grouped into four clusters. Clusters 1 and 2 are characterized by transitions between the third row and the response (AOIs number 3 and 6). Cluster 3 is characterized by a period of fixations on the first three rows, followed by transitions to the response. Cluster 4 is the most variable, having characteristic pattern of progression from the top to the bottom of the game with frequent transitions to the response in between. Overall, clusters 1 and 2 align with the predicted systematic patterns, whereas clusters 3 and 4 align with the top-to-bottom pattern. The distinction between the clusters 3 and 4 is that patterns in the cluster 3 usually terminate very quickly after fixating the third row (which contains sufficient information to deduce the correct solution), whereas patterns in cluster 4 do not seem to have this pattern.

The clustering suggested that the groups also differ in terms of the number of fixations. This would be consistent with the view that the current items are

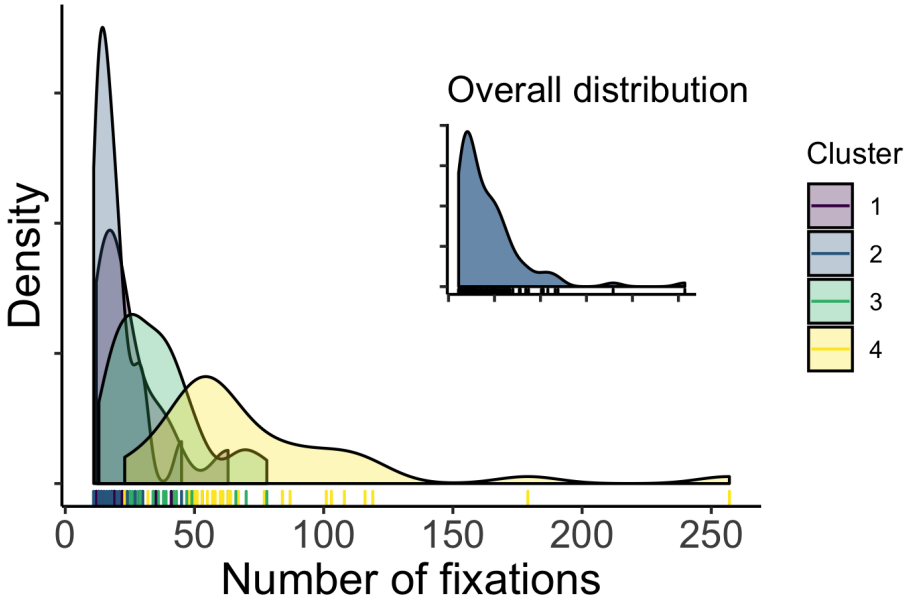


Figure 3.5: Distribution of fixations of the four clusters.

relatively easy to solve when using the systematic search (i.e., focusing on the most informative feedback).

Figure 3.5 shows the distribution of the number of fixations for each cluster, as well as the marginal distribution over all data. We conducted exploratory analyses by fitting the fixation counts with multilevel negative binomial model using R package `brms` (Bürkner, 2017, 2018) to see whether the apparent differences between the clusters are statistically supported (see `osf.io/87ahz/`). The results indicated that the cluster 4 has the highest number of fixations, the cluster 3 has the second highest, and the cluster 1 and cluster 2 are comparable, see Figure 3.6. However, trying to uncover the groups based on the fixation counts would be a hard task, judged by the apparent absence of multimodality of the overall distribution of fixation counts.

Lastly, we set out to investigate whether the clusters are associated with different probability of a correct answer, although the current set of items does not allow a lot of room for modelling on this part as the items are relatively easy (i.e., the percentage of correct answers is 88.5%). In particular, the first three clusters had perfect or near perfect performance on these items, whereas

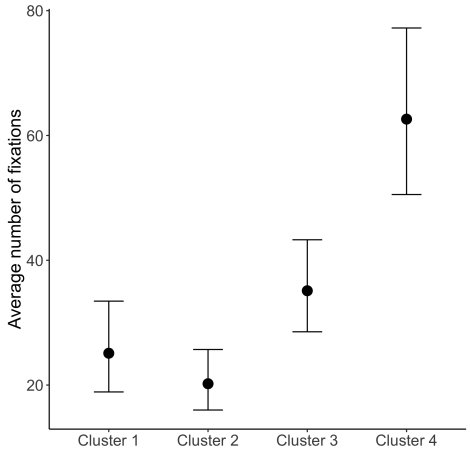


Figure 3.6: The average number of fixations of each group with 95% credible intervals.

	cluster			
	1	2	3	4
n	13	23	29	31
# correct	13	22	29	21
% correct	1	0.957	1	0.677

Table 3.4: Descriptives of the correct answers for each of the clusters.

only 68% of the patterns in the fourth cluster resulted in correct response, see Table 3.4.

To sum up, we were able to cluster the real DMM data using transition matrices; in accordance with the expectations, we found two general patterns – one that is characterized by a systematic selective scanning of the most informative feedback and another characterized by a search pattern starting at the top of the item, proceeding downwards. However, the clustering results were not entirely conclusive regarding the number of clusters and it is possible that more subpatterns are hidden (i.e., one that starts as the top to bottom pattern and switches once the most informative feedback is processed). The patterns differ in the lengths of the sequences, and the four clusters seem to have different probability of correct answers. Specifically, the first two clusters have high

chance of a correct answer as they appear to focus on the feedback which is sufficient to solve the item. The third cluster also has a high success rate, suggesting that it might be capturing the processes when a participant solves the item in non-systematic way (i.e., from top to bottom), but deduces the correct solution once arrived to the most informative feedback.

3.4 Application: Progressive Matrices

In the previous example, we have shown that classifying eye movement sequences using clustering of transition matrices is possible, even if the data are relatively sparse. To illustrate the use of clustering transition matrices in a different context, we present a reanalysis of data collected by Laurence et al. (2018). The data contain eye-tracking recordings of participants who solved Wiener-Matrizen Test 2 (WMT-2; Formann & Piswanger, 1979; Formann, Waldherr, & Piswanger, 2011). The WMT-2 is structurally similar to the Raven's Progressive Matrices (RPM), as both consist of a 3×3 matrix containing images with varying features, where the bottom-right item is missing, and a 2×4 response alternatives matrix. The goal of the task is to identify which item from the response alternatives matrix belongs to the missing part of the 3×3 matrix, such that the varying features complete a logically consistent pattern.

Vigneau et al. (2006) proposed that two distinct general strategies – constructive matching and response elimination (Bethell-Fox et al., 1984) – can be employed when solving the Raven's Progressive Matrices (RPM). The former is a systematic strategy of evaluating the matrices to deduce the only correct solution, which is then found in the response area. In contrast, response elimination is a strategy of successively considering different responses and evaluating whether they are consistent with the information given by the matrices or not. The two strategies should manifest through different eye movement patterns, as constructive matching would yield systematic transitions by rows (or columns), whereas the response elimination would show a pattern of frequent transitions from the matrix and the response area. Following the seminal work of Vigneau et al. (2006), numerous studies followed up the hypothesis to replicate its findings, using mostly summary statistics from the eye-tracking data (Laurence et al., 2018; Loesche et al., 2015; Vakil & Lifshitz-Zehavi, 2012). More

recently, a different approach has been applied for describing cognitive strategies taking into account higher order dependencies in the transition patterns (Hayes et al., 2011, 2015). Here, investigate this hypothesis using clustering of transition matrices.

3.4.1 Laurence et al. (2018) data

The data analyzed here were collected and reported previously by Laurence et al. (2018). The data are generated by 34 participants who solved 18 items (+3 practice items) from Wiener-Matrizen Test 2 (WMT-2; Formann et al., 2011). The data contain the responses (correct/incorrect) and the processed eye-tracking data: the fixations were classified into 10 AOIs (`osf.io/sgyk3/`). The areas 1–9 correspond to the individual cells in the matrix, starting from top-left entry, and filling the matrix row-wise (e.g., 1 – top-left; 3 – top-right; 7 – bottom-left, up to 9 – bottom-right). The area 10 is the response matrix area, containing all eight options for selecting the solution. The data is organized as ordered sequences of fixations on the areas of interest for each participant and each item. If a fixation did not fall into either of the designated areas of interest, we excluded that fixation from the data, which resulted in deleting 4338 fixations out of the total number of 91267, leading to a 95% inclusion rate.

3.4.2 Methods

Because it has been argued that in the context of Raven's Matrices, one should remove repeated fixations within one AOI, as the frequency of repeats is quite high (especially within the response matrix; Hayes et al., 2011), we use the clustering technique both on data where the repeated fixations were included (i.e., using the full data), as well as clustering data after removing the repeated fixations, essentially removing 35880 transitions (about 44.7%). Almost half of the excluded transitions (15185) were based on the repeated fixations within the response alternatives matrix.

The rest of the procedure was as follows. First, each fixation sequence was converted to a transition matrix. The $34 \times 18 = 612$ transition matrices were reshaped into vectors and stored in a 612×100 data matrix. This matrix was subjected to k-means clustering estimating 1 to 10 clusters to inspect the scree

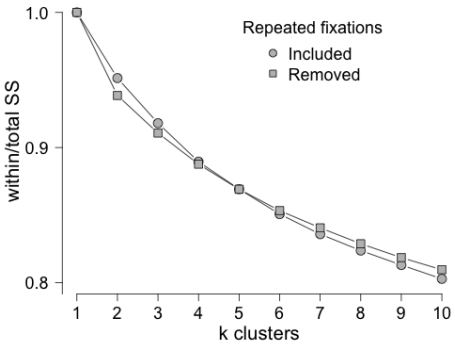


Figure 3.7: Scree plots from the k-means clustering for the data with repeated fixations (dots) and without repeated fixations (squares).

plots.

3.4.3 Results

The scree plots from the k-means on data with excluded repeated fixations provided a modest support for the presence of two groups, whereas the scree plot on the full data remained inconclusive, see Figure 3.7. In line with previous literature (Hayes et al., 2011, 2015), we further discuss the results based on the k-means solution with two clusters; solutions with higher numbers of clusters yielded qualitatively comparable results (see `osf.io/h3nc7/`).

On the full data, 271 (42 %) sequences were classified into the first cluster, whereas 293 (48 %) sequences were classified into the first cluster using the data without the repeated fixations. Overall, the agreement between the two classifications was high: 516 out of the total 612 sequences (84 %) were assigned into the same cluster regardless whether the repeated fixations were excluded or not. Figure 3.8 shows the mean transition matrices of the two clusters. The transition matrix of the first cluster suggests a similar pattern that has been previously described by Hayes et al. (2011), interpreted as the constructive matching strategy, indicating high probabilities of transitioning to left or right relative to the current fixation, which suggests a general pattern of inspecting the matrices within individual rows. However, the interpretation of the second cluster is less clear. First, the probabilities of transitioning left or right remain

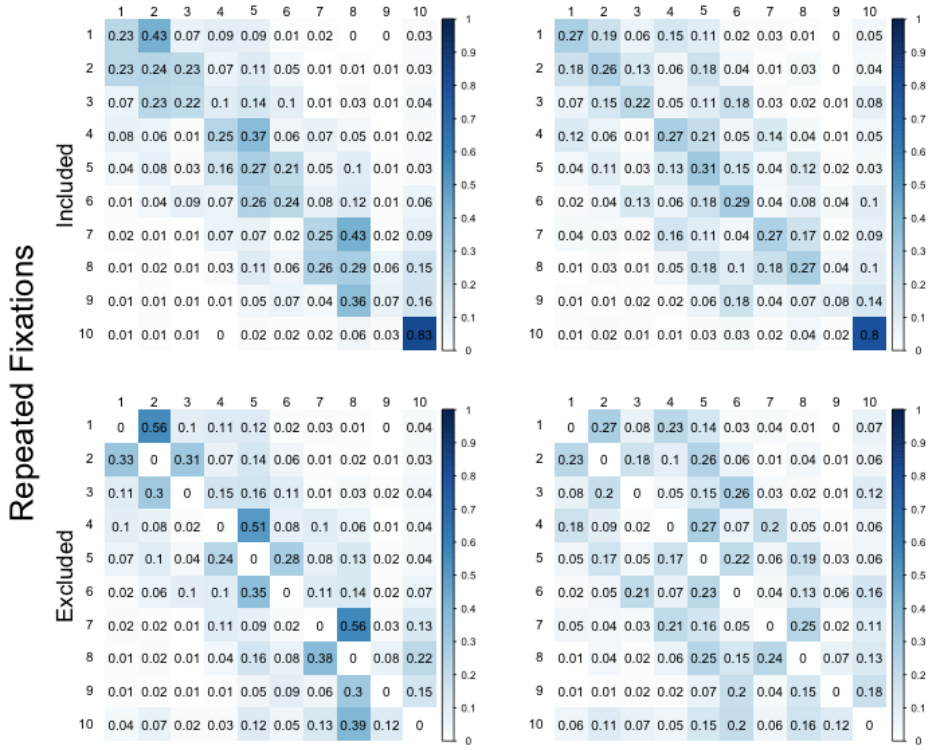


Figure 3.8: Average transition matrices of the two clusters. Left panel shows matrices of the first cluster with (top) and without (bottom) repeated fixations, right panel shows matrices of the second cluster with (top) and without (bottom) repeated fixations.

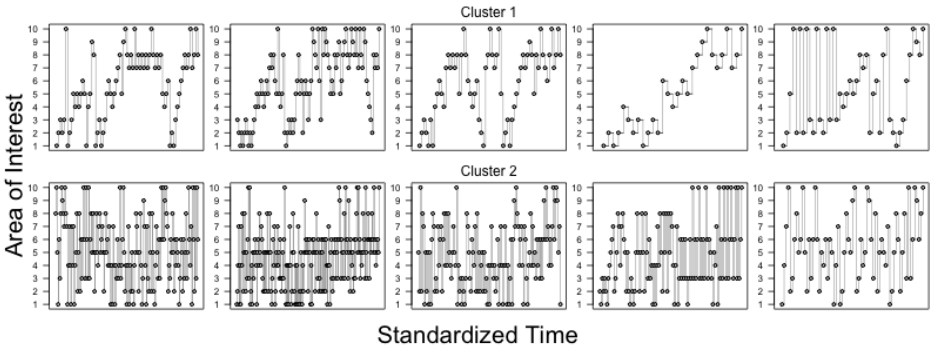


Figure 3.9: Top row shows five scanpaths that have been assigned to cluster 1, bottom row shows five scanpaths that have been assigned to cluster 2 by all three clustering methods. Repeated fixations are removed.

quite high, but there is also an increased probability to transition up or down, suggesting inspection of the matrices within columns. Second, under the expectation that the second cluster is related to the response elimination strategy, we would expect higher, and more uniformly distributed probabilities on column 10 (transition probabilities to the response area), but also in row 10 (transition probabilities from the response area). Although this is generally the case, the differences compared to the first cluster are rather small, which does not corroborate strongly that this cluster can be interpreted as the response elimination strategy.

Figure 3.9 shows examples of the scanpaths that have been assigned to one or the other cluster. The first cluster is characterized by frequent transitions from left to right within rows (i.e., $1 \rightarrow 2 \rightarrow 3$, etc), whereas the second cluster also shows frequent transitions within columns (i.e., $1 \rightarrow 4$, $2 \rightarrow 5$, etc).

The approach in the previous studies focusing on strategies in Progressive Matrices (Laurence et al., 2018; Loesche et al., 2015; Vakil & Lifshitz-Zehavi, 2012; Vigneau et al., 2006) is to inspect, for example the number of toggles (transitions between the matrix and the alternatives), or the rate of toggling (number of toggles divided by the response time). Here, we inspected whether the two uncovered clusters differ in the length of the sequences, number of toggles, or rate of toggling (in this case defined as the number of toggles divided by

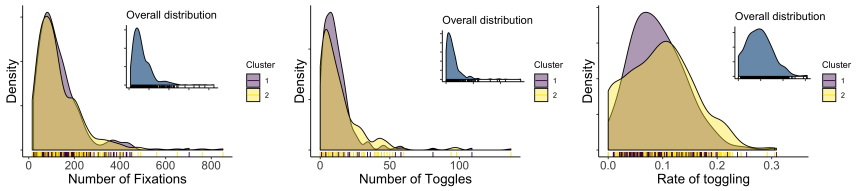


Figure 3.10: Distribution of the number of fixations (left), number of toggles (middle), and rate of toggling (right) of the two clusters.

the number of transitions). Figure 3.10 shows that the differences between the clusters are not very pronounced in either of these measures. We did not test the differences further. However, the results suggest that neither of the clusters relate to the hypothetical response elimination pattern.

Regardless of the interpretation of the clusters, we inspected their relation to performance. We fitted an exploratory multilevel logistic model using R package `brms` (Bürkner, 2017, 2018) predicting whether the answer was correct or incorrect with a fixed and random slope for clusters, random intercept for participants and items (see `osf.io/wvy23/`). The analyses revealed that the differences between the clusters vary substantially and the average effect is not very pronounced; the second cluster performed slightly better, but the results are inconclusive. Following the focus of the original article, we fitted an exploratory model which also takes into account item types (i.e., Rule Type items, Rule Direction items, and Graphical Component Nature items; Laurence et al., 2018) and their interactions with the clusters (see `osf.io/adt89/`). Figure 3.12 summarises the main results. On a descriptive level, the first cluster performs slightly better on the Rule Type items, and the second cluster performs slightly better on the Rule Direction and Graphical Component Nature items. However, these differences were very small and inconclusive given the limited sample size. We found that there was some systemacity between the cluster assignment and participants; that is, some participants were assigned consistently to one cluster over another; the number of these participants was larger than what would have been expected if participants switched between patterns randomly. Thus, we also explored the possibility that the amount of switching between the two patterns could be related to performance. However, we did not find any notable patterns. For more details, see `osf.io/2zkj8/`.

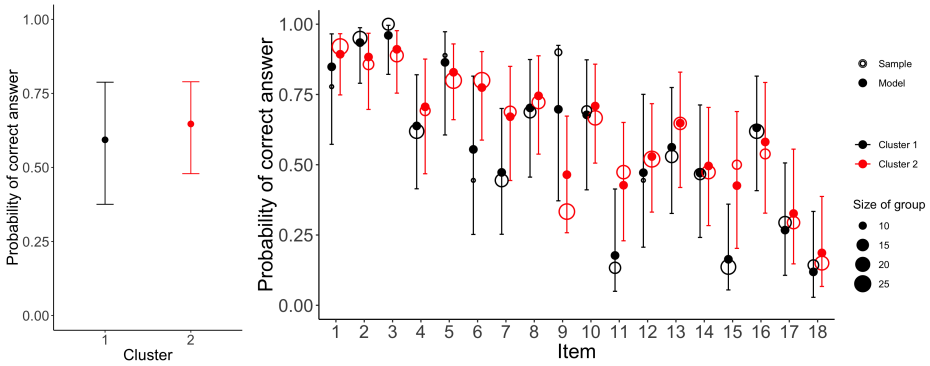


Figure 3.11: Left panel shows the marginal average probability of a correct answer of each cluster for the two clusters. Right panel shows the probability of a correct answer for each cluster and each item separately. The circles denote the observed proportion of correct answers (and the size of the circle represents the number of data points), whereas dots denote the mean of the posterior distribution. Error bars correspond to 95% credible intervals.

To sum up, we found two clusters in solving progressive matrices. Contrary to the results from previous literature (Hayes et al., 2011, 2015; Laurence et al., 2018; Vigneau et al., 2006), we did not find a clear pattern that would correspond to the response elimination strategy. However, the two clusters would roughly correspond to patterns, one of which is predominantly driven by transitions within rows, whereas the other is characterised by mixtures of transitions within rows and within columns. To our knowledge, the second pattern is rarely discussed in the literature as a viable alternative to solve the matrices. It is not impossible that other, more nuanced sub-strategies remained hidden in our analysis, for example, switching between different patterns (i.e., row-wise, column-wise, and matrix-response transitions), instead of using these patterns as a pure cognitive strategies.

3.5 Conclusion & Discussion

In this article, we centralize the idea of classification scanpaths where we can assume that different strategies to solve a cognitive task could elicit different types of gaze behavior. To this end, using an unsupervised method for cluster-

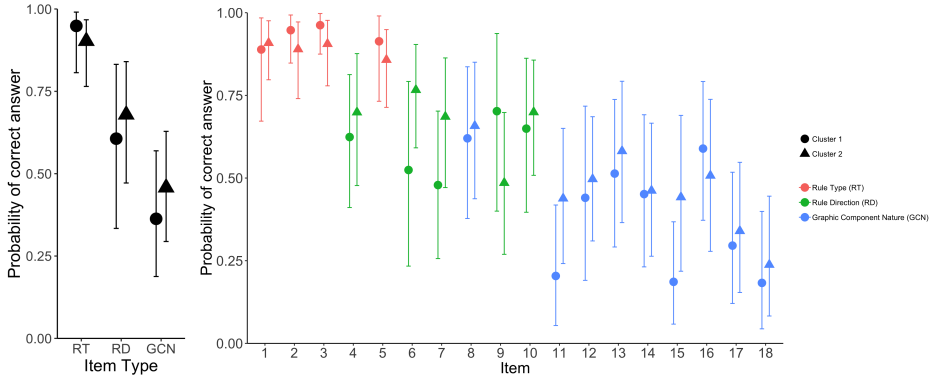


Figure 3.12: Left panel shows the interaction between the three item types and the cluster with respect to the probability of correct answer. Right panel shows the probability of a correct answer for each cluster and each item separately. Error bars correspond to 95% credible intervals.

ing transition matrices, we can discover groups of similar eye movement patterns without the need to assume that the groups differ on some other variable (e.g., performance in the task). This is of special interest in contexts where the groups are hypothesized and have to be inferred from the data, as well as the relationship of the group to the other variables is hypothesized and needs to be empirically tested. This problem arises frequently in the discussion of strategies in solving cognitive tasks, which we presented with two examples using the Deductive Mastermind game and Progressive Matrices task.

In the Mastermind example, we showed that we can retrieve patterns that correspond to systematic search for the most informative feedback, compared to less systematic scanning patterns guided by the order of the feedback presented. Such patterns that were predicted based on the logical reasoning analysis of the items in Gierasimczuk et al. (2013). In this example, the differences between the groups were detectable by visual inspection, which allowed us to conduct a realistic simulation study. Hence the classification should be relatively easy. From our point of view, this is a virtue of our example: showing that an automatic method arrives at the same conclusion as working through the data manually should assure us that the method is indeed valid. Furthermore, the application of the method to the real data revealed one pattern where the participant solves the item in a relatively non-systematic way, but switches to

the systematic pattern once arriving at the most informative feedback, whereas another pattern suggests that the participant attempts to solve the item in a relatively non-systematic way, and does not recognize that the third row is sufficient to arrive at the conclusion.

The second application used data from Progressive Matrices items. We found a pattern corresponding to the one described in the previous literature (solving analytically the matrices by progressing through the rows, Hayes et al., 2011, 2015), and one additional pattern that can be roughly described as progressing through the matrices within their columns, a pattern that has not been reported previously. Inspecting solutions with more clusters yielded qualitatively comparable results suggesting that we were unable to detect any additional patterns except these two. The patterns we found did not show differences on various summary measures (derived from eye movements, but also the performance in the task), thus, it would be hard to disentangle these patterns using supervised or semi-supervised methods, which have been predominantly used in earlier attempts to discover strategies in similar cognitive tasks (Hayes et al., 2011, 2015; Loesche et al., 2015; Vigneau et al., 2006). Contrary to the previous literature, we did not find a pattern that would correspond to the response elimination strategy. It is possible that the chosen representation of eye movement patterns (i.e., transition matrices) is unable to detect the response elimination pattern. Another option could be that the response elimination pattern occurs rarely as a pure strategy, but is rather emerging as a short phase during solving the items, after more systematic phases (e.g., that a person falls back on the response elimination after he or she fails to deduce the correct solution using analytic matching). If this is the case, our method could miss this pattern as it assumes that the eye-movements follow one pattern throughout solving the individual item (i.e., it is not possible to detect switches between patterns during solving the task). We believe that a comprehensive re-analysis of existing data sets (Hayes et al., 2011, 2015; Laurence et al., 2018; Loesche et al., 2015; Vigneau et al., 2006) using a range of different methods, or a (large-scale) replication study might be appropriate to find the common ground for the findings.

Our choice of the specific representation of the eye movement data, and the method for clustering as well as the distance metric is up for a debate. Dif-

ferent analytic choices could yield different results, depending on the questions and context of the analysis. We used transition matrices because the predicted strategies should differ in the transition matrices, hence, it should be possible to identify them as such. However, using transition matrices requires predefined areas of interest, and thus the method is limited only to applications where these areas can be defined without many arbitrary decisions. In these situations, transition matrices are simple to construct and interpret, although this should be done with caution. Some authors (Hayes et al., 2011) suggested that looking only at first-order transition probabilities is too much of a simplification of the eye movements data. Further, even very different scanpaths can have similar transition matrix (von der Malsburg & Vasishth, 2011). Thus, there is an intrinsic epistemological asymmetry – it is easier to discover qualitatively different groups of eye movement patterns than to provide evidence that some hypothesised pattern is missing (as is the case of our application on the Progressive matrices). To some extent, this asymmetry would likely occur regardless of the representation of eye movements as there will be potentially always some aspect of the data that has been left unmodelled. Individual researchers thus need to make informed decisions what representation of eye movement data to use, and if possible, commit to the analysis in advance to enable confirmatory analyses (de Groot, 2014). Exploratory analyses using different analytic approaches and eye movement representations can be then used to complement, expand, or challenge the confirmatory findings and their theoretical underpinnings (Jaeger & Halliday, 1998) – especially if methods that build upon different assumptions lead to different results. We hope that the method we demonstrated in this article enriches the analysis toolbox for latent inverse-Yarbus problems and will offer new insights, as we showed in our two examples.

The k-means clustering method based on minimizing squared Euclidean distances was chosen based on purely pragmatic reasons. It may be thought that the k-means is not the most appropriate method for clustering transition matrices, as it corresponds to the simplest form of mixture model for multivariate normal data - whereas transition matrices are essentially multivariate vectors of probability simplicia. Furthermore, the selection of the number of retained clusters with scree plots is somewhat arbitrary, and the k-means assumes that

the groups are of equal size, leading to a bias (and potentially incorrect classification) if that is not the case. These limitations can be tackled with more advanced modeling, either by specifying a full (hidden) Markov model and use clustering techniques on them (Chuk, Chan, & Hsiao, 2017; Chuk, Crookes, et al., 2017), modeling the data as mixtures of categorical time-series where the transition matrices can be thought of as collections of multinomial variables (Pamminger & Frühwirth-Schnatter, 2010), or mixture modeling of even more complex time-series models (e.g., Berchtold & Raftery, 2002). Whereas either of these methods would probably do more justice to the data, we believe that simpler methods such as the k-means might be useful. Computing transition matrices is a simple task and the k-means is implemented as a basic algorithm in most of the statistical software, can be run without extensive modelling experience and knowledge, and thus is widely available to all researchers. Thus, the method we proposed can prove to be a simple alternative to assess hypotheses about qualitatively different groups of scanpaths, or explore whether the data set comprises of homogeneous eye movements patterns. Furthermore, the method is able to capture the patterns on single item basis, which we have also shown using simulations. This enables us to potentially investigate within-person variability in the cluster assignment (e.g., due to effects of learning). Further, even within the simple approach of k-means, there may be possible important improvements, such as using clustering based on different distance measures, different criteria for selection of the number of clusters (e.g., Tibshirani, Walther, & Hastie, 2001), or regularized k-means or k-means with variable selection to tackle the dimensionality of the data and identifying features important for detecting differences between the clusters (e.g., Chormunge & Jena, 2018; Sun, Wang, Fang, & others, 2012).

While the method's advantages perhaps facilitate its use in wide range of application, it provides only limited options for modeling the eye movement data in more flexible manner. In particular, we cannot fix certain parameters to balance over-fitting and under-fitting, nor can we take into account hierarchical structure of the data (i.e., participant and item characteristics). This limitation proved to be important in our Mastermind example, where the non-systematic, top to bottom strategy should more or less exhibit similar pattern across all items, whereas the systematic strategies should exhibit different pat-

terns depending on the structure of the feedback. This is why we limited our example to only four items where the systematic strategy should elicit the same pattern. We could partially solve the problem by recoding AOIs on some items to conform to the same expected transition matrix, but it would not solve the problem in general. On the other hand, more flexible approaches to modeling the transition patterns would enable us to fix the strategies across items for one cluster, but let vary the strategies across items for another. Furthermore, more advanced modeling techniques could be used to identify or extend models of response behavior that assume latent states of different cognitive processes (e.g., Dutilh, Wagenmakers, Visser, & van der Maas, 2011; Molenaar, Oberski, Vermunt, & De Boeck, 2016; van Maanen, Taatgen, van Vugt, Borst, & Mehlhorn, 2015), some of which were partially motivated by the results of eye-tracking studies on cognitive tasks (Molenaar & de Boeck, 2018). However, we believe that even simple methods such as the method proposed in this article provides new ways to analyse data and derive new hypotheses, as well as think about novel directions of the eye-tracking applications.

Open Practices Statement

The data and analysis code are openly available at `osf.io/wvzs9/`. All analyses are exploratory and not preregistered.

It's a good run, but it's a poor run, if you know what I mean?

–Michael Owen

Chapter 4

Hidden Markov Models of Evidence Accumulation in Speeded Decision Tasks

This chapter is published as Kucharský, Š., Tran, N.-H., Veldkamp, K., Raijmakers, M., and Visser, I. (2021). Hidden Markov models of evidence accumulation in speeded decision tasks. *Computational Brain & Behavior*, *4*, 416–441. doi: 10.1007/s42113-021-00115-0

Abstract

Speeded decision tasks are usually modeled within the evidence accumulation framework, enabling inferences on latent cognitive parameters, and capturing dependencies between the observed response times and accuracy. An example is the speed-accuracy trade-off, where people sacrifice speed for accuracy (or vice versa). Different views on this phenomenon lead to the idea that participants may not be able to control this trade-off on a continuum, but rather switch between distinct states (Dutilh et al., 2011).

Hidden Markov models are used to account for switching between distinct states. However, combining evidence accumulation models with a hidden Markov structure is a challenging problem, as evidence accumulation models typically come with identification and computational issues that make them challenging on their own. Thus, an integration of hidden Markov models with evidence accumulation models has still remained elusive, even though such models would allow researchers to to capture potential dependencies between response times and accuracy within the states, while concomitantly capturing different behavioral modes during cognitive processing.

This article presents a model that uses an evidence accumulation model as part of a hidden Markov structure. This model is considered as a proof of principle that evidence accumulation models can be combined with Markov switching models. As such, the article considers a very simple case of a simplified Linear Ballistic Accumulation. An extensive simulation study was conducted to validate the model's implementation according to principles of robust Bayesian workflow. Example reanalysis of data from Dutilh et al. (2011) demonstrates the application of the new model. The article concludes with limitations and future extensions or alternatives to the model and its application.

4.1 Introduction

Evidence accumulation models (EAMs) have become widely popular for explaining the generative process of response times and response accuracy in elementary cognitive tasks (N. Evans & Wagenmakers, 2019). The strength of EAMs is their ability to accurately describe the speed-accuracy trade-off in speeded decision paradigms. The speed-accuracy trade-off is the conundrum that typically occurs when participants are instructed to make faster decisions, thereby increasing their proportion of errors (Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010; Luce, 1991; Wickelgren, 1977). The trade-off implies that in some situations, people can be slow and accurate, whereas fast and inaccurate in other situations. The dependency between response times

and responses generally frustrates interpretation of response time and accuracy at face value. EAMs aim to capture and explain this dependency between response times and accuracy, and enable inference on the latent cognitive constructs and a mechanistic explanation of the observed response time and accuracy. Thus, such analyses often enable us to tell, for example, whether slowing down is caused by increased response caution, increased difficulty or decreased ability of the respondent (N. Evans & Wagenmakers, 2019; van der Maas et al., 2011).

The traditional view of the speed-accuracy trade-off is that of a continuous function. That is, people are able to control their responses on the entire continuum from "slow and accurate" to "fast and inaccurate". This is an intrinsic assumption of EAMs which makes it possible to manipulate parameters associated with "response caution" to make more or less accurate (on average) decisions by slower or faster (on average) responding. Under such a view, it is in principle possible to hold average accuracy to any value between a chance performance and a maximum possible accuracy (often near 100%), by adjusting how fast one needs to be.

An opposing view is that of a "discontinuity" hypothesis (Dutilh et al., 2011), which states that people are not able to trade accuracy for response time on a continuous function, but rather switch between different stable states. The discontinuity hypothesis in speeded decision-making is strongly associated with thinking about two particular response modes: a stimulus controlled mode and a guessing mode (Ollman, 1966). Under the stimulus controlled mode, one is maximizing response accuracy while sacrificing speed; whereas under the guessing mode, choices are made at random for the sake of responding relatively fast. Hence, there are two modes of behavior under discontinuity hypothesis. Such dual behavioral modes are present in many models of cognitive processing (e.g., dual processing theory; J. Evans, 2008).

The discontinuity hypothesis has an increasing relevance in the speeded decision paradigm because it is able to explain specific observed relationships between decision outcomes and reaction times that standard EAMs cannot account for (Dutilh et al., 2011; Molenaar et al., 2016; van Maanen, Couto, & Lebreton, 2016). One of the most elaborate theoretical and empirical investigations of the "discontinuity" hypothesis is the phase transition model for the

speed-accuracy trade-off (Dutilh et al., 2011), which added several more predictions regarding the dynamics of switching between the controlled and guessing state. These phenomena can be modeled using hidden Markov models (HMM, Visser, 2011; Visser, Raijmakers, & van der Maas, 2009). Dutilh et al. (2011) used HMMs to model their data such that response time and accuracy are independent conditional on the state. Specifically, the model assumed that the responses are generated from a categorical distribution and response times from the lognormal distribution, independently of each other. Thus, the speed-accuracy trade-off is described only by assuming one slow and accurate state, and one fast and inaccurate state. However, at least under the controlled state, evidence accumulation presumably takes place to generate the responses, and so can lead to continuous speed-accuracy trade-off typical for EAMs, although within a smaller range than assumed under the continuous hypothesis. Thus, inference on the latent cognitive constructs given by the EAM might be the preferred option, but is neglected under the current HMM implementations of the phase transition model. Combining EAM with HMM would thus result in a model that is discontinuous on the larger scale (between state speed-accuracy trade-off), and continuous on the smaller scale (within state speed-accuracy trade-off), representing a third theoretical possibility beyond purely continuous and purely discontinuous models (Dutilh et al., 2011).

Fitting an HMM combined with an EAM would enable researchers to test specific predictions coming from the phase transition model as well as utilizing the strength of the EAM framework to account for the continuous speed-accuracy trade-off within the states. The ability of EAMs to infer the latent cognitive constructs liberates researchers from defining the states solely in terms of their behavioral outcomes. For instance, instead of describing the controlled state on the observed behavioral outcomes only (i.e., "slow and accurate"), EAMs allows researchers to form a mechanistic explanation of the observed behavioral outcomes using the latent cognitive constructs (i.e., "high response caution and high drift rate"). Further, capturing residual dependency between the observable variables conditionally on the latent states could improve performance of an HMM in terms of classification accuracy.

However, fitting EAMs can be a challenging endeavor, especially for more complicated models that allow for various sources of within and between trial

variability, which often exhibit strong mimicry between different parameters, and as such belong to the category of "sloppy models" (Apgar, Witmer, White, & Tidor, 2010; Gutenkunst et al., 2007). More complicated models, such as leaky competitor models, are not analytically tractable, and subject to highly specific simulation-based fitting methods (N. Evans, 2019). Thus, combining EAMs with HMMs, which themselves come with several computational (e.g., evaluation of the likelihood of the whole data sequence, Visser, 2011) and practical (e.g., label switching, Spezia, 2009) challenges, is highly demanding. The only successful applications of HMMs in these tasks is in combination with models that cannot capture possible residual dependencies, usually log-normal models or shifted Wald models for response times (Dutilh et al., 2011; Molenaar et al., 2016; Timmers, 2019). Yet, even the supposedly simplest complete model of response times and accuracy — the Linear Ballistic Accumulation model (LBA, S. D. Brown & Heathcote, 2008) — has proven to be difficult to combine with an HMM structure or even as a simple independent mixture (Veldkamp, 2020); this may not come as a surprise considering the general identifiability issues of the standard LBA model (N. Evans, 2020).

Given the potential of complex cognitive models to suffer from computational issues, it is important to present evidence that the model implementation is correct and that the procedure used to fit the model on realistic data (in terms of plausible values but also size) indeed succeeds in recovering the information that is used for inferences. The importance of validating models in terms of practical applicability is ever more increasing with the growing heterogeneity of approaches for fitting complex models, as well as modern approaches to build custom models tailored to specific purposes.

This need is taken seriously in this article which implements and validates a simple (constrained) version of the LBA model as part of an HMM. This model makes it possible to capture the discontinuity of the speed-accuracy trade-off by the HMM part, while concomitantly striving to capture the residual dependency between speed and accuracy within the states. Further, the model retains the fundamental inferential advantages of an EAM framework, but is analytically tractable and stable enough to be used with standard, state-of-the-art, modeling tools. To our knowledge, this is the first working combination of an HMM and an EAM, and serves as a proof of concept.

The structure of this article is as follows. First, the model is described in conceptual terms to explain the core assumptions and mechanics. Second, a simulation study summarises all steps that were followed when building and validating the model in accordance with a robust Bayesian workflow (Lee et al., 2019; Schad et al., 2019; Talts, Betancourt, Simpson, Vehtari, & Gelman, 2018). The model validation is followed with an empirical example to demonstrate the full inferential power of the model on experimental data. The article concludes with discussion and future potential directions towards improving the model.

4.2 Model

The general architecture of the model for response times and choices that we adopt here is the same as for the Linear Ballistic Accumulator (LBA, S. D. Brown & Heathcote, 2008). In the standard LBA, each response option is associated with its own evidence accumulator. Each accumulator rises linearly towards a threshold from a randomly drawn starting point, with its own specific drift rate, drawn from some distribution (commonly a normal distribution that is truncated at zero). The first accumulator that reaches its decision threshold triggers the corresponding response. Figure 4.1 explains the basic mechanics of typical LBA model.

Although the LBA became a popular choice for analyzing response times and accuracy, more recently evidence has surfaced suggesting practical identifiability issues of the standard LBA model — especially when trying to quantify differences in parameters such as decision boundary or drift rates between experimental conditions (N. Evans, 2020). Given that HMMs can be viewed as way to quantify differences between "conditions" (states) which themselves need to be inferred from the data, (lack of) identifiability of the standard LBA in combination with HMMs is a concern (especially in the upper bound of the starting point Timmers, 2019; Veldkamp, 2020).

However, there exists a number of potential remedies to solve the identifiability issue of the standard LBA. These remedies involve constraining the LBA model in some way while retaining as much flexibility of the model as possible to account for different patterns in the data, and to still allow inferences on the most fundamental parts of the evidence accumulation decision

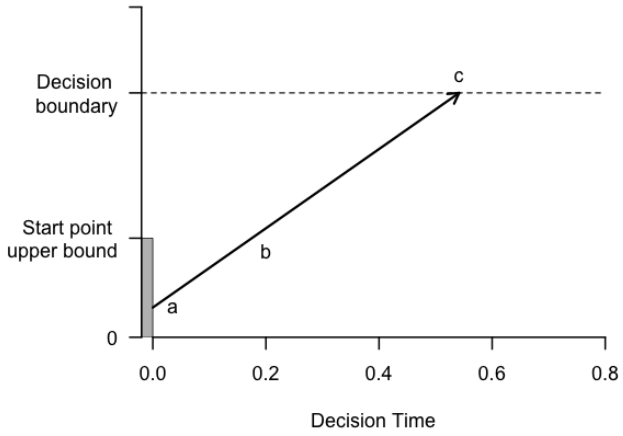


Figure 4.1: Linear Ballistic Accumulator (S. D. Brown & Heathcote, 2008). Each response outcome has an independent accumulator. For simplicity, the plot shows only one accumulator. (a) starting point for each accumulator is generated from Uniform distribution between zero and the upper bound of the starting point. (b) accumulator is launched from the starting point and with a drift rate that is generated from normal distribution with a mean drift and standard deviation of drift rate. (c) decision is made based on which accumulator hits the decision boundary first. Final response time is the sum of the decision time (the time it took the first accumulator reach the boundary) and a non-decision time (a fixed time for encoding the stimuli and motoric response).

process (e.g., speed of accumulation, response caution, etc). For example, a relatively well established set of constraints is to ensure that the average drift rates across accumulators are equal to some constant value (e.g. a scaling value of 1, Donkin, Brown, Heathcote, & Wagenmakers, 2011; N. Evans, 2020; Visser & Poessé, 2017). Such constraints may be accompanied by implementing equality constraints on parameters such as the upper bound of the starting point or the standard deviation of the drift rates. In the context of different conditions, even more stringent (equality) constraints are possible, such as equating parameters (such as drift rate for the "error" response) across conditions (N. Evans, 2020).

This article aims to provide a proof of concept that EAMs and HMMs can

be combined into a single model. The present application simplifies the LBA model to a bare minimum and acts as a sanity check – in case even very minimalist EAM models cannot be employed as part of a HMM model, there is little reason to expect that more complex, complete and computationally demanding models of decision making will be more successful.

The bare minimum, simple instance of LBA is achieved in this article by setting several constraints on the parameters. For practical reasons, we will refer to this model as sLBA, a short for "simplified Linear Ballistic Accumulator". Most significantly, the model implemented in this article fixes all starting points at zero, effectively removing the variability of the starting point. As commonly done in the LBA, we constrain the drift rates to sum to unity. In addition to that, the drift rates are assumed to have equal standard deviations across accumulators. Full details on the model, its likelihood and identifiability are described in Appendix 4.A, additional helpful derivations can be found in Nakahara, Nakamura, and Hikosaka (2006). Figure 4.2 explains the model in additional detail.

The simplification achieved by removing the variability of the starting point makes the model coarsely similar to the LATER model (Linear Approach to Threshold with Ergodic Rate, R. H. S. Carpenter, 1981; Noorani & Carpenter, 2016), with the difference that the current model explicitly evaluates the likelihood of observing the first accumulator that reached the threshold according to the general race equations (see Heathcote & Love, 2012), and contains additional parameters (such as non-decision time). Therefore, it enables researchers to model accuracy in addition to response times as opposed to the LATER model (see Ratcliff, 2001, for critique of LATER for inability to do so).

The constraints employed in this application greatly reduce the complexity compared to the standard LBA model. Specifically, our model for responses and response times on a two choice task contains the following parameters: the average drift rate for the correct (ν_1) and incorrect (ν_2) responses, the standard deviation of the drift rates (σ), the decision threshold (α), and the non-decision time (τ). The latter three parameters are equal for both accumulators.

The purpose of simplifying the LBA model is to employ it as a distribution of response times and responses in an HMM. Specifically, the current model assumes two latent states: A "controlled" state ($s = 1$) and a "guessing" state

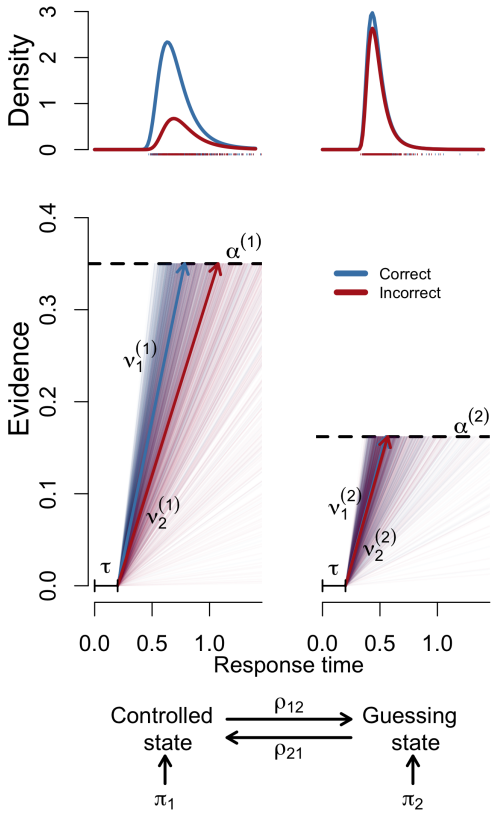


Figure 4.2: HMM combined with sLBA. Bottom panel: Latent controlled and guessing states evolve as a Markov chain, with initial state probabilities π_1 and π_2, and transition probabilities ρ_{12} and ρ_{21}. Middle panel: Non-decision time τ shifts the response times. Correct and incorrect responses launch an accumulator (starting at 0), with a drift rate drawn from a truncated Normal distribution with mean drift rate ν and a standard deviation σ. The plot shows the average drift rates as thick arrows, and realisations of the random process as thin lines to represent the randomness of the process. Accumulator that reaches the decision boundary α first launches corresponding response. Average drift rates and decision boundary can differ between the states. Top panel: Under the controlled state (left), the expected response times are larger than under the guessing state (right), but the accuracy is higher (i.e., the decision boundary is reached by the correct accumulator more often).

($s = 2$). These states evolve according to a Markov chain, which is characterized by the initial (π_1 and π_2) and transition state probabilities ρ_{ij}, where the first index i corresponds to the outgoing state and j corresponds to the incoming state: For example, ρ_{12} is the probability that the participants switch from the controlled state to the guessing state.

Traditionally, these states would be equipped by their own distribution of response times and responses, possessing their own parameters. That is, we could use the LBA model for each latent state of the HMM, and estimate the drift rate for the correct responses for the first state $\nu_1^{(1)}$, second state $\nu_2^{(2)}$, and similarly for all of the parameters. However, we further reduce the complexity of the model by equating some parameters between states. Specifically, we assume that the difference between the guessing state and the controlled state is evoked by differences between average drift rates and decision thresholds. The rest of the parameters are held equal across the states. Thus, equality constraints $\sigma^{(1)} = \sigma^{(2)}$ and $\tau^{(1)} = \tau^{(2)}$ are used to further simplify the model.

Additionally, there are some notable considerations regarding the controlled and guessing states, which will later help setting priors and preventing label switching. Specifically, the controlled state has higher average drift rate for the correct response than the guessing state ($\nu_1^{(1)} > \nu_1^{(2)}$, and consequently $\nu_2^{(1)} < \nu_2^{(2)}$ due to the sum-to-one constraint of the drift rates, see Appendix 4.A) at the expense of having higher decision threshold ($\alpha^{(1)} > \alpha^{(2)}$). Further, if the second state truly is guessing, the drift rates under this state should be roughly the same: $\nu_1^{(2)} \approx \nu_2^{(2)} \approx 0.5$.

4.2.1 Implementation

We implemented the HMM and LBA model in a probabilistic modeling language Stan (B. Carpenter et al., 2017); specifically, v2.24.0 release candidate of `CmdStan` (`github.com/stan-dev/cmdstan/releases/tag/v2.24.0-rc1`, Stan Development Team, 2020). In this version of Stan, several new functions were introduced that implement the forward algorithm for calculating the log-likelihood of the data sequence, while marginalizing out the latent state parameters (for easy introduction, see Visser, 2011), which makes estimating HMM models in Stan much easier, computationally cheaper, and less error-prone than before (which required manual coding of the forward al-

gorithm). The sLBA distribution of response times and responses was custom coded in the Stan language. We executed `CmdStan` from the statistical computing language R (R Core Team, 2020) using the R package `cmdstanr` (Gabry & Češnovar, 2020). The code is available at `github.com/Kucharssim/hmm_slba`.

4.2.2 Label switching

Finite mixture models and Hidden Markov models share the characteristic that the likelihood of the models is typically invariant to the permutation of the latent state labels (Jasra, Holmes, & Stephens, 2005; Spezia, 2009). This means that fitting the model can result in different estimates, depending on towards which state configurations the fitting procedure leads to. In the current context of guessing and controlled state, it is not possible on the basis of the model likelihood alone to state whether model 1 should be controlled or guessing state and vice versa – both options lead to the same likelihood value. There are several perspectives on dealing with potential label switching, perspectives that differ in terms of what types of applications and inferential paradigms one follows. For example, in maximum likelihood paradigm, label switching is not a severe problem as the analyst can simply relabel the states after the model has been fitted, based on how the parameter estimates can be interpreted. In Bayesian framework (especially with MCMC), the problem is more complicated as the label switching can manifest in different ways, and can also depend on the sampler (and its settings) one uses to obtain the estimates of the entire posterior distribution. Common remedies of label switching are, for example 1) Change the model so that emission distributions under each state are uniquely identified, 2) establish parameter inequalities which leads to identifying the labels, 3) use of informative priors that lead to better identification of the apriori constraints, 4) some form of state relabeling of the posterior samples, among others. Usually, various remedies are combined together as the solutions do not work in generality for all possible mixture problems and applications.

In the current application, we heavily rely on approach 3), whereby specifying informative priors leads to soft identification of the state labels, i.e., associating slow and accurate responding with a (controlled) state 1 and fast and inaccurate responding with a (guessing) state 2. However, it is important that

even which informative priors, one is only increasing the apriori probability of some state configuration, but does not render other configurations impossible. In fact, the other state configurations are still *valid* modes of the joint posterior space, albeit less plausible according to the prior specification. In some applications (estimating marginal likelihood in order to conduct model comparison; Frühwirth-Schnatter, 2004), it is actually desirable to make sure that the sampler is switching between state labeling freely, to ensure that the MCMC sampling efficiently explores the joint posterior in its entirety. In purely estimation settings (which is the case of this article which is not concerned by model comparison), one does not need to ensure that all valid modes of the posteriors are explored efficiently, as long as the main mode is explored well, which, among others, entails checking whether the labels did *not* switch, either within- or between- the MCMC chains.

4.3 Simulation study

In order to investigate the quality of inferences we draw from the model, a simulation study was conducted. Specifically, we conducted the simulation in accordance with a principled Bayesian workflow (Schad et al., 2019). The simulation study consists of 1) prior predictive checks to identify priors that reflect our domain specific knowledge, 2) a computational faithfulness check to test correct posterior distribution approximation, 3) model sensitivity analysis to investigate how well the estimated posterior mean of parameter matches the true data generating value, and the amount of updating (i.e., how much are the parameters informed by the data). Additionally, as is the case in classical model validation simulation, we report standard parameter recovery results, including coverage probabilities of credible intervals.

4.3.1 Prior predictives

Choosing prior distributions is an integral part of the Bayesian model-building process because the prior should reflect theoretical assumptions and cumulative knowledge about the parameter space as well as aid model convergence (Gershman, 2016; Vanpaemel, 2011). Ideally, the priors should be informed and

constrained by a large collection of previous studies (Tran, van Maanen, Heathcote, & Matzke, 2020; Zwet & Gelman, 2022) to yield more efficient sampling and plausible estimates. In the current study, we selected prior distributions to constrain parameter values to reasonable regions of the parameter space (e.g., non-decision time must be positive, therefore we used an exponential distribution) and to nudge the model towards convergence. Concomitantly, our prior distributions were informed by the large collection of literature on evidence accumulation models applied to lexical and perceptual decision tasks (Tran et al., 2020). Interested readers who want to apply our models to different experimental tasks or non-standard populations, might want to consult the corpus of literature specific to the application to adjust the prior distributions.

To place priors that reflect our expectations about data from the tasks to which the model will be applied, we conducted prior predictive simulations. In particular, we first set out to generate 1,000 data sets each of 200 trials, which is generally a lower bar for running speeded decision tasks. Then, the following expectations of the generated data are defined, specified in terms of summary statistics across the 200 observations per data set. Throughout, response times are measured and reported in seconds. In case response times are measured in different units, the priors should be re-scaled appropriately.

Latent state distribution.

First, we expect that the number of trials participants spend in one or another state will be relatively even, and that it is very rare that participants would complete all 200 trials in a single state. The evenness is achieved by composing a symmetric initial state probabilities vector $\boldsymbol{\pi}$ and a symmetric transition matrix $\mathrm{P} = \begin{bmatrix} \boldsymbol{\rho_1} \\ \boldsymbol{\rho_2} \end{bmatrix}$. Further, we assume that the states are relatively sticky, therefore there will be a tendency to stay in the current state rather than switching to another state. Specifically, the average run length is expected to be approximately between 5–10, and that in at least 50% of the simulations the proportion of the trials under the controlled state ranges between 30% to 70%.

We chose the following priors

$$\begin{aligned}\boldsymbol{\pi} &\sim \text{Dirichlet}(5, 5)\\ \boldsymbol{\rho_1} &\sim \text{Dirichet}(8, 2)\\ \boldsymbol{\rho_2} &\sim \text{Dirichet}(2, 8).\end{aligned}$$

The initial state probabilities are assigned a symmetric Dirichlet prior. The hyperparameters slightly favor probabilities closer to 0.5. Usually, the initial state probabilities are not the focus of inference as they depend mostly on just the first trial. Thus, slightly informative priors were chosen to help the model to converge. For the transition probabilities, Dirichlet priors that favor "sticky" states were chosen. Specifically, the mean probability of staying under the current state is 0.8. There is still considerable uncertainty about how sticky the two states are: 90% of the prior mass for the probability of persisting in the current state lies between 0.63 and 0.94.

The results of the prior predictive simulation showed that the median of the average run length is 6.25, IQR[4.35, 9.524]. The distribution of the average run length is positively skewed. Although it could be expected in many experiments that run lengths could be higher, the priors would have to be much more informative (pushing the probability of staying in a current state closer to one) than the current settings. However, that would give only a very narrow range of the values used for validating the models. Therefore, the current setting of the prior is a compromise between prior expectations about the data and the need to validate the model on a wider range of parameter values. Regarding the percentage of trials in the controlled state, the distribution over the 1,000 simulations had a median of 0.51, IQR[0.35, 0.67].

Response and response time distributions.

We expect that the distributions of the responses will be the following. Under the controlled state, the proportion of correct responses is well above chance; we assume that under the controlled state, there is almost zero probability that a person would have accuracy smaller than 50%, and that it is possible to achieve relatively high accuracy on average ($\approx$75%). Under the guessing state, we assume that the average accuracy is exactly 50%.

For the distributions of the response times, we have the following expectations. First, the response times under the controlled state are on average slower than responses under the guessing state. Second, the responses under the guessing state are relatively rapid: responses in simple perceptual decision tasks can be faster than 1 sec on average. Third, the majority of response times does not exceed 5 sec (Tran et al., 2020).

Based on these considerations and prior predictive simulations, the following prior specification for the LBA parameters were identified as suitable:

$$
\begin{aligned}
\boldsymbol{\nu}^{(1)} &\sim \text{Dirichlet}(14, 6) \\
\boldsymbol{\nu}^{(2)} &\sim \text{Dirichlet}(10, 10) \\
\alpha^{(1)} &\sim \text{Gaussian}(0.5, 0.1)_{(0,\infty)} \\
\alpha^{(2)} &\sim \text{Gaussian}(0.25, 0.05)_{(0,\infty)} \\
\sigma &\sim \text{Gaussian}(0.4, 0.1)_{(0,\infty)} \\
\tau &\sim \text{Exponential}(5)
\end{aligned}
$$

Figure 4.3 and Table 4.1 summarise the prior predictive distribution of the accuracy (proportion of correct answers) under the two states separately. As desired, the accuracy under the controlled state is well above chance, whereas under the guessing state it clusters around 50%. There is considerable variability under both states, leaving the possibility for the model to learn from the data.

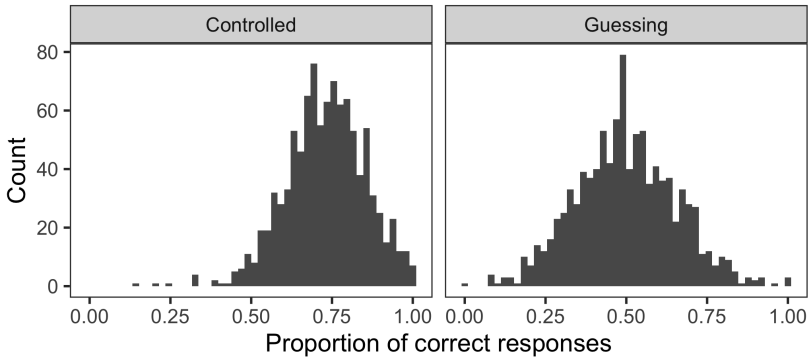


Figure 4.3: Prior predictive distribution of the response accuracy (proportion of correct answers).

			Quantile				
State	Mean	SD	2.5%	25%	50%	75%	97.5%
Controlled	0.73	0.12	0.48	0.65	0.73	0.81	0.96
Guessing	0.50	0.16	0.21	0.39	0.50	0.60	0.81

Table 4.1: Descriptives of the prior predictive distribution of the response accuracy (proportion of correct answers).

				Quantile				
State	Response	Mean	SD	2.5%	25%	50%	75%	97.5%
Controlled	Correct	0.92	0.28	0.49	0.73	0.87	1.03	1.57
Controlled	Error	1.09	0.34	0.59	0.87	1.03	1.26	1.82
Guessing	Correct	0.60	0.24	0.28	0.44	0.55	0.70	1.19
Guessing	Error	0.60	0.23	0.27	0.44	0.55	0.70	1.18

Table 4.2: Descriptives of the prior predictive distribution of the average response times.

Figure 4.4 and Table 4.2 summarise the prior predictive distributions of the average response times for correct and incorrect responses under the two states separately. As desired, the average response times are slower under the controlled state than under the guessing state. The majority of the average response times under the guessing state are below 1 sec, whereas under the controlled state cluster around 1 sec. There are no large differences between response times for correct and incorrect responses under the two states separately, although the average response times for incorrect responses under the controlled state show higher variance than for the correct responses. However, this phenomenon might be caused by the fact that there are more correct responses than incorrect responses under the guessing state, resulting in higher standard errors for the averages of the incorrect responses.

The prior distributions specified above may seem extremely informative, introducing "subjective" bias to the analysis. However, we believe the prior distributions are justified by our prior predictive simulations and based on cumulative characterizations of psychological processes underlying a lexical deci-

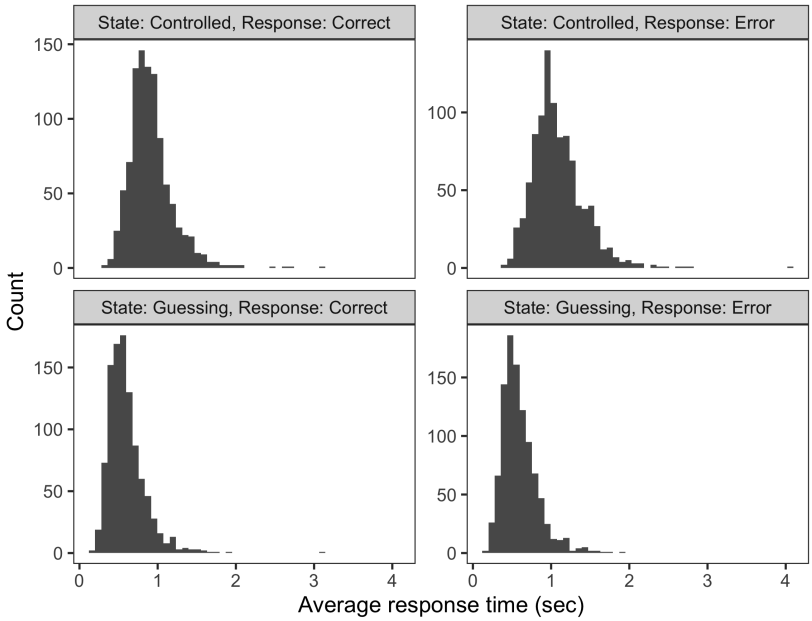


Figure 4.4: Prior predictive distribution of the average response times.

sion and a perceptual decision task of EAMs (Tran et al., 2020). Further, prior distributions may be also regarded as constraining the parameter space to plausible values (Kennedy, Simpson, & Gelman, 2019; Tran et al., 2020; Vanpaemel, 2011), similarly as a traditional statistician would decide on ranges of parameters for a simulation study. In the current study, the prior distributions actually cover slightly more volume of the parameter space than is typical in simulation studies of similar type (e.g., Donkin et al., 2011; Visser & Poessé, 2017). Lastly, priors on the parameters in both states (e.g., $\alpha^{(1)}$ and $\alpha^{(2)}$) are used to primarily separate the latent states from each other, and associate the first state with the controlled state (and conversely the second state with the guessing state). Using informed priors in such occasions prevents label switching problems, and gently nudges the model towards convergence.[1] However, the prior specification

[1]There are other techniques to identify states and prevent label switching (Jasra et al., 2005). For example, a common approach is to put an order constraint on the model parameters, for example, $\alpha^{(1)} < \alpha^{(2)}$, by using a transformation $\alpha_2 := \alpha_1 + \exp(\theta)$. Such a "hard" order restriction is effective in dealing with label switching, but makes it harder to reason about the prior specification. Further, "hard" order restrictions can hinder computing normalizing

does not ensure that the labels do not switch at all. When fitting the models, we performed additional checks using the posterior samples to check whether the labels indeed converged to the modes of the posteriors we intended.

4.3.2 Computational faithfulness

There are many ways in which model implementation can fail, especially in case of Bayesian models requiring MCMC. Possible problems might arise due to error in specification of the likelihood (or just insufficiently robust implementation), the use of difficult parameterizations, or a simple coding error. Another problem may arise when the model combined with the priors and the data result in a very complex parameter space for the MCMC algorithm to navigate, which may lead to inefficient exploration of the target posterior distribution. Such issues can lead to biased estimates, underestimating the uncertainty of parameters, or simply wrong inferences.

For the endless possibilities in which model implementation can fail, there was a lot of recent advancement in techniques that aim to check for *computational faithfulness* of a model — in the context of the Bayesian framework, this means testing whether the proposed MCMC procedure yields valid approximations of the posterior distributions (Schad et al., 2019). One established technique is Simulation-based calibration (SBC, Talts et al., 2018). As the model that we propose in this article is definitely suspect for computational problems, we use SBC to check our model implementation (although it could be argued that such checks should be done by default for non-standard models at least). Since these checks are not yet the standard in cognitive modeling literature (Schad et al., 2019), we briefly summarise the rationale behind SBC here, although the interested reader should refer to excellent articles by Talts et al. (2018) and Schad et al. (2019).

To check whether the method used for approximating the posterior distribution $\pi(\theta|\tilde{y})$ is correct, the following steps can be done: (1) draw from the prior distribution $\tilde{\theta} \sim \pi(\tilde{\theta})$, (2) draw a data set from the model using the generated values of the parameters, $\tilde{y} \sim \pi(\tilde{y}|\tilde{\theta})$, and (3) fit the model on the generated data to obtain the posterior distribution $\pi(\theta|\tilde{y})$. The draws from such

constants, in case one is eager to quantify the marginal likelihood (evidence) of the model (Frühwirth-Schnatter, 2004, 2019).

an obtained distribution, across many repeated replications of this procedure, should give back the prior distribution of the parameters $\pi(\theta)$. In short, SBC builds on the fact that (Talts et al., 2018)

$$\pi(\theta) = \int \int \pi(\theta|\tilde{y})\pi(\tilde{y}|\tilde{\theta})\pi(\tilde{\theta})d\tilde{y}d\tilde{\theta}, \tag{4.1}$$

which means that we can recover analytically the prior distribution on model parameters $\pi(\theta)$ by averaging the posterior distribution $\pi(\theta|\tilde{y})$ weighted by the prior predictive distribution $\int \pi(\tilde{y}|\tilde{\theta})\pi(\tilde{\theta})d\tilde{\theta}$. In order to check whether the prior distribution is indeed recovered, for each repetition, we compare the draw from the prior (that generated the data) to the samples from the posterior, and count the posterior samples that are smaller than the draw from the prior. If these two distributions are the same, every rank (i.e., the count of posterior samples that are smaller than the generating parameter value) would be equally likely – yielding an approximately uniformly distributed rank statistic (Talts et al., 2018).

Using the already created ensemble of 1,000 prior predictive data sets in section 4.3.1, each of the data sets was fitted using Hamiltonian Monte Carlo supplied by Stan (B. Carpenter et al., 2017). Due to computational constraints (typical run of a model averages roughly about 500 sampling iterations per minute on Apple's MacBook Air edition 2017), each model ran only with one chain for 500 warmup and 1,000 sampling iterations. Starting points were generated by drawing independent samples from the priors. In case the model label switched, the model was reran (at maximum five times). Model switching was detected by comparing the true (generative) states to the estimated states (identified using modal assignment based on mean state probabilities using the forward-backward algorithm). This resulted in non-label switching MCMC samples for 945 data sets out of the total 1,000. Since only 783 repetitions achieved acceptable values of the (split-half) Gelman-Rubin $\hat{R}$ statistic (Gelman & Rubin, 1992) between 0.99 and 1.01 for all of the parameters, we selected several data sets at random from non-converged cases and refitted them with 4 chains, 1,000 warmup and 1,000 sampling iterations. The new model fits had good $\hat{R}$ for all parameters, suggesting that the unsatisfactory convergence diagnostics were a consequence of the small number of MCMC iterations dur-

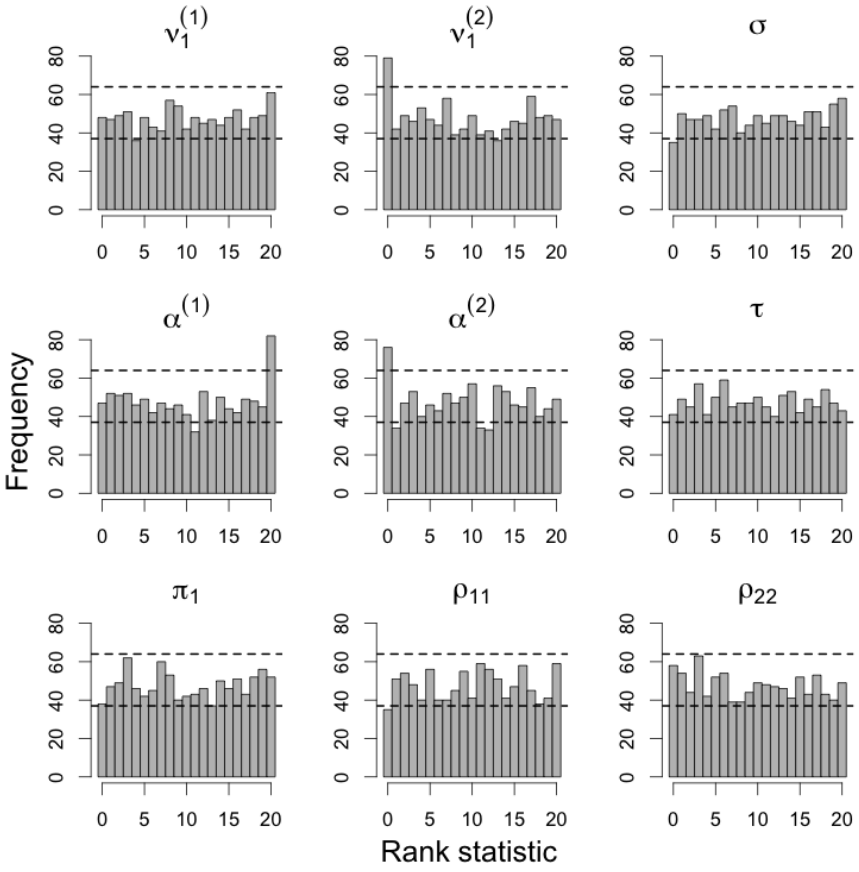


Figure 4.5: Simulation based calibration: Histogram of the rank statistic. The dashed lines correspond to the lower and upper limits of the 95% interval under the null hypothesis that the rank statistic is uniformly distributed.

ing the simulation. We excluded from the results only the repetitions that label switched, but kept those that did not yield satisfactory convergence diagnostics. Because the SBC rank statistic is sensitive to potential autocorrelation of the chain, the posterior samples were thinned by a factor of 50 — leading to the rank statistic ranging between 0 and 20.

Figure 4.5 shows the histogram of the SBC rank statistic for each of the parameter separately. Figure 4.6 shows the difference between the cumulative distribution and the theoretical cumulative distribution of a uniformly distributed variable (Talts et al., 2018).

The results show that none of the parameters exhibit typical patterns present in case that the posterior approximation is under-dispersed or over-dispersed compared to the true posterior (which would manifest as a $\cup$ or $\cap$ shape of the rank distribution, Talts et al., 2018). Further, the distribution of rank statistics for most of the parameters seem consistent with a uniform distribution, suggesting that the posterior approximation is very close to the true posterior. However, three parameters seem potentially problematic: the rank statistic for $\alpha^{(1)}$, $\alpha^{(2)}$, and $\nu_1^{(2)}$ show an excess of frequencies at 20 and 0, respectively, suggesting that $\alpha^{(1)}$ approximation could be underestimating the true posterior, whereas $\alpha^{(2)}$ and $\nu_1^{(2)}$ approximations could be overestimating the true posterior. However, this observation could also arise if the thinning was not efficient to reduce the autocorrelation of the chain (autocorrelation can result in excess of ranks at the edge of the distribution Talts et al., 2018). Additionally Figure 4.6 reveals that the rank distribution for ρ_{22} also potentially deviates from the uniform distribution. However, this deviance is not associated with any typical problem in posterior approximations, lacking a meaningful interpretation.

SBC gave us assurance that our model is capable of approximating the posterior distribution for most of the parameters. Three potentially problematic parameters remain, although the deviance from the expected results it small. Potential explanations for these deviances could be the constraints to resolve label switching (which could cause the truncation of the parameters for one state near values for the same parameter from the other state), or unsuccessful reduction of the auto correlations of the MCMC chains (which could be solved by running the procedure for more iterations and use higher thinning.)

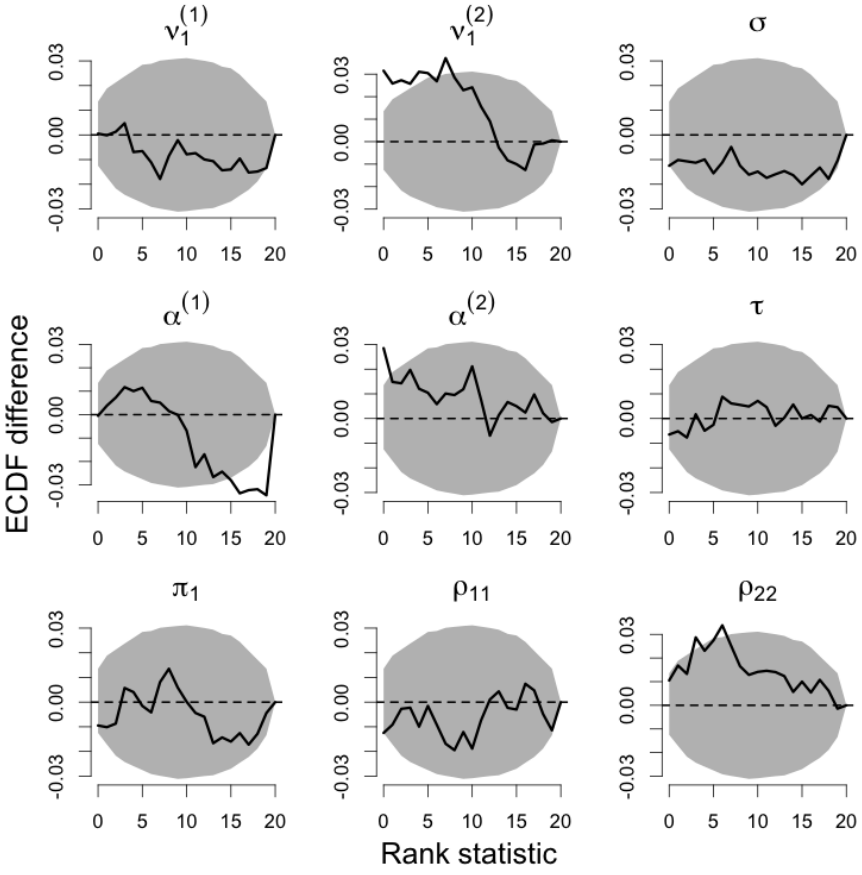


Figure 4.6: Simulation based calibration: ECDF of the rank statistic minus the ECDF of a uniformly distributed variable. The shaded area corresponds to the 95% interval under the null hypothesis that the rank statistic is uniformly distributed.

4.3.3 Model sensitivity

Next, the goal was to investigate for each parameter, (1) how well the posterior mean matches the true data generating value of the parameter, and (2) how much uncertainty is removed when updating the prior to the posterior. This is useful to investigate the bias-variance trade-off for each parameter, and to adjust our expectations regarding how much we can learn about parameters, given a data set of a specified size (in this simulation, number of trials = 200).

To answer (1), posterior z-scores for each parameter are defined as:

$$z = \frac{\mu_{\text{posterior}} - \tilde{\theta}}{\sigma_{\text{posterior}}}, \tag{4.2}$$

that is, the difference between the posterior mean and the true parameter value is divided by the posterior standard deviation. The posterior z-scores tell us how far the posterior expectation is from the true value, relative to the posterior uncertainty. The distribution of the posterior z-scores should have a mean close to 0 (if not, the posterior expectation is a biased estimator).

To answer (2), posterior contraction for each parameter is defined as:

$$\text{contraction} = 1 - \frac{\sigma^2_{\text{posterior}}}{\sigma^2_{\text{prior}}}. \tag{4.3}$$

If the posterior contraction approaches one, the variance of the posterior in negligible compared to the variance of the prior, indicating that the model learned a lot about the parameter of interest. Conversely, if the posterior contraction is close to zero, there is not much information in the data about the parameter, resulting in the inability to reduce the prior uncertainty.

These two variables are plotted against each other in a scatter plot, which provides useful diagnostic insights (Schad et al., 2019). Specifically, for each parameter, and each simulation which did not label switch, the posterior z-scores and posterior contraction are plotted on the y-axis and x-axis, respectively. Figure 4.7 shows the diagnostic plot for the nine parameters with equal axes between them to enable comparison between parameters.

All of the parameters cluster around z-scores of 0 (dashed horizontal line), suggesting that neither of the parameters exhibits systematic bias. However,

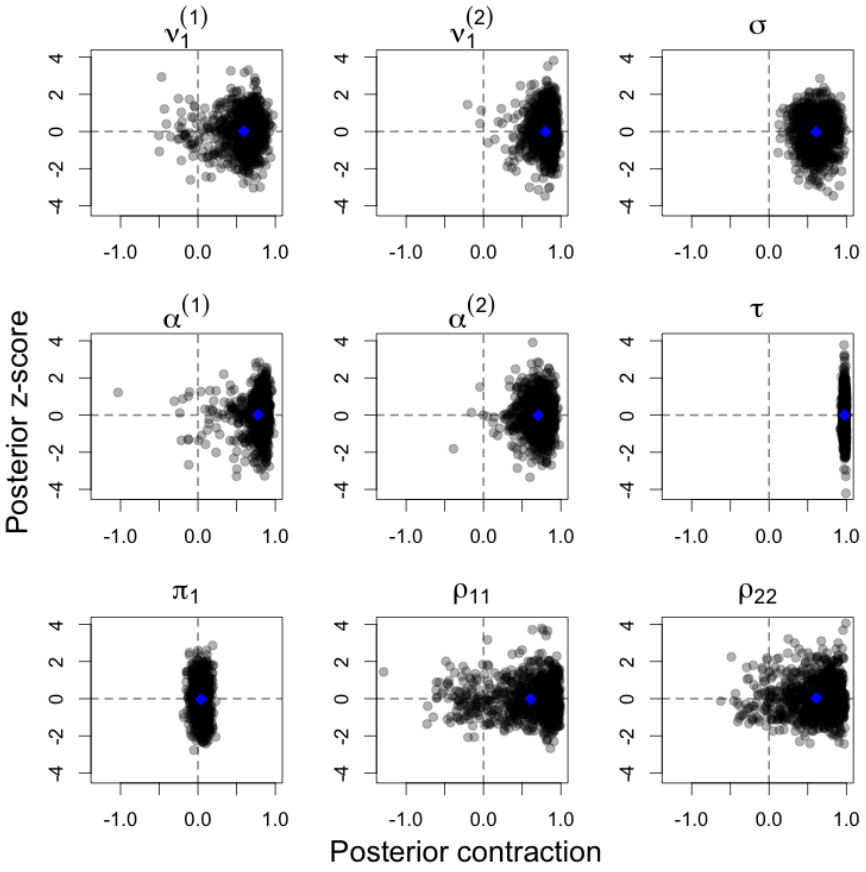


Figure 4.7: Model sensitivity plot for all nine parameters. Blue diamond shapes depict the means of the distributions.

there are large differences between parameters in terms of posterior contraction. The most contraction is present for the non-decision time τ, followed by the rest of the LBA parameters. We could expect that the contraction would increase with the number of trials. The worst results concern the initial state probability π_1: The posterior contraction basically stays at zero. However, this is expected as the initial state probability is affected mostly by just the first trial, and as such, there is not much information in the data about it. Increasing the number of trials would not help to identify this parameter, only repeated experiments would.

In general, the sensitivity analyses suggest that the amount of learning about the parameters of interest could be satisfactory given the typical experimental designs (our simulation was based on 200 trials per experiment, whereas typical decision tasks experiments could count multiples of that number), especially for the LBA parameters.

4.3.4 Parameter recovery and coverage probability

Traditional simulation studies aim to validate statistical models and assess the quality of a point estimator of a given parameter of interest. Additionally, such simulations are accompanied by assessment procedures. This section adheres to this tradition: for each of the parameters (that are not a linear combination of others) we report the standard "parameter recovery" results.

The simulation was done using two estimation techniques: the maximum a posteriori (MAP) estimation, and the posterior expectation (i.e., the mean of the posterior distribution). MAP is useful in situations where researcher needs to obtain estimates quickly, and does not need to express the uncertainty in the estimates. As the rest of the article focuses on full Bayesian inference, MAP results are presented only in the Supplementary Information. Pearson's correlation coefficient between the estimated parameter value and its true values serves as a rough indicator of parameter recovery. High correlations indicate that the model is able to pick up variation in the parameter. Additionally, scatter plots visualizing the relationship between the true and estimated parameter values show the precise relationship between the true and estimated values of the parameters.

We also investigate the coverage performance of the central credible inter-

vals. For each parameter, the frequency with which 50% and 80% central credible intervals contain the true data generating value was recorded. The confidence levels are relatively low compared to traditionally reported values, because we have only 1,000 MCMC samples per parameter due to computational constraints, which results in low precision in the tails of the posterior distributions (i.e., the tail effective sample size was generally too low).

Posterior expectation

Figure 4.8 shows the scatter plot between the true (x-axis) and estimated (y-axis) values (i.e., means of the posteriors) for the nine free parameters in the model: the drift for the correct choice under the controlled state ($\nu_1^{(1)}$), the drift for the correct choice under the guessing state ($\nu_1^{(2)}$), the standard deviation of drifts (σ), the decision boundary under the controlled ($\alpha^{(1)}$) and guessing ($\alpha^{(2)}$) state, the non-decision time (τ), the initial probability of the controlled state (π_1), the probability of dwelling in the controlled (ρ_{11}) and the guessing (ρ_{22}) state. The correlations for the LBA parameters range from high ($r = 0.77$ for $\nu_1^{(1)}$) to nearly perfect ($r = 0.99$ for τ) and the point lie close to the identity line, suggesting good recovery of the LBA parameters. An exception is the parameter σ, which shows a pattern of underestimating the true values, if the true value is relatively high.

As for the parameters characterizing the evolution of the latent states, the recovery of the initial state probability is sub optimal ($r = 0.22$). This is expected, as there is not much information in the data about this parameter (it mostly depends on the state of the first trial), and so it is highly dependent on the prior. This parameter is not to be interpreted, however, unless the model is fitted on repeated trial sequences (so that there are more "first" trial observations). The recovery of the two "dwelling" probabilities are satisfactory.

Coverage of the credible intervals

Using the MCMC samples, we computed the 50% and 80% central credible intervals for each parameter under each fitted model (that did not label switch), and checked whether the true value of the parameter lies within that interval. Table 4.3 shows that the relative frequencies with which the CIs cover the true value is very close to the nominal value of the confidence level. Thus, we did not

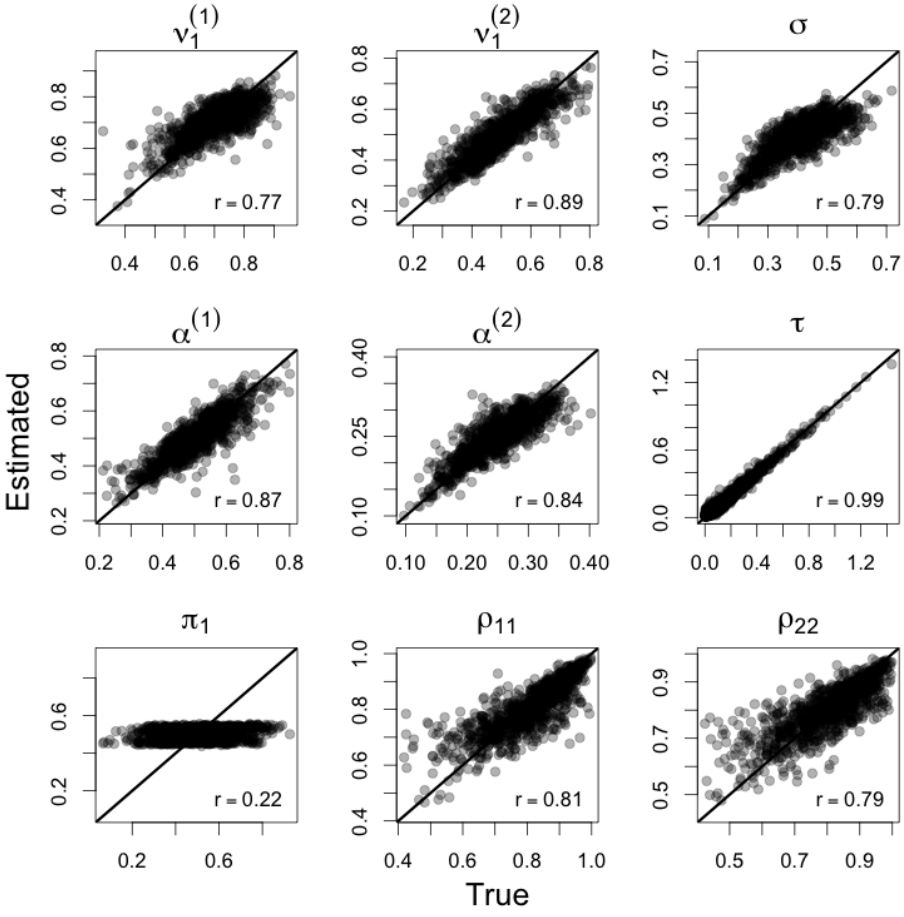


Figure 4.8: Parameter recovery using posterior expectation. Correlation plots between the true values (x-axis) and the estimated values (y-axis). The slope line shows the identity function.

	50% CI Coverage	80% CI Coverage
$\nu_1^{(1)}$	0.52 [0.49, 0.55]	0.79 [0.76, 0.82]
$\nu_1^{(2)}$	0.48 [0.45, 0.51]	0.79 [0.76, 0.82]
σ	0.51 [0.48, 0.54]	0.82 [0.80, 0.85]
$\alpha^{(1)}$	0.49 [0.45, 0.52]	0.78 [0.76, 0.81]
$\alpha^{(2)}$	0.51 [0.48, 0.54]	0.81 [0.79, 0.84]
τ	0.50 [0.47, 0.53]	0.81 [0.79, 0.84]
π_1	0.49 [0.45, 0.52]	0.80 [0.78, 0.83]
ρ_{11}	0.52 [0.49, 0.56]	0.83 [0.81, 0.86]
ρ_{22}	0.51 [0.48, 0.54]	0.80 [0.77, 0.82]

Table 4.3: The relative frequency with which 50% and 80% credible interval contained the true parameter value. The numbers in the brackets correspond to the 95% Jeffreys credible interval for binomial proportion (L. D. Brown et al., 2001).

observe that the credible intervals would be poorly calibrated with respect to their frequentist properties. It is important to keep in mind, though, that this is not a proof of well calibrated CIs in general (e.g., for all possible parameter values and all confidence levels).

4.3.5 Conclusion

We followed general recommendations for a principled Bayesian workflow for building and validating bespoke cognitive models (Kennedy et al., 2019; Schad et al., 2019; Tran et al., 2020). Knowledge about data typical in two-choice speeded decision tasks was used to define the prior distributions on the model parameters. The MCMC procedure yielded accurate approximations of the posterior distributions using simulation-based calibration. SBC further yielded good results except for three parameters for which slight bias could have potentially occurred. Model sensitivity analysis revealed that the model is able to learn about the parameters of interest while not introducing substantial systematic bias to the estimates. The standard parameter recovery resulted in acceptable results. Further, the 50% and 80% credible intervals had coverage probabilities at their nominal levels. Results of the simulation study hence suggest that fur-

ther work on improving the model is not absolutely necessary before applying it to real data.

4.4 Application: Dutilh et al. (2011) study

This section demonstrates the use of our model on a real data set from an experiment reported by Dutilh et al. (2011). In this experiment, 11 participants took part in a lexical-decision task (participants A–C in Experiment 1a and participants D–G in Experiment 1bL) and perceptual decision task (participants H–K in Experiment 1bV). Despite the fact that the experiments are based on a different modality, the analysis stayed the same as the data have the same structure regarding the application of the HMM. Specifically, participants were asked to give answers on a two-choice task with varying degrees of pay-off for response time and response accuracy: the sum of the pay-off was a given constant, but the difference between them varied, thus leading to trials preferring accuracy (high reward for getting the answer correctly) to trials preferring speed (high reward for responding fast). Dutilh et al. (2011) originally fitted a two state HMMs where the emission distribution for the response times was assumed log-normal, and the distribution for the responses a categorical (i.e., assuming independence of response times and accuracy after conditioning on the state). Here, the EAM HMM model is applied to each of the participants separately, and the model fit is assessed using posterior predictives.

4.4.1 Method

We fitted each participants' data using the model described in section 4.2 and priors developed in section 4.3.1. Specifically, for each participant, we ran eight MCMC chains with a 1,000 warmup and 1,000 sampling iterations using Stan (B. Carpenter et al., 2017), with the tuning parameter δ_{adapt} increased to 0.9. Starting points were randomly generated from the prior. Some initial values yielded likelihoods that were too low, leading to failure of the chain initialization. If seven out of the eight chains failed to initialize, the model was reran. If at least two chains managed to run, we inspected the Gelman-Rubin potential scale reduction factor $\hat{R}$ (Gelman & Rubin, 1992), traceplots of the MCMC

chains, and parameter estimates, to detect possible label switching. Label switching was identified if $\bar{\alpha}^{(1)} < \bar{\alpha}^{(2)}$ or $\bar{\nu}_1^{(1)} < \bar{\nu}_2^{(2)}$ (the two conditions coincided in 100% of the cases). If label switching occurred, we reran the eight chains. Once we were able to run at least two chains without label switching, we proceeded to fit data from another participant.

4.4.2 Results

Model fit for two participants needed to be run three times and for one participant five times due to seven chains failing to initialize. Further, models needed to be rerun twice for one participant and three times for four participants due to between chain label switching. The final fits for two participants ended with two valid chains, for six participants with three valid chains, and for three participants with four valid chains. Therefore, the number of posterior samples used for inference ranged between 2,000 and 4,000. None of the models yielded divergent transitions. All $\hat{R}$ statistics range between 0.99 and 1.01, and traceplots of the MCMC chains show typical caterpillar shape without a visible drift. Thus, the final model fits do not exhibit convergence issues.

For each participant, we performed several fit diagnostics, to assess whether (and how) the model misfits the data. In the interest of brevity, results for only the first participant from each of the sub-experiments are shown (i.e., participant A, participant D, and participant H). The rest of the results can be found online at `github.com/Kucharssim/hmm_slba/tree/master/figures`.

First, we simulated the posterior predictives for response times and accuracy and plotted them against the observed data. Figure 4.9 shows the posterior predictive distribution for the response times summarised as 80% and 50% quantiles of the posterior predictive distribution for each trial (light red and dark red, respectively), and the median of the posterior predictive distribution (red line). The black line shows the observed response times at a particular trial. Figure 4.10 shows the posterior predictive distribution for the responses. Specifically, the red line shows the predicted probability of a correct response for a particular trial, whereas the black dots points the observed responses. For ease of the visual comparison, the observed responses were smoothed by calculating their moving average with a window of 10 trials, which is shown as a black line.

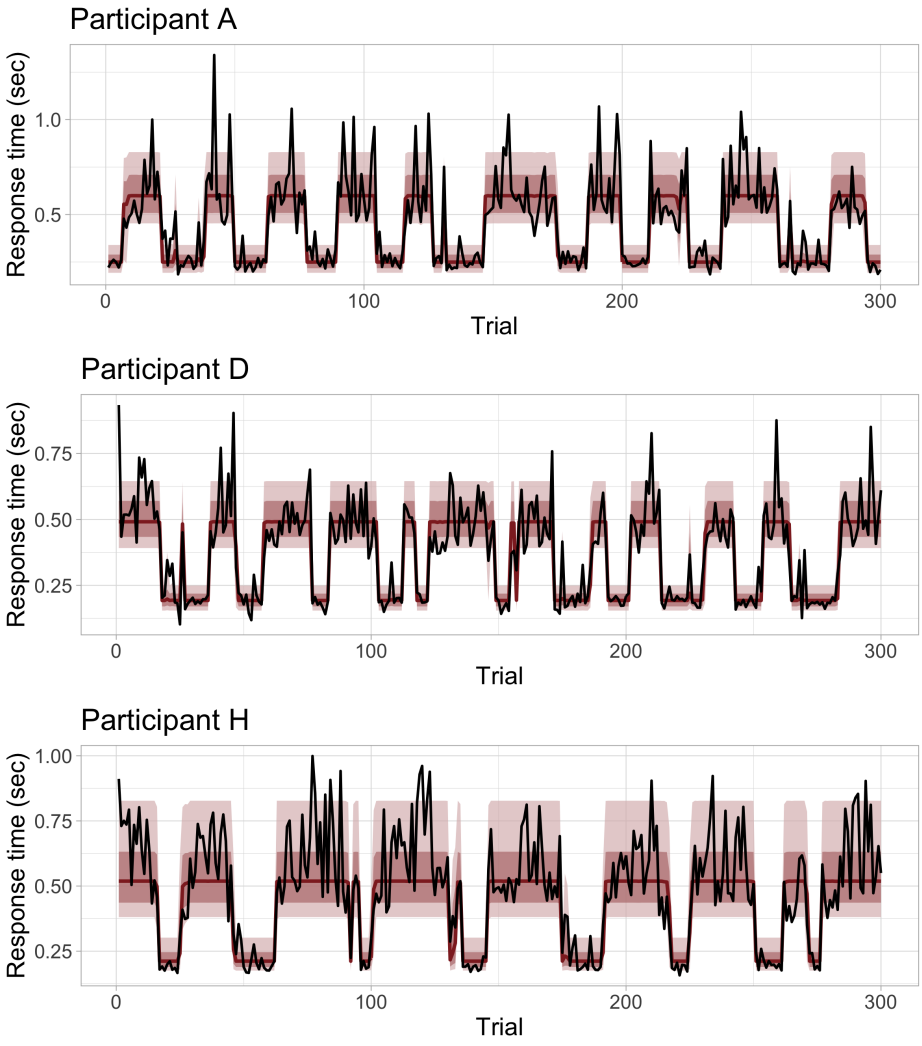


Figure 4.9: Posterior predictives for the response times for three participants. Only the first 300 trials are shown.

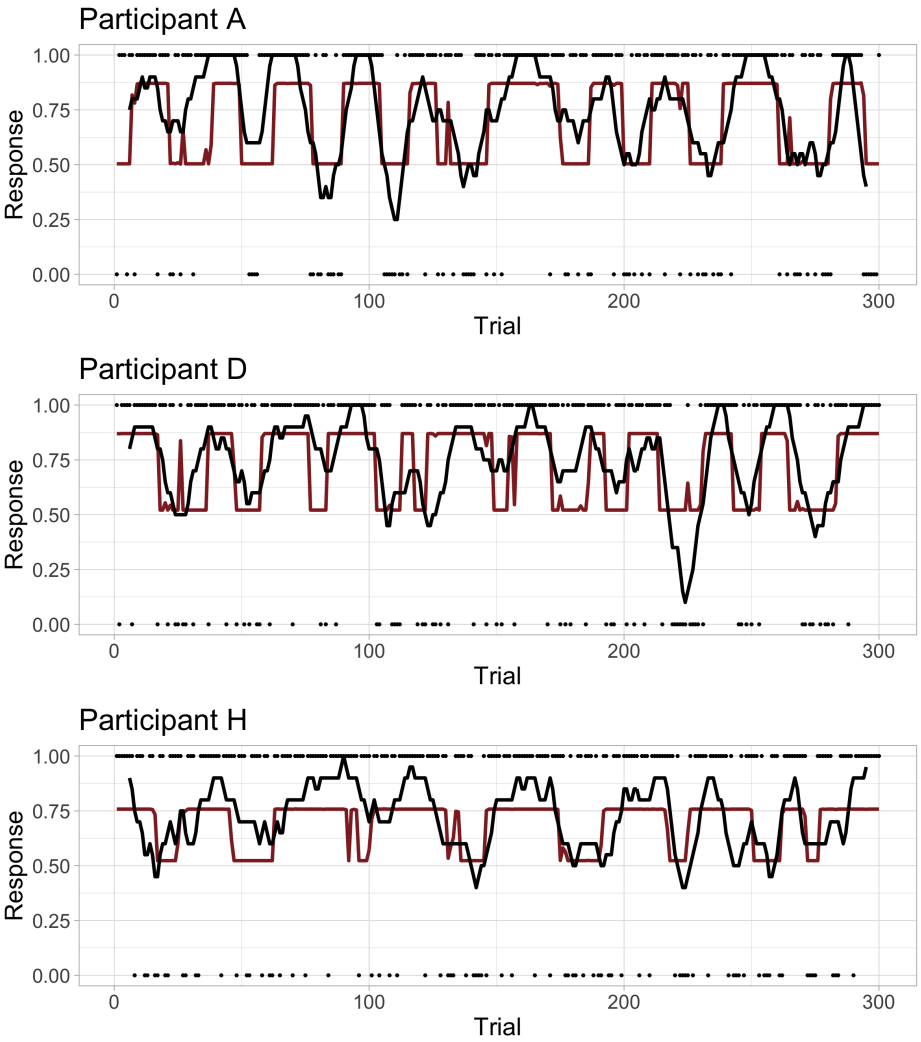


Figure 4.10: Posterior predictives for the responses for three participants. Only the first 300 trials are shown.

In general, the posterior predictives capture the observed data well. Specifically, the model is able to replicate the bi-modality of the response times and captures the runs of trials with predominantly correct responses relatively well. The model also seems to capture correctly that the response times under the guessing (fast) state have smaller variance than under the controlled state. However, for some participants, there seem to be many outliers (i.e., slow responses) that are not predicted by the model, suggesting that the model of the response times has perhaps tails that are too thin.

We also assessed how well the model predicts the response time distributions for correct and incorrect responses. Figure 4.11 shows the observed response times of the correct and incorrect responses as histograms, overlaid with the predicted density of the response times — shown as a black line and 90% CI band. Further, the blue and red lines show the densities under the guessing and controlled state, respectively. Figure 4.12 shows the observed and predicted cumulative distribution functions conditioned on the state and response.

The distribution plots show good model fits, as the bi-modality of the response times is captured correctly, as well as the proportions of correct and incorrect answers under the states. However, for some participants, there are clear signs of a slight misfit. For example, the predicted distribution of the response times of incorrect answers under the controlled state is shifted slightly to the right compared to the empirical distribution (this shift is the most visible for participant H). Further, there is a general tendency of the model to overestimate the variance of the response times under the guessing state, which might be a consequence of equating the standard deviation of the drift rate (σ) across all accumulators and states. Another possibility would be to enable bias, by setting different decision boundaries for each of the accumulators. These alterations to the model would increase its flexibility and should be validated using simulations - therefore, such additions should be the focus of future projects. In general, the tendency of the model to imply slightly slower incorrect responses than the data suggests, could be also caused by the fact that the number of incorrect responses under the controlled state is low, generally about 10% of the trials (see Figure 4.12). It is possible that the likelihood is then dominated by the distribution of the correct responses and the distributions of the responses under the guessing state, thus favoring a better fit towards them.

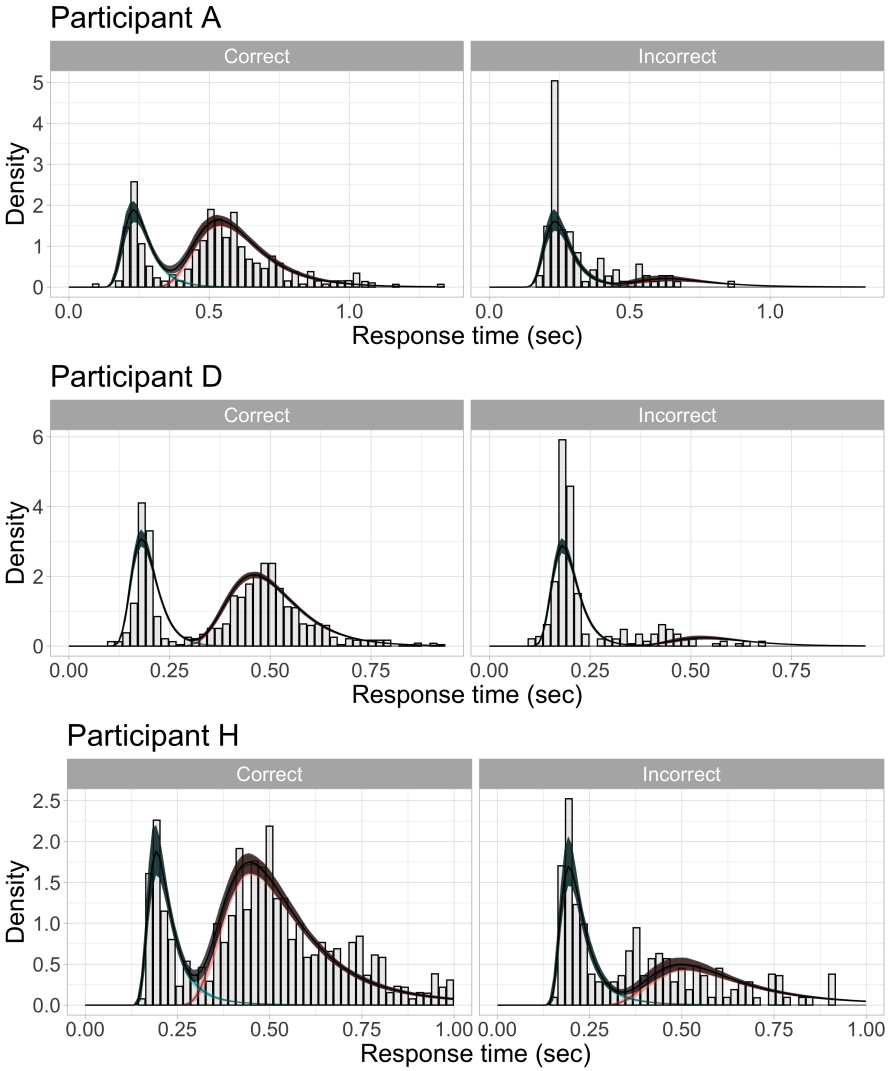


Figure 4.11: Observed and predicted response times distribution of correct and incorrect responses.

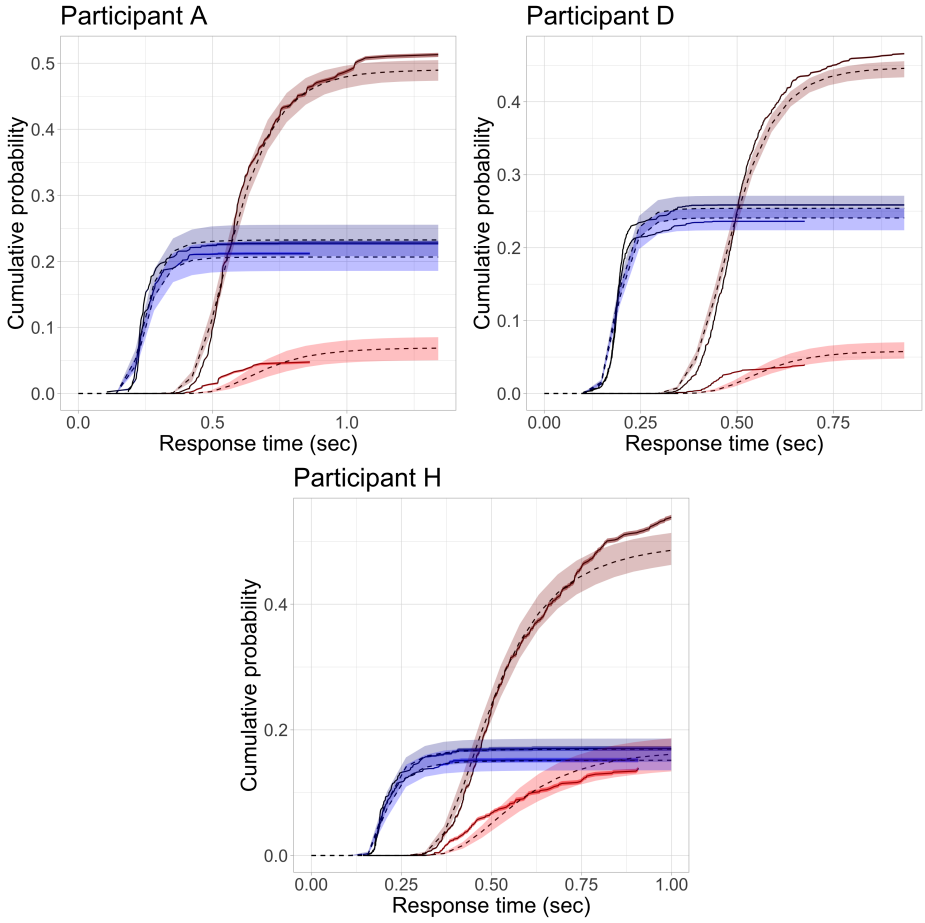


Figure 4.12: Observed and predicted cumulative distribution conditioned on the state (blue=guessing, red=controlled) and response (dark=correct, light=incorrect)

Parameter estimates for each participant are attached in Appendix 4.B. Posterior contraction for all participants was close to one for most of the parameters, indicating that there occurred substantial updating of the priors through the observed data, in line with the simulation results which showed strong updating of priors despite relatively modest number of trials ($n = 200$) in the simulations. An exception was the parameter π_1 which does not update much, a result that was expected following the simulation results as well. Although there seems to be variability between participants' parameter estimates, there are common patterns that to some degree apply to all participants. Generally, the states of the HMMs are sticky, with a probability of remaining in the current state at about 90% of the trials for both of the states. This percentage is (likely) dependent on the experimental design of (Dutilh et al., 2011) who varied the pay-off balance in a structured way depending on the participant's actions, and should not be interpreted as a general tendency of people to stick in the current state to exactly this extent.

As for the parameters that were held fixed across states and accumulators, the non-decision time τ is negligible for the majority of participants; the longest non-decision time occurred for participant B with about 0.11 sec (110 msec), with some participants as short as about 0.01 sec (10 msec). Non-decision time is largely informed by the fastest responses in the data (i.e., the shortest response time gives the upper bound of the parameter). It is possible that loosening up equality constraint between the states would reveal that non-decision time is larger under the controlled state than under the guessing state, representing additional encoding time and executing a motoric response after a decision is made; which could also slightly improve the model fit especially regarding the relatively more variable response times under the controlled state. Relatively surprising were the values of the standard deviation of the drift rates σ, with posterior means ranging between 0.13 and 0.27 — quite smaller than specified by the priors ($\sigma \sim \text{Gaussian}(0.4, 0.1)_{(0,\infty)}$) — suggesting that the variability of the response times is smaller than implied by the prior. Future studies should pay specific attention to variability of the response times in prior predictive simulations.

Shorter response times in the actual data compared to the prior predictive expectations resulted also in a relative mismatch between the prior settings for

the decision boundaries under the two states. Specifically, the posterior means of the decision boundary under the controlled state ranged between 0.24 and 0.37 (whereas the prior was set $\alpha^{(1)} \sim \text{Gaussian}(0.5, 0.1)_{(0,\infty)}$). The posterior means of the decision boundary under the guessing state was as low as between 0.08 and 0.18 (prior $\alpha^{(2)} \sim \text{Gaussian}(0.25, 0.05)_{(0,\infty)}$).

As expected, the average drift rate of the correct response under the guessing state is usually very close to 0.5, implying 50% accuracy. Under the controlled state, the posterior mean of the average drift rate of the correct response ranged between 0.58 – 0.65. This is slightly smaller than the prior expectation (which on average expects about 0.7), although it still leads to relatively high accuracy (at minimum 75%, and leading to accuracy as high as 90%) due to the small standard deviations of the drift rates.

Thanks to the fact that our model is an EAM model, it is possible to inspect the pattern of the discontinuous speed-accuracy trade-off within and between participants in terms of the latent cognitive parameters that control speed of the evidence accumulation (ν) and the response caution (α). Figure 4.13 shows this between state trade-off and reveals striking similarity between participants.

4.5 Conclusion & Discussion

This article presented a robust implementation of a model that combines an EAM with an HMM structure. To our knowledge, this is the first successful implementation combining both structures in one model. The model was built to capture the two state hypothesis following from the phase transition model of the speed-accuracy trade-off (Dutilh et al., 2011) — that there is a guessing and a controlled state between which participants switch. This hypothesis can be represented by an HMM structure. Compared to previous HMM applications on speeded-decision tasks, our model uses an EAM framework for the joint distributions of the responses and response times, and thus enables inference on latent cognitive parameters, such as response caution or drift rate (N. Evans & Wagenmakers, 2019).

The model was validated using extensive simulations and by applying it to real data. The simulations suggested that the model implementation was robust and did not show pathological behavior. Further, the model achieved

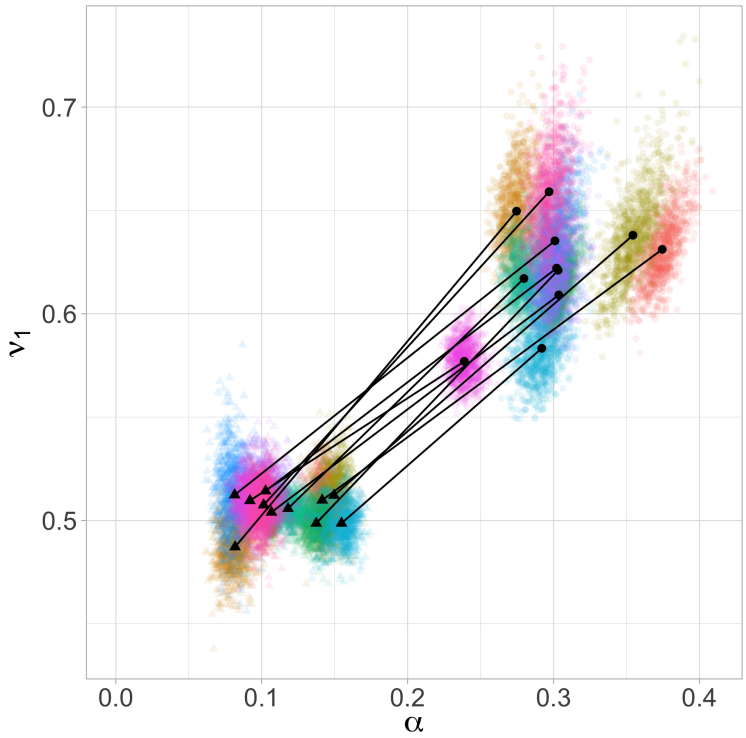


Figure 4.13: Speed-accuracy trade-off for all participants in the Dutilh et al. (2011) data set. Black dots show the posterior mean of each participants' decision boundary ($\alpha^{(1)}$) and drift rate for the correct response ($\nu_1^{(1)}$) under the controlled state, triangles the same but under the guessing state. Lines connect the posterior means for separate participants. Colored points show the samples from the joint posterior distributions.

good parameter recovery and coverage probabilities of the credible intervals. In the empirical example, the model was fitted to eleven participants who partook in the Dutilh et al. (2011) study. The results demonstrate that the model shows a good fit to the data and is able to capture most of the patterns in the data. However, the model also showed a slight systematic misfit because the predicted error responses under the controlled state were slower than that of the data (a typical example of a phenomenon known as fast errors; Tillman & Evans, 2020). The results suggested quite strong consistency between participants in terms of the speed-accuracy trade-off — suggesting that the inac-

cessibility region (i.e., a region of speed of accumulation and response caution which "cannot be accessed", resulting in switching between two discrete states) predicted by the phase transition model could be qualitatively similar across participants (see Figure 4.13).

We used a full Bayesian framework in this article, and with it comes the perks of defining the prior distributions on the parameters. Setting well behaved priors is important in any Bayesian application as they define the subset of the parameter space that generates data that are expected in a particular application of the model. Because the EAMs can cover a lot of heterogeneous experimental paradigms (with heterogeneous scales of the data), it is important to decide on priors in respect to the specific application of the model, preferably after consulting related research literature, careful reasoning about the experimental design and the particular parameterization of the model. The empirical analysis pointed to some discrepancies between empirical parameter estimates and their priors that highlight misalignment between the priors and the data. Ideally, such discrepancies would be minimized to avoid a prior-data conflict (possibly leading to problems with estimation, Box, 1980; M. Evans & Moshonov, 2006). In our application, the discrepancy between the priors and the data arose mainly because we apriorily expected longer and more variable response times than was the case in the Dutilh et al. (2011) study. For the purpose of model validation through extensive simulation, such discrepancy is not a critical problem as the simulation covered cases with potentially more variability and outliers (which usually cause problems in fitting), thus exposing the model to a robustness test.

It is important to reiterate that the priors in this model also serve another purpose: to solve the label switching problem. As is commonly the case in HMMs, the current model is identified only up to the permutation of the state labels. The priors in this article were used to nudge the model towards one specific permutation — to associate the first state with the controlled response, and the second state with the guessing response. Such use of the priors was possible because we specifically assumed the controlled and guessing state, and followed the implications from the theory about them (Dutilh et al., 2011). In case the expectation regarding the state identity is more vague (e.g., when expecting only that the distributions might be multimodal), such use of priors

becomes much more problematic on both the conceptual and practical level. On the other hand, some prior specifications could have been even more informative in the current application. For instance, under the controlled state, the drift rate for the correct response should be higher than the drift rate for the incorrect response as the other alternative would imply that the respondent's performance is below chance level.

Despite our efforts to solve label switching using informative priors, the issue of switching labels still persists, albeit to a lesser degree than without informative priors. Specifically, the use of soft order constraints (by specifying prior distributions that heighten prior probability of a specific state configuration) does not ensure that the labels do not switch at all. To this end, we were forced to perform additional checks of label switching to ensure that the model converged to the solution we preferred, and refitting the model if if did not. Virtually the same estimation results would have been obtained if traditional order constraints were used, by effectively truncating the parameter space to the region which corresponds to the appropriate state interpretation, although in case one would want to perform model comparison using marginal likelihoods, the decision of whether or not to use order restriction would make a difference. Implementing order restrictions would also make it harder to reason about theoretically justified prior specification. For the sake of simplicity, this article did not focus on developing such approach, as its focus was to demonstrate the possibility of combining EAMs with HMMs at least in estimation context. Developing proper ways how to identify the model using order constraints, set reasonable priors, and compute marginal likelihoods would be additional ways how to take the current modeling framework towards more general applications.

One of the future applications would be to actually put the continuous and discontinuous debate under a test. In this article, we presented a model that assumes both discontinous, between state trade-off, and continuous, within state trade-off inherent to the EAM. Utilizing Bayesian framework makes it naturally attractive to use marginal likelihoods to compare simple EAMs, HMM combined with an EAM, and a HMM that assume local independence of response times and accuracy, to assess which of the hypotheses are supported by the data. Although methods for estimating marginal likelihoods for EAMs are

available (N. Evans & Brown, 2018; Gronau, Heathcote, & Matzke, 2019), the HMM extensions will lead to further problems, as estimating marginal likelihoods for finite mixture models and HMMs is a notoriously difficult problem (Frühwirth-Schnatter, 2004). Nevertheless, combining clever constraints (so as to prevent label switching) and development of principled priors would enable the use of efficient techniques for estimating marginal likelihoods such as bridge sampling and its extensions (Gronau et al., 2019; Gronau, Singmann, & Wagenmakers, 2017, 2020; Meng & Wong, 1996), which are now becoming more available than ever. Of course, multi-model inference would also benefit from simulation-based calibration approaches build on similar principles as that of single model inference shown in this article (Schad, Nicenboim, Bürkner, Betancourt, & Vasishth, 2021).

An alternative to identifying the HMMs using the priors is to assume functionally different emission distributions under the states. For example, as Dutilh et al. (2011) point out, it is questionable to assume that guessing requires evidence to make a response. Therefore, using an EAM to represent the guessing state probably leads to model misspecification, as under guessing there is no evidence accumulation (about the correct response). Such misspecification could be fixed, for example, by assuming that the response time of guessing is just a simple response time (Luce, 1991), and model it appropriately by a single accumulator independent of the response (which would be a categorical variable with proportion of correct answer fixed at 0.5). In the context of the phase transition model, such an assumption could further improve the model.

Additional advantages of utilizing Bayesian inference and implementation in Stan is the relative ease with which the model could be extended from single-participant model to multiple-participants model and let the individual parameters be estimated in a hierarchical structure. Hierarchical models have the advantage that they can improve individual estimates by pooling information across the sample. Such approach would also improve the amount of information used for estimating the prior probability of the starting state, which is poorly identified in the single-participant model.

In this article, we used a minimal linear ballistic model to ensure computational stability of the model. However, such a model can hardly be considered adequate for characterizing all phenomena of the speeded-decision paradigm,

and the current results already revealed some ways in which the current model misfits the data. Thus, it is desirable to find ways how to extend or improve the current model, while ensuring that the quality of inferences and implementation does not decline. One alternative to improve the current model is to use the full LBA model where the variability of the starting point is not fixed at zero (S. D. Brown & Heathcote, 2008). Another would be to build on a different evidence accumulation mechanism (such as replacing the ballistic accumulation with sequential sampling models) — for example, the Diffusion Decision model (DDM, Ratcliff & McKoon, 2008) or the Racing diffusion model (Tillman, Van Zandt, & Logan, 2020). Regardless of which framework will be in the end more successful in combination with a HMM, we believe it is important to start with a minimal existing model that captures the most crude phenomena from the speeded-decision framework, and expand from there. In the case of a DDM, that would be to start with the simplest four parameter model because is can be implemented in a fast and robust way (Navarro & Fuss, 2009; Wabersich & Vandekerckhove, 2014) and generally focus on the most important sources of variability at first (Tillman et al., 2020). Then — provided that model validations are satisfactory — it is possible to add more parameters. In each stage of the model building, it is important to stick to the model validation procedures, some of which were demonstrated in the current article.

Further development and additions to the model should probably also be combined with simplifications. Such simplifications, as for example, simplifying the distribution under the guessing state (as discussed above) can provide more computational stability and provide degrees of freedom to extend the model under the controlled state.

The current model provides a proof-of-principle of a combination of an EAM with an HMM, and as such can lead to further interesting applications and extensions, as it opens new possibilities regarding modeling continuous and discontinuous patterns of response times and accuracy in a single modeling framework. Although the current article focused solely on speeded decision tasks, questions about the continuous and discontinuous relations between response times and accuracy is ubiquitous in higher cognitive applications as well, including study of more complex cognitive tasks and development of strategies used to solve these tasks (Hofman, Visser, Jansen, Marsman, & van der

Maas, 2018; Raijmakers, Schmittmann, & Visser, 2014; van der Maas & Jansen, 2003). An interesting feature of higher level cognitive tasks that might be relevant to explore using the current framework is the emergence of more efficient strategies, that lead to qualitatively better response accuracy as well as shorter response times. Such strategies have been described in many applications, such as multiplication tasks (Hofman et al., 2018), Mastermind game (Gierasimczuk et al., 2013; Kucharský et al., 2020), or Progressive matrices tasks (Laurence et al., 2018; Vigneau et al., 2006). Combination of HMM with EAM in this context would enable uncovering different relations between response times and accurate depending on whether we look within or between strategies — it is possible to imagine that an efficient strategy would be faster and more accurate than less efficient strategy, but within those strategies separately, we will see the traditional speed-accuracy trade-off whereby increasing response caution increases accuracy at the cost of speed, which would be captured by the EAM part of the model.

Open Practices Statement

The code and data used in this article are publicly available at `github.com/Kucharssim/hmm_slba`. The analyses were not preregistered.

Appendix

4.A Derivation of the simplified LBA model

Here, we provide the derivation of the likelihood function for the simplified LBA model. We assume that each choice option is associated with an accumulator of evidence. These accumulators are independent of each other and the first accumulator that reaches its decision threshold launches the decision associated with it. This leads to general race equations (Heathcote & Love, 2012), the probability density of observing response a with the reaction time rt comprises of the probability density that an accumulator associated with response a hits the threshold at time rt times the probability that none of the other accumulators has hit the threshold at an earlier time point:

$$\text{sLBA}(\text{rt}, \text{a}|\nu, \sigma, \alpha, \tau) = f(\text{rt}|\nu_a, \sigma_a, \alpha_a, \tau_a) \times \prod_{k \neq a} \left[1 - F(\text{rt}|\nu_k, \sigma_k, \alpha_k, \tau_k)\right], \tag{4.4}$$

with ν_a the mean drift rate, σ_a the standard deviation of drift rate, α_a the decision boundary, and τ_a the non-decision time for the accumulator a.

The density of the passage time for each accumulator $f(\text{rt})$ is specified as follows:

$$\begin{aligned} \text{rt} &= \tau + t \\ t &= \frac{\alpha}{\delta} \\ \delta &\sim \text{Gaussian}(\nu, \sigma)_{(0,\infty)}. \end{aligned} \tag{4.5}$$

We assume that the passage time is a sum of the non-decision time and the decision time t, where the decision time is a result of a linear rise of evidence towards

a decision threshold α, at a drift rate δ drawn randomly from a Gaussian distribution with mean ν and standard deviation σ, truncated at 0 on the lower bound. The truncation is assumed because we do not allow for the possibility of a non-response (i.e., that all drifts in a particular trial are negative, thus never cross the decision threshold). We do not assume any randomness in the parameters τ, α, ν and σ, hence, the only missing piece in deriving $f(\mathrm{rt})$ is the change of variables $\mathrm{rt} = \tau + \alpha/\delta$.

First, we derive the density of the latent drift (δ), which is defined as a truncated normal distribution for $\delta \geq 0$ and zero otherwise:

$$g(\delta|\nu,\sigma) = \frac{1}{\sigma} \times \frac{\phi\left(\frac{\delta-\nu}{\sigma}\right)}{1-\Phi\left(\frac{-\nu}{\sigma}\right)}, \tag{4.6}$$

where $\phi(.)$ is the pdf and $\Phi(.)$ the cdf of the standard normal distribution, respectively.

Next, we determine the density of the variable t, which arises as a scaled reciprocal truncated normal variable for $t \geq 0$ and zero otherwise (see also Nakahara et al., 2006):

$$h(t|\nu,\sigma,\alpha) = \frac{\alpha}{t^2} \times g\left(\frac{\alpha}{t}|\nu,\sigma\right) \tag{4.7}$$

Finally, to obtain the density of the passage time rt, we shift the distribution of the decision time t by τ, which results in the following pdf:

$$f(\mathrm{rt}|\nu,\sigma,\alpha,\tau) = h(\mathrm{rt}-\tau|\nu,\sigma,\alpha) = \frac{\alpha}{(\mathrm{rt}-\tau)^2} \times g\left(\frac{\alpha}{\mathrm{rt}-\tau}|\nu,\sigma\right), \tag{4.8}$$

for $\mathrm{rt} > \tau$ and zero otherwise.

The cumulative probability function of the passage times, $F(\mathrm{rt}|\nu,\sigma,\alpha,\tau)$, is relatively easier to compute, by realizing that the only source of randomness in this model is the distribution of the latent drift δ. Thus,

$$\begin{aligned} P(\mathrm{rt} \leq X) &= P(\delta \leq Y) \\ Y &= \frac{\alpha}{X-\tau}, \end{aligned} \tag{4.9}$$

which leads to

$$F(\text{rt}|\nu, \sigma, \alpha, \tau) = G\left(\frac{\alpha}{\text{rt} - \tau}|\nu, \sigma\right), \tag{4.10}$$

where $G(.|\nu, \sigma)$ is the cdf of a normal distribution truncated at zero.

Identifiability and a minimal model

If we had only response time data without choices (e.g., from a single choice response time task), the entire likelihood would be given by the distribution of the passage times for a single accumulator $f(\text{rt}|\nu, \sigma, \alpha, \tau)$. Such distribution is a ballistic analogue to the shifted Wald distribution (otherwise known as inverse Gaussian distribution) of response times (Anders et al., 2016; Chhikara & Folks, 1988), and would similarly require fixing one of the parameters ν, σ, or α to achieve identifiability.

Once we have multiple choice tasks, it is possible to estimate more parameters per accumulator, as is the case for the LBA (S. D. Brown & Heathcote, 2008). However, some identifiability constraints still need to be put in place. In this paper, we use the following set of identifiability constraints:

$$\sum_i \nu_i = 1,$$
$$1 \geq \nu_i \geq 0.$$

That is, we use the sum-to-one constraint common for the LBA model (S. D. Brown & Heathcote, 2008; Visser & Poessé, 2017), and make it even slightly more severe by assuming that no average drift rate can be negative. The second, additional constraint is convenient for Bayesian implementation as it allows using Dirichlet priors on the drifts.

The simplified LBA model can be achieved by additionally assuming that the non-decision time is equal between the accumulators – usually EAM models assume that non-decision time is by definition the time spend on processes that are not related to the decision – such as encoding and executing motoric responses (N. Evans & Wagenmakers, 2019). Further, we may equate σ and α between the accumulators. The minimal model for a two choice task would

then contain five parameters: $\theta = (\nu_1, \nu_2, \sigma, \alpha, \tau)$, of which four of them are "free" (ν_1 and ν_2 are collinear due to the sum-to-one constraint). In general, the simplified LBA model would have $K + 3$ parameters (of which $K + 2$ are free), where K is the number of response options (accumulators).

4.B Parameter estimates of the Dutilh et al. (2011) data

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.63	0.63	0.02	0.61	0.66	1.001	3319	2939	0.975
$\nu_1^{(2)}$	0.51	0.51	0.01	0.49	0.53	1.000	4090	2860	0.992
$\alpha^{(1)}$	0.37	0.37	0.01	0.36	0.39	1.003	2540	2191	0.991
$\alpha^{(2)}$	0.14	0.14	0.00	0.13	0.15	1.002	2069	2483	0.992
σ	0.16	0.16	0.01	0.15	0.18	1.000	2250	2672	0.991
τ	0.01	0.01	0.01	0.00	0.02	1.003	1602	1690	0.999
π_1	0.46	0.46	0.15	0.22	0.70	1.001	4497	2559	0.051
ρ_{11}	0.92	0.92	0.02	0.88	0.95	1.001	4483	2909	0.973
ρ_{22}	0.89	0.90	0.02	0.85	0.93	1.002	3901	2381	0.960

Table 4.4: Descriptives of the posterior draws for Participant A from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.65	0.65	0.02	0.62	0.68	1.004	1837	1704	0.960
$\nu_1^{(2)}$	0.49	0.49	0.01	0.47	0.51	1.000	3065	2167	0.990
$\alpha^{(1)}$	0.27	0.27	0.01	0.26	0.29	1.000	1979	1934	0.994
$\alpha^{(2)}$	0.08	0.08	0.01	0.07	0.09	1.005	1168	1061	0.978
σ	0.18	0.18	0.02	0.16	0.21	1.003	1271	1370	0.975
τ	0.11	0.11	0.01	0.08	0.13	1.005	1127	1063	0.996
π_1	0.45	0.45	0.14	0.22	0.70	1.001	3029	1997	0.070
ρ_{11}	0.90	0.90	0.02	0.87	0.93	1.001	3038	1897	0.978
ρ_{22}	0.84	0.84	0.03	0.80	0.89	1.001	3049	2364	0.946

Table 4.5: Descriptives of the posterior draws for Participant B from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.64	0.64	0.02	0.61	0.68	1.001	2190	1837	0.953
$\nu_1^{(2)}$	0.51	0.51	0.01	0.49	0.53	1.000	2883	2091	0.987
$\alpha^{(1)}$	0.35	0.35	0.01	0.34	0.37	1.002	1985	1831	0.986
$\alpha^{(2)}$	0.15	0.15	0.01	0.14	0.16	1.001	1693	1564	0.984
σ	0.17	0.17	0.01	0.15	0.19	1.001	2022	1734	0.984
τ	0.01	0.01	0.01	0.00	0.03	1.002	1358	1622	0.998
π_1	0.46	0.46	0.14	0.23	0.69	1.001	3171	2226	0.120
ρ_{11}	0.91	0.92	0.02	0.88	0.94	1.000	3279	1883	0.968
ρ_{22}	0.87	0.88	0.03	0.82	0.92	1.002	2925	2082	0.937

Table 4.6: Descriptives of the posterior draws for Participant C from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.61	0.61	0.01	0.60	0.62	1.000	2911	2213	0.994
$\nu_1^{(2)}$	0.50	0.50	0.00	0.50	0.51	1.004	3268	1746	0.998
$\alpha^{(1)}$	0.30	0.30	0.00	0.30	0.31	1.000	2889	1793	0.998
$\alpha^{(2)}$	0.11	0.11	0.00	0.10	0.11	1.001	1391	1591	0.999
σ	0.13	0.13	0.00	0.12	0.14	1.000	2095	2116	0.998
τ	0.00	0.00	0.00	0.00	0.01	1.001	1131	1488	1.000
π_1	0.54	0.54	0.15	0.29	0.78	1.000	3930	2281	0.027
ρ_{11}	0.90	0.90	0.01	0.88	0.92	1.000	3998	2251	0.987
ρ_{22}	0.90	0.90	0.01	0.88	0.92	1.000	3513	1906	0.987

Table 4.7: Descriptives of the posterior draws for Participant D from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.62	0.62	0.01	0.60	0.65	1.001	2303	2036	0.978
$\nu_1^{(2)}$	0.50	0.50	0.01	0.49	0.51	1.000	2858	2045	0.996
$\alpha^{(1)}$	0.30	0.30	0.01	0.29	0.32	1.000	1530	1782	0.994
$\alpha^{(2)}$	0.14	0.14	0.01	0.12	0.15	1.001	957	1785	0.983
σ	0.15	0.14	0.01	0.13	0.16	1.001	1458	1674	0.990
τ	0.02	0.01	0.01	0.00	0.04	1.001	899	987	0.997
π_1	0.46	0.45	0.14	0.23	0.70	1.000	2769	1768	0.079
ρ_{11}	0.85	0.85	0.02	0.80	0.88	1.002	2862	1848	0.959
ρ_{22}	0.85	0.85	0.02	0.81	0.89	1.000	2668	1749	0.957

Table 4.8: Descriptives of the posterior draws for Participant E from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.62	0.62	0.01	0.60	0.64	1.002	1999	2295	0.984
$\nu_1^{(2)}$	0.51	0.51	0.01	0.50	0.51	1.001	3617	2235	0.998
$\alpha^{(1)}$	0.28	0.28	0.01	0.27	0.29	1.003	1206	1413	0.994
$\alpha^{(2)}$	0.12	0.12	0.01	0.11	0.13	1.004	893	803	0.975
σ	0.16	0.16	0.01	0.14	0.18	1.003	1023	974	0.987
τ	0.05	0.05	0.01	0.02	0.07	1.004	874	815	0.995
π_1	0.45	0.45	0.14	0.23	0.70	1.004	2860	1943	0.102
ρ_{11}	0.91	0.91	0.01	0.88	0.93	1.002	2486	1753	0.986
ρ_{22}	0.91	0.91	0.01	0.89	0.93	1.001	2647	1798	0.988

Table 4.9: Descriptives of the posterior draws for Participant F from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.58	0.58	0.01	0.56	0.61	1.001	3076	2370	0.977
$\nu_1^{(2)}$	0.50	0.50	0.01	0.48	0.51	1.000	2903	2069	0.993
$\alpha^{(1)}$	0.29	0.29	0.01	0.28	0.31	1.001	1334	1996	0.990
$\alpha^{(2)}$	0.15	0.16	0.01	0.14	0.17	1.001	1109	1461	0.978
σ	0.17	0.17	0.01	0.15	0.19	1.000	2029	1885	0.986
τ	0.02	0.01	0.01	0.00	0.04	1.001	1049	1069	0.997
π_1	0.46	0.46	0.14	0.23	0.69	1.001	2437	1881	0.114
ρ_{11}	0.89	0.89	0.03	0.84	0.93	1.000	2661	2149	0.953
ρ_{22}	0.88	0.89	0.03	0.84	0.93	1.001	2175	2087	0.946

Table 4.10: Descriptives of the posterior draws for Participant G from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.64	0.63	0.02	0.61	0.67	1.000	2787	2656	0.959
$\nu_1^{(2)}$	0.51	0.51	0.02	0.48	0.54	1.001	3673	2653	0.978
$\alpha^{(1)}$	0.30	0.30	0.01	0.29	0.32	1.000	2982	2630	0.991
$\alpha^{(2)}$	0.08	0.08	0.01	0.07	0.09	1.002	1922	1484	0.977
σ	0.27	0.27	0.02	0.23	0.31	1.001	2001	1784	0.938
τ	0.09	0.09	0.01	0.06	0.10	1.002	1825	1520	0.997
π_1	0.55	0.55	0.14	0.30	0.77	1.002	4024	2054	0.067
ρ_{11}	0.94	0.94	0.01	0.92	0.96	1.002	3678	2633	0.989
ρ_{22}	0.88	0.88	0.02	0.84	0.92	1.003	3683	2612	0.958

Table 4.11: Descriptives of the posterior draws for Participant H from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.62	0.62	0.02	0.60	0.65	1.001	1423	1521	0.968
$\nu_1^{(2)}$	0.51	0.51	0.01	0.50	0.53	1.000	2217	1289	0.991
$\alpha^{(1)}$	0.30	0.30	0.01	0.29	0.32	1.000	1851	1275	0.993
$\alpha^{(2)}$	0.10	0.10	0.01	0.09	0.12	1.001	899	934	0.978
σ	0.26	0.25	0.02	0.22	0.30	1.001	1074	1177	0.944
τ	0.06	0.06	0.01	0.04	0.08	1.001	854	789	0.996
π_1	0.55	0.55	0.15	0.30	0.80	1.000	2496	1225	-0.001
ρ_{11}	0.91	0.91	0.01	0.89	0.93	1.001	2211	1267	0.986
ρ_{22}	0.90	0.90	0.02	0.88	0.93	1.000	2047	1255	0.982

Table 4.12: Descriptives of the posterior draws for Participant I from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.58	0.58	0.01	0.56	0.59	1.000	4004	3489	0.992
$\nu_1^{(2)}$	0.51	0.51	0.01	0.50	0.52	1.002	4176	2785	0.995
$\alpha^{(1)}$	0.24	0.24	0.01	0.23	0.25	1.001	2186	2528	0.996
$\alpha^{(2)}$	0.09	0.09	0.01	0.08	0.10	1.001	1731	1602	0.984
σ	0.18	0.18	0.01	0.16	0.20	1.002	2166	2552	0.988
τ	0.05	0.06	0.01	0.04	0.07	1.002	1674	1606	0.997
π_1	0.45	0.45	0.14	0.22	0.69	1.001	4561	2501	0.103
ρ_{11}	0.94	0.94	0.01	0.92	0.96	1.000	3888	2076	0.991
ρ_{22}	0.89	0.89	0.02	0.86	0.92	1.002	4567	2907	0.977

Table 4.13: Descriptives of the posterior draws for Participant J from Dutilh et al. (2011).

				Quantile			ESS		
Parameter	Mean	Median	SD	5%	95%	$\hat{R}$	Bulk	Tail	Contraction
$\nu_1^{(1)}$	0.66	0.66	0.02	0.63	0.69	1.001	1492	1412	0.953
$\nu_1^{(2)}$	0.51	0.51	0.01	0.49	0.53	1.000	1757	1341	0.990
$\alpha^{(1)}$	0.30	0.30	0.01	0.28	0.31	1.000	1590	1497	0.992
$\alpha^{(2)}$	0.10	0.10	0.01	0.09	0.11	1.002	769	778	0.985
σ	0.21	0.21	0.02	0.19	0.24	1.000	944	1039	0.973
τ	0.04	0.05	0.01	0.03	0.06	1.002	708	725	0.997
π_1	0.46	0.46	0.15	0.22	0.70	1.000	2083	1334	0.040
ρ_{11}	0.91	0.92	0.02	0.88	0.94	1.002	1898	1417	0.977
ρ_{22}	0.92	0.92	0.02	0.89	0.94	1.000	2218	1371	0.978

Table 4.14: Descriptives of the posterior draws for Participant K from Dutilh et al. (2011).

Part II

Addressing Imperfections

Aim for the sky and you'll hit the ceiling. Aim for the ceiling and you'll stay on the floor.

–Bill Shankly

Chapter 5

Bayesian Sample Size Planning for Developmental Studies

This chapter is published as Visser, I., Kucharský, Š., Levelt, C., Stefan, A. M., Wagenmakers, E.-J., and Oakes, L. (2023). Bayesian sample size planning for developmental studies. *Infant and Child Development*, e2412. doi: 10.1002/icd.2412

Abstract

Running developmental experiments, particularly with infants, is often time consuming and intensive, and recruitment of participants is hard and expensive. Thus, an important goal for developmental researchers is to optimize sampling plans such that neither too many nor too few participants are tested given the hypothesis of interest. One approach that enables such optimization is the use of Bayesian sequential designs. The use of such sequential designs allows data collection to be terminated as soon as the evidence is deemed sufficiently strong, without compromising the interpretability of the test outcome. In this tutorial, we illustrate how to plan a Bayesian sequential testing design prior to data collection by the method of Bayes factor design analysis - the Bayesian equivalent of power analysis - and discuss the relevance of this for developmental psychologists. The tutorial provides a step-by-step guide to perform such analyses, and the methods are illustrated using commonly used statistics in a typical infant looking time paradigm such that researchers can easily adapt these methods for their studies.

5.1 Introduction

One myth of infant research is that it is a science of large effects (Oakes, 2017). Small sample sizes have been accepted as the norm, in part, because it was believed that effect sizes are larger than what seems to be the case now, thus, smaller sample sizes appeared acceptable (Schäfer & Schwarz, 2019). Researchers have difficulty collecting data from large samples of infants. Each infant tested reflects hours of time and effort by the research team. Infants are recruited via birth records, hospitals, advertisements in newspapers, or public outreach. Eligible infants have to be scheduled both at a time that is convenient for the infants' family, and at a time where research staff is available. After the session it must be determined whether the infant was actually eligible and whether the data collected are of sufficient quality and quantity to be included. Typically, data from a large minority of the sessions are discarded due to fussiness, low data quality, or too many missing trials for instance. Finally, the usable data may need to be coded and processed before they can be analyzed. Thus, collecting data from infants is hard, and as a result, researchers have historically targeted the smallest reasonable sample size (Oakes, 2017; Peterson, 2016).

Planning for what appears to be the smallest reasonable sample size often yields underpowered studies (Bergmann et al., 2018; Oakes, 2017). For example, as recently as 2015, the modal sample size in published research using looking time methods with infants was about 14 infants per cell (and invariably there are only few trials per infant). Given the observed typical effect sizes in these studies, this sample size does not provide sufficient statistical power (Oakes, 2017). However, the problem of low statistical power is not confined to increased type II error rates; underpowered studies also increase the rate of false discoveries in a research field. That is, when samples are small, this not only increases the chance that an existing effect will not be discovered, it also increases the chance that a detected effect is in fact non-existent (Colquhoun, 2014). As a result, there has been a call for increasing sample sizes in infant research in general. Yet, this one-size-fits-all solution is problematic for researchers who already have difficulty in recruiting and testing infants in their studies. What is required is a way of increasing the likelihood of arriving at informative results without the need to run huge numbers of participants in every single study. In other words, there is a need for efficient sampling plans. In this paper, we show how the tandem of Bayesian Sequential Testing (BST) and Bayes Factor Design Analysis (BFDA) can deliver just that. Beyond the option of efficient sample size planning, applying Bayesian statistics rather than frequentist statistics brings other advantages as well (see e.g. van de Schoot et al., 2014). Although our discussion here is focused on infants, these issues apply similarly to developmental research more broadly, and the methods we introduce and illustrate in the following apply to infant research and any developmental research alike.

5.1.1 Goals & overview

In this tutorial we detail how BST and BFDA can be combined to plan and execute highly efficient studies in infant and developmental psychology research. We first describe how to use BST to analyse data as it comes in, until a predefined criterion is met. Next, we show how BFDA can be used to analyze crucial study characteristics, such as the expected number of participants needed. Combining these methods delivers a powerful tool to optimize sampling plans and increase efficiency in carrying out developmental research. Note that BFDA, the main focus of this paper, is also informative when sequential testing is im-

possible or impractical.

It is important to stress that this type of analysis is useful in settings in which the goal of the research is to test a previously stated hypothesis. The methods are not applicable when the goal of a particular study is merely exploratory. Application of the methods presented in this tutorial requires that the researcher starts with a research question, a hypothesis, and a statistical test that can be used to test the hypothesis. Such a hypothesis can of course be pre-registered (Davis-Kean & Ellis, 2019; Nosek, Ebersole, DeHaven, & Mellor, 2018), and the BFDA can help strengthen the case for the proposed (pre-registered) sampling plan.

We have a number of goals with this tutorial. Our first goal is to briefly introduce basic concepts from Bayesian statistics, provide pointers to further reading for those unfamiliar with these concepts, and provide a gentle introduction to the conceptual framework of BST and BFDA. The second goal is that after studying the material in this tutorial, the interested reader can start applying these methods in their own research. To facilitate this, we first present a step-by-step guide to BST and BFDA, and then illustrate their use in an example study applying a t-test. We provide analysis code in R (R Core Team, 2020) both in text and in Online Supplementary Materials (available at `osf.io/wak9e/`). The analysis code can be easily adapted to the goals of the researcher (see `github.com/nicebread/BFDA/blob/master/package/doc/BFDA_manual.pdf` for the BFDA manual and examples). Our final goal is to discuss the usefulness, applicability, and extension of the presented methods.

Note that there are other tutorials that discuss the use of Bayesian analysis more broadly and BST more specifically in infant and developmental psychology. Notably, Mani et al. (2021) provide an excellent introduction to sequential testing and extensively discuss a number of applications of BST in early word learning in infants. Marsman and Wagenmakers (2017) introduce Bayesian analysis more generally for a developmental psychology audience and also (briefly) discuss BST. van de Schoot et al. (2014) introduces Bayesian analysis more broadly for developmental psychologists. Although for readers familiar with this material the next section may be repetitive, we include an introduction to these topics here to provide the necessary background for readers who are less familiar with those other sources before we start discussing BFDA.

The methods of BST and BFDA are not specific to infant and developmental research. However, the problems of underpowered studies are exacerbated in these types of research for a number of reasons. Infant development and development generally is a period of large change resulting in large individual differences, and it is exactly those that we are after in developmental psychology. The developing cognitive system also results in large intrinsic noise or variability and our measurement instruments often lack precision, again resulting in large variability in the data. Finally, infants and children generally show more exploration behavior, again resulting in larger variability in the data than would be expected when testing adults. For all these reasons, developmental psychology in particular may benefit from adopting these methods.

5.1.2 Bayesian Sequential Testing

When preparing a study, researchers determine their sampling plan. In common frequentist practice, this typically means deciding on a fixed number of participants prior to data collection, usually based on a power analysis or previous studies using similar methods (J. Cohen, 1988). The approach in the Bayesian framework is rather different. Bayesian analysis is all about learning from data, that is, updating our beliefs when faced with new information. This learning process can be repeated indefinitely as additional data comes in (for details, see box 'Bayesian updating'). Therefore, the Bayesian statistical framework allows for flexible sampling plans, and sequential sampling in particular. Instead of deciding on a fixed number of participants to test, a sampling plan in the Bayesian framework means adopting a particular 'stopping criterion'. This often takes the form of deciding on the strength of the evidence that is required to reach a conclusion about the tested hypotheses. The strength of evidence is expressed in a quantity called the "Bayes factor" (Jeffreys, 1961) — see the box 'Bayesian evidence' for details. A stopping criterion could therefore be formulated as: 'stop collecting more data when the Bayes factor reaches the value of 10 or more'. BST thus entails the following steps: i) test a participant, ii) compute the Bayes factor, iii) check whether stopping criteria have been met and stop if that is the case, otherwise repeat the steps. Note that other considerations can go into determining a stopping criterion such as a minimal required sample size, see discussion in Mani et al. (2021); we return to this issue in the

section 5.2.1.

Bayesian updating

Bayesian inference relies on updating our beliefs when presented with new data. For more in-depth introductions we recommend the 2018 special issue from Psychonomic Bulletin & Review (Vandekerckhove, Rouder, & Kruschke, 2018), as well as several textbooks (Hudson, 2021; Lee & Wagenmakers, 2013). In general, Bayes' rule allows one to update plausibility assessments based on predictive success: hypotheses that predicted the data relatively well gain credibility, whereas hypotheses that predicted the data relatively poorly suffer a decline, as can be seen from Bayes' rule (e.g., Wagenmakers, 2020; Wagenmakers, Morey, & Lee, 2016):

$$\underbrace{p(\theta|\text{data})}_{\text{Posterior}} = \underbrace{p(\theta)}_{\text{Prior}} \times \underbrace{\frac{p(\text{data}|\theta)}{p(\text{data})}}_{\text{Updating factor}} . \tag{5.1}$$

Here, θ represents a parameter in the model, for example the true (i.e., population) correlation coefficient between two variables. We can start by quantifying the plausibility of different values of θ values through a prior probability distribution $p(\theta)$. In this distribution, we can assign higher probability density to parameter values that we deem more plausible than to values that we deem less plausible before seeing new data, thereby formulating what we already know about the parameter (e.g., from estimates of a correlation in previous experiments). We call this our "prior". As we observe new data, we update this prior distribution which becomes our posterior distribution. Specifically, values of θ that predicted the data well, increase in plausibility (i.e., obtain a *higher* probability density), and values of θ that predicted the data poorly, decrease in plausibility (i.e., obtain a *lower* probability density). This Bayesian learning process can continue indefinitely. As we collect data, our posterior distribution of observations becomes the prior distribution for the analysis of the next observations.

The core idea of sequential sampling and stopping criteria is already commonly implemented in studies on learning and in particular in habituation studies in infants (see, e.g., Kucharský, Zaharieva, Raijmakers, & Visser, 2022; Oakes, 2017). In those situations, the stopping criterion is applied to a series of trials from an individual. The habituation phase is stopped, and the test phase started, when there is enough evidence to conclude that an infant has habituated, or that a child has learned the task at hand to a sufficient extent. A similar reasoning is possible on the level of the study, where the stopping criterion is applied to data from subsequent participants (rather than subsequent trials) to decide whether there is evidence in support of the tested hypothesis, and hence to stop collecting data (i.e., recruiting new participants). BST can be used for this purpose precisely; determining a good strategy for using and planning BST comes from applying BFDA.

Crucially, conclusions drawn from Bayesian analyses are not dependent on a pre-defined sampling plan, and are immune to optional stopping (Edwards, Lindman, & Savage, 1963; Lindley, 1993; Rouder, 2014). This is unlike frequentists approaches, in which the interpretation of the p-value becomes invalid through data "peeking" during data collection (Armitage, McPherson, & Rowe, 1969), unless the critical test statistic has been appropriately adjusted to preserve the correct error rate under sequential testing (O'Brien & Fleming, 1979; Pocock, 1977; Stallard, Todd, Ryan, & Gates, 2020; van de Schoot et al., 2014). We return to this issue in the Discussion.

BST allows for flexible data collection procedures in which a researcher monitors the accumulating evidence, and stops sampling precisely when compelling evidence for either hypothesis has been reached (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). Instead of deciding a-priori on the number of participants to test, the researcher decides what would be compelling evidence in support of a hypothesis. Thus, stopping rules in sequential Bayes factor designs are determined based on the researcher's definition of strong evidence (see e.g. Kass & Raftery, 1995, for guidelines about strength of evidence). For example, if a researcher deems a Bayes factor of BF=10 compelling for their particular research context, they would terminate data collection once a Bayes factor of $BF_{10} = 10$ (i.e., supporting the alternative hypothesis) or $BF_{10} = 1/10$ (i.e., supporting the null hypothesis) is reached, regardless of how many sub-

jects were tested to achieve that goal. In addition to evidence thresholds, researchers may include other criteria that allow to take into account resources available and thus optimize the balance between the desired evidence to achieve with the current study, and the practical constraints the researchers face when executing the study.

Sequential hypothesis tests are particularly advantageous in research contexts where data collection is costly or time-consuming. They have repeatedly been shown to be more efficient than tests with comparable error rates that fix sample sizes before the experiment (Schönbrodt et al., 2017; Wald & Wolfowitz, 1948). In particular, the sample size required to meet a stopping rule as identified in a BST is smaller than the sample size identified a-priori from a power analysis for the frequentist fixed-N equivalent of the test (Schönbrodt & Wagenmakers, 2018). From a practical perspective, this means that by using BST procedures, conclusions about hypotheses can be reached earlier, which helps to save resources.

Bayesian evidence

Bayesian updating is applicable to hypothesis testing. In psychology, the most popular models are the null hypothesis $\mathcal{H}_0 : \delta = 0$ and the alternative hypothesis $\mathcal{H}_1 : \delta \neq 0$, where δ denotes effect size. Note that for Bayesian inference, a fully specified $\mathcal{H}_1$ model is required. This means that it is not sufficient to simply formulate an inequality, e.g., $\mathcal{H}_1 : \delta \neq 0$, as it is typically done in frequentist inference. Instead, researchers need to make a probabilistic commitment regarding the values that the parameter is likely to take on in the form of a prior distribution, see Box "Bayesian updating". When comparing two hypotheses, it is convenient to apply the odds form of Bayes' rule:

$$\underbrace{\frac{p(\mathcal{H}_1|\text{data})}{p(\mathcal{H}_0|\text{data})}}_{\text{Posterior odds}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{Prior odds}} \times \underbrace{\frac{p(\text{data}|\mathcal{H}_1)}{p(\text{data}|\mathcal{H}_0)}}_{\text{Bayes factor}}, \tag{5.2}$$

In this example, the prior odds reflect the a-priori plausibility of $\mathcal{H}_0$ relative to $\mathcal{H}_1$. As prior odds are often not formally quantified and indi-

vidual researchers may disagree on the prior plausibility of individual hypotheses, they are usually set to 1 (i.e., a position of equipoise), or left unspecified, so that every reader of a Bayesian analysis can insert their own subjective a-priori plausibilities. The focus in Bayesian hypothesis testing is on the quantification of the evidence, that is, the degree to which the data bring about a change from prior to posterior odds. The evidence is the relative predictive performance of $\mathcal{H}_0$ vs. $\mathcal{H}_1$ (i.e., does the evidence favor $\mathcal{H}_0$ or $\mathcal{H}_1$), and it is generally known as the *Bayes factor* (Jeffreys, 1935, 1939, 1961; Kass & Raftery, 1995).

The Bayes factor subscripts indicate what model is in the numerator and denominator of Equation 5.2, so a subscript of 10 means that the model $\mathcal{H}_1$ is in the numerator and the model $\mathcal{H}_0$ is in the denominator. For instance, $\text{BF}_{10} = 3$ indicates that the observed data are 3 times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$. When $\text{BF}_{10} = 0.2$, the observed data are 0.2 times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$. Because people find this difficult to parse, Bayes factors lower than 1 are usually presented after switching numerator and denominator; here we would have $\text{BF}_{01} = 1/0.2 = 5$, so that the observed data can be said to be 5 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$.

Although the Bayes factor quantifies the evidence in a graded fashion, in order to facilitate interpretation and communication, Harold Jeffreys proposed a classification scheme in which Bayes factors from 1 to 3 are termed "weak evidence", Bayes factors from 3 to 10 are "moderate evidence", Bayes factors from 10 to 30 are "strong evidence", Bayes factors from 30 to 100 are "very strong evidence", and those larger than 100 are "extreme evidence" (e.g., Jeffreys, 1961, Appendix B; Lee & Wagenmakers, 2013; Wasserman, 2000).

In summary, the Bayes factor is a relative measure of predictive success. This means that the Bayes factor can be used to discriminate between "absence of evidence" (when Bayes factors are near 1, such that the data do not provide much diagnostic value) and "evidence of absence" (when Bayes factors favor $\mathcal{H}_0$; see also Keysers, Gazzola, & Wagenmakers, 2020). Moreover, when the evidence is found not to be sufficiently

compelling, additional data may be collected in an attempt to sharpen our knowledge concerning the relative predictive performance of the rival hypotheses – a straightforward application of the Bayesian learning cycle.

5.1.3 Bayes Factor Design Analysis

In designing a BST study, the stopping criterion is crucial. The stopping criterion influences the design characteristics of a study. For example, if more evidence is required to stop data collection, say, a Bayes factor of 20 instead of a Bayes factor of 10, it will usually take longer to reach this amount of evidence, leading to studies with larger sample sizes on average. But how long will it take to reach this amount of evidence? And how often will researchers have to terminate data collection because they ran out of resources, that is, reached their maximum number of participants, rather than having obtained conclusive evidence? One way to establish this is to use Bayes Factor Design Analysis (BFDA; Schönbrodt & Wagenmakers, 2018; Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019). By using BFDA, the researcher can gain insights into the effects of their specified stopping criteria *before* running a study, similar to a power analysis investigating the effect of a chosen sample size on statistical power. As in a frequentist power analysis, the (expected) population effect size influences the results of a BFDA, specifically, the expected sample sizes under different stopping rules.

BFDA is a simulation-based method for design analysis. In a simulation, artificial data for virtual participants is generated from a user-defined population under the design of interest (see Arnold, Hogan, Colford, & Hubbard, 2011, for an introduction to simulation-based design analyses). Here, the population mainly refers to the expected effect size, for example, a specific correlation coefficient or Cohen's δ for a difference between two groups or two conditions. The design refers to the type of statistical test as well as the stopping rule that is employed.

By simulating the same scenario many times (a so-called Monte Carlo simulation), methods like BFDA can determine the design characteristics. The idea is that by simulating not one, but many samples under a given design, and

applying the hypothesis testing procedure to each of the samples, a researcher obtains a distribution of plausible results. In a next step, the researcher can further investigate this distribution to determine the design characteristics. In a sequential Bayesian design, the most important design characteristics are the expected sample size, and the probability that the sequential process ends at the upper or lower Bayes factor threshold, respectively. If the Bayes factor at termination shows evidence for the incorrect model (e.g., in favor the null model when the simulated effect size is unequal to zero), this is termed "misleading evidence". By investigating the results of a BFDA, a researcher can identify the probability of misleading evidence for a certain design, sometimes also called the "error rates" of the design. Error rates are one possible design characteristic that can be evaluated using a BFDA, but BFDA is a flexible approach that allows researchers to investigate every aspect of the design they are interested in (Stefan et al., 2019).

Before a study is conducted, researchers can opt to conduct a BFDA to investigate the design characteristics of their planned design. However, it is also possible to conduct multiple BFDAs with different stopping rules in an iterative process. This allows researchers to plan a design that fulfills the requirements of a certain research scenario, for example, in terms of error control or evidence strength (Stefan, Schönbrodt, Evans, & Wagenmakers, 2022).

In the following sections we provide two examples for a BFDA including commented code that can be easily adapted for different research scenarios. For an in-depth tutorial-style overview of BFDA, we refer the reader to Stefan et al. (2019), and to the manual of the `R` package `BFDA` (Schönbrodt & Stefan, 2018).

5.2 Planning for BST and BFDA

There are several steps and decisions that need to be made in order to run a BST and a BFDA, which are discussed in turn below. Please refer to Figure 5.1 for an overview of the steps in this section. Ideally, these decisions are made prior to collecting the data, possibly made completely transparent in a preregistration or registered report (Crüwell, Stefan, & Evans, 2019).

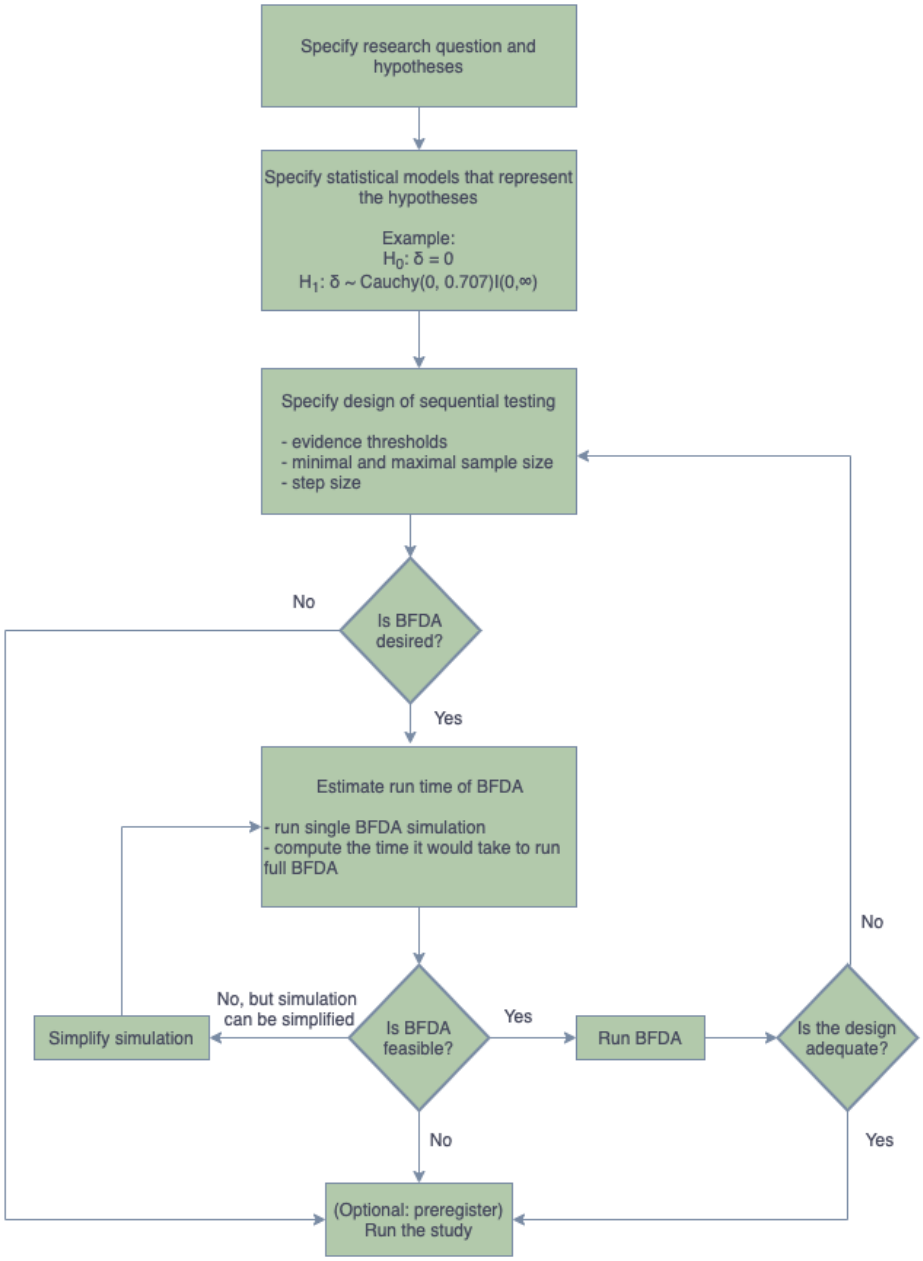


Figure 5.1: Flowchart summarizing the recommended steps when planning a Bayesian sequential study.

5.2.1 BST specifications

Research question and hypotheses

We recommend that researchers carefully set the goal of the study, including the research question they wish to answer, and the competing hypotheses that constitute different theoretical standpoints. When the research goal is more exploratory or the aim is to arrive at precise estimates of an effect size for example, these methods are not applicable or useful. Without clarifying the goals of the study and the hypotheses one wishes to test, there is no reason to run a sequential hypothesis test.

Statistical models that represent the hypotheses

Once the research goals have been set, researchers need to specify the statistical analysis that will be used to answer the research questions. When the goal is to determine whether a particular effect is present or absent (e.g., do infants show a novelty preference? do infants learn a rule?), researchers typically use null hypothesis testing. In the Bayesian framework, null hypothesis testing is achieved by specifying two models: one that represents the null hypothesis, and one that represents the alternative hypothesis. The alternative hypothesis, or alternative model, requires that a prior distribution of the effect of interest be defined to specify the range of plausible values of the effect, under the assumption that the effect exists. The null hypothesis can be specified as a point-null hypothesis, as commonly used in frequentist analyses, or as an interval null hypothesis, where the boundaries of the interval can, for example, be determined by arguments about the smallest effect size that would be of interest (R. D. Morey & Rouder, 2011). Note that BFDA can be used to analyse the sensitivity of the design when altering such factors.

During this stage, it is advisable to research and plan how the analysis will be carried out. Most common analyses (e.g., *t*-test, ANOVA, regression, etc.) are implemented in user-friendly statistical software (e.g., JASP), or come as packages with programming languages (e.g., `BayesFactor` (R. Morey & Rouder, 2018) in R).[1]

[1]For more complicated models one may turn to using probabilistic programming languages, e.g., JAGS (Plummer, 2003) or Stan (B. Carpenter et al., 2017), or use additional soft-

Sequential testing design

The next step is to establish the design of the sequential testing study. First, stopping criteria must be defined. Stopping criteria are any rules that determine whether or not to collect additional data at any specific point in the study. More precisely, they specify the conditions that determine conclusion of the data collection. For example, data collection may be concluded once a certain evidence threshold, say, a Bayes factor of 10 or 1/10, has been reached (Schönbrodt et al., 2017). Evidence thresholds can also be combined. For example, in a case where multiple research questions are of interest, researchers can define that the data collection will end once all of the questions are answered satisfactorily, i.e. reached their respective evidence thresholds.

Stopping criteria may also be informed by practical restrictions and resource constraints. For example, a researcher interested in an effect that is present in a particular age group might only have restricted access to participants in that age group. Sequential Bayes factor designs allow researchers to incorporate these practical restrictions into the planned design. Specifically, it is possible to integrate a maximum sample size in the stopping criteria of the test. For example, when it is impossible or impractical to collect a data set of more than 50 participants, the researcher may decide to stop the data collection when the evidence thresholds have been met, or when a maximum sample size of 50 participants is reached (Schönbrodt et al., 2017). Importantly, when stopping the data collection for whatever reason – whether hitting the evidence threshold or the maximum sample size – the interpretation of the Bayes factor always remains the same: it is the relative evidence for one hypothesis over another, and no adjustment or justification is needed. Although it may be necessary under some circumstances, specifying a maximum sample size has the disadvantage that the evidence accrued when stopping can be non-decisive.

Another practical restriction that may occur is that participant recruitment cannot be stopped immediately when the stopping criterion is met, or that the statistical test cannot be performed after every single participant (Friede & Kieser, 2006). For example, participants may be assigned overlapping time slots

ware and packages that make building special purpose models easier, e.g. the graphical modeling module in JASP (JASP Team, 2021) or use packages that include a wider variety of designs such as the R-package *brms* (Bürkner, 2017)).

in a lab schedule, or data collection, entry, and processing may be bundled for several participants for efficiency reasons. Sequential designs also allow flexible adaptations for such scenarios. Specifically, rather than calculating the Bayes factor after every participant, Bayes factors can be calculated in batches of participants according to how the data comes in. The stopping criteria would then be evaluated after each batch of participants instead of after each individual participant (Schönbrodt et al., 2017). However, calculating the Bayes factor after, say, every 10th participant changes the design characteristics (i.e., the error rates and the resulting expected sample sizes). It is therefore important to be mindful of these design choices when evaluating the design using BFDA. We will provide an example in the following section.

If the data are convincing, stopping thresholds can be hit very early in sequential designs, say, after less than five observations per cell. Some researchers find this discomforting because they aim for a minimum degree of precision in the estimated parameters (Kruschke & Liddell, 2018). Due to sampling variability, an effect size estimate may be unreliable for small samples, and hence one may want to require a minimal sample size. That is similarly the case when individual differences are expected and of interest; to study such individual differences, a minimal sample size is also advisable. Note that this is similarly the case for frequentist sampling plans. BST and BFDA allow for the specification of a minimum sample size in addition to (or instead of) a maximum sample size. The minimum sample size then determines the number of data points that will be collected at minimum, regardless of the strength of the evidence in the data (Kelley & Rausch, 2006; Maxwell, Kelley, & Rausch, 2008; Schönbrodt et al., 2017).

5.2.2 Bayes factor design analysis

The next step is to investigate the characteristics of the selected design by running a BFDA. To specify the design, one must know the expected effect size, the statistical test that is used to test the hypothesis under study, and the stopping criteria in case of a sequential design. To be clear, BFDA is *not required* to run a sequential study design, but it is an optional step that can help reassure researchers that their approach is reasonable given the available resources. In sum, researchers can use a BFDA to determine whether the design needs to be

changed before data collection commences. Note that running a BFDA (e.g. with a fixed n) can still provide useful information about the feasibility of the study also when the design does not involve sequential testing.

Typically, a BFDA would be run such that it reflects the candidate design that is entertained during the planning stage of the experiment. In basic scenarios that rely on running standard statistical tests (e.g., t-tests, test of correlations, etc.), obtaining results of BFDA is straightforward using the `BFDA` package in R (Schönbrodt & Stefan, 2018), or using the BFDA interactive website `shinyapps.org/apps/BFDA/` (Stefan et al., 2019). In situations where the current design or statistical tests are not readily implemented in a dedicated software package, researchers need to set up the analysis themselves. If the specified models are computationally complex, it is possible that running a full BFDA is impractical, as running the analysis would take a prohibitive amount of time and resources. Below, we provide some recommendations for alternative solutions.

Estimating the time needed to run the simulation

The BFDA simulations take time to run, and the amount of time will vary depending on the computer system available, the complexity of the statistical model and the study design. Thus, a good strategy is to first conduct a small number of simulations to get an estimate of the amount of time to run the full BFDA simulation. Consider the following example: A researcher plans to run a sequential study that has a maximum sample size of 100, a stepsize of 1 (so the test will be conducted after each subject), and plans to run 1000 "replications" of the hypothetical experiment under each hypothesis. This BFDA would involve simulating 2000 datasets (1000 for the null hypothesis, and 1000 for the alternative hypothesis). In addition, because each run of the analysis might require computing the Bayes factor up to 100 times (if the maximum number of subjects was reached before the pre-specified Bayes Factor was achieved), running the BDFA could amount to computing the Bayes factor $2000 \times 2 \times 100 = 400,000$ times on different datasets. Clearly, if computing a Bayes factor takes more than couple of seconds, conducting the BFDA becomes prohibitively time consuming.

However, some insights can be gained even if the full BFDA can not be

run. For example, one can decrease the number of hypothetical replications (e.g., from 1000 to 100). This will provide some information, but fewer repetitions decreases the precision of the Monte Carlo results, and therefore make the results of the BFDA less informative for planning the study (A. S. Cohen, Kane, & Kim, 2001); even so, running the BFDA with 100 replications instead of 1000 will of course provide more information than not running it at all; results should always be interpreted taking into account the accuracy of the estimates.

Another alternative is to run the BFDA only for the maximum sample size, instead of simulating the full sequential sampling process. This drastically decreases computation time because the statistical analysis does not have to be re-computed at every step of the sequential process. For example, the BFDA described earlier would require only 2000 evaluations of the statistical test, rather than up to 400,000. This simplified approach will not provide an estimate of the expected sample size or of the error rates of the design. However, it will give an indication of whether or not it is plausible that the study will terminate earlier than at the maximum sample size. For example, if nearly 100 percent of the simulations are highly informative (i.e., large Bayes factors), it is likely that the stopping criteria will be met before the maximum sample is reached. This is a desirable outcome in a BFDA because it shows that it is likely to obtain compelling evidence with the available resources, and that it is highly plausible that the planned experiment will not use up the available resources. If, however, the resulting BFDA shows that even with the maximum sample size it would be unlikely to obtain strong evidence, it might be necessary to re-think the study design.

Of course, it is possible to skip BFDA entirely. BFDA is useful for assessing whether the current study is feasible and to set expectations regarding the study, but it is not a required step for conducting a sequential testing study, as the interpretation of the Bayes factor remains the same regardless of the characteristics of the sampling plan.

5.2.3 Run the study!

After designing the study, and optionally performing the BFDA, data collection may proceed. Typically, a study would be run using the sequential design

that was specified prior to the data collection. However, Bayesian inference in principle allows for continuous updating of knowledge as the evidence provided by the data accrues; strictly speaking, there is no need in Bayesian analyses to pre-specify a sampling plan or stick to the designed sampling plan, because the results of Bayesian analyses can be interpreted independent of the sampling plan. However, it is recommended to be transparent about changes in the design of the study to avoid the suspicion of result hacking (Crüwell et al., 2019).

5.3 Illustration: Rule learning

In this section we illustrate the principles and practices described so far on an example, and present the R-code that is required. The code is available on GitHub and can be downloaded and adapted by researchers wishing to apply the presented methods to their own research designs. To make the examples very concrete, we use the case of rule learning in infancy as a typical developmental study. We first briefly describe the rule learning paradigm and the typical hypotheses that are being tested using the paradigm. These hypotheses are then translated to statistical tests that are analysed using BFDA.

5.3.1 Rule learning in infants

In studies of rule learning, infants are first familiarized with patterned sequences of syllables and then tested whether they can generalize the patterns to sequences built from novel syllables. In a seminal paper, Marcus, Vijayan, Rao, and Vishton (1999) introduced this algebraic rule learning paradigm and showed that 7-month-olds were able to do this. Infants were familiarized with sequences like dewiwi, wejiji, lididi for two minutes. In this case, infants may learn the rule XYY, as each of the sequences involves one syllable followed by a second syllable repeated twice. During test, infants were exposed to congruent sequences (e.g. bapopo), in which the familiar rule or pattern was preserved, and to incongruent (e.g. bapoba) sequences, in which the pattern was changed, in this case to XYX. These test sequences were composed of novel syllables. Using the headturn paradigm (Nelson et al., 1995), infants showed more interest in incongruent, i.e. novel, than congruent patterns, indicating that they had generalised the rule. Rule learning is thought to be a fundamental ability in the acquisition

of language and it has been studied mostly in that context Rabagliati, Ferguson, and Lew-Williams (see the meta-analysis in 2019, for an overview of language and non-language related rule learning studies).

The rule learning paradigm is simple and versatile enough to be used with different types of stimuli, different age groups, and even different species. As a result, it has been used in many follow-up studies to answer questions about the nature of rule learning. For example, it has been shown that rule learning is not domain specific, as infants can infer rules from e.g. visual stimuli (S. P. Johnson et al., 2009; Saffran, Pollak, Seibel, & Shkolnik, 2007) and chords(Dawson & Gerken, 2009) too. Rule learning is facilitated by stimuli that are meaningful, i.e. relevant to their everyday experience, like speech (Rabagliati et al., 2019), or multimodal stimuli (Frank, Slemmer, Marcus, & Johnson, 2009). Direct repetition also appears to be facilitating, as the XYY pattern is learned earlier than the XYX pattern, both with speech and visual stimuli (Gervain, Berent, & Werker, 2012; S. P. Johnson et al., 2009). Up until now some form of algebraic rule learning has been shown only in a single non-human animal species, the budgerigar (Spierings & Ten Cate, 2016). However, there are questions about the robustness of the effect. Rabagliati et al. (2019) in a meta-analysis of over 90 studies found an effect size (Hedges' g) of 0.25 for 7-month-olds overall, an enhanced effect when they were familiarized with meaningful stimuli, and a zero effect size when non-meaningful stimuli like tones or abstract visual objects were used. The average effect size was estimated as g = 0.25, 95% CI = [0.09, 0.40] (Rabagliati et al., 2019). A recent study in 4 Dutch babylabs found no evidence for rule learning in a close replication of the original paradigm (Geambașu et al., 2022).

5.3.2 Rule learning BST specifications

Research question and hypothesis

The main hypothesis in rule learning studies is that looking time differs for sequences with a novel versus a familiar structure–or sequences that preserve versus violate the learned rule. In particular, the expectation is that there are longer looking times to sequences with a novel structure than to sequences with familiar structure. The null hypothesis states that there is no difference in looking

time to novel and familiar trials. The alternative hypothesis states that the looking times to novel trials will be longer than looking times to familiar trials. With these specific hypotheses, we have all the elements needed to conduct a BFDA and BST.

Statistical models that represent the hypotheses

Once we determined the main hypothesis to test, we represent it in the form of statistical models that can be compared with a Bayes factor. The canonical test for the current hypothesis is the t-test. As in frequentist analysis, the data need to be prepared in the appropriate way by averaging the looking times for the novel trials and the familiar trials (but see van Doorn, Aust, Haaf, & Wagenmakers, 2021). The experimental effect is then computed as the difference between these two averages for each participant. The standardized effect size obtained in the sample is finally computed as the mean difference score divided by its standard deviation. This standardized effect size is the effect size of interest in the analysis that we present here and its population value is denoted as δ.

In Bayesian statistics the t-test is formulated as follows. We define the null hypothesis for the parameter of interest that states that the true effect size is exactly zero:

$$\mathcal{H}_0 : \delta = 0$$

Under the alternative hypothesis, the effect size is allowed to be different from zero. Similar to frequentist analysis, one needs to decide additionally whether the effect is expected to be positive, negative, or can go into either direction. This is done by assigning a prior distribution to the effect size.

There are many options on how to define the prior distribution. For situations where no previous data or knowledge is available about the size of the effect, so-called default prior distributions were derived (Rouder, Speckman, Sun, Morey, & Iverson, 2009) that meet general mathematical desiderata (Etz, 2018; Ly et al., 2020; Ly, Verhagen, & Wagenmakers, 2016b). These defaults are commonly used throughout the literature as well as implemented in popular statistical software (e.g., JASP Team, 2021; R. Morey & Rouder, 2018). For the t-test presented in this example, the popular default prior distribution for

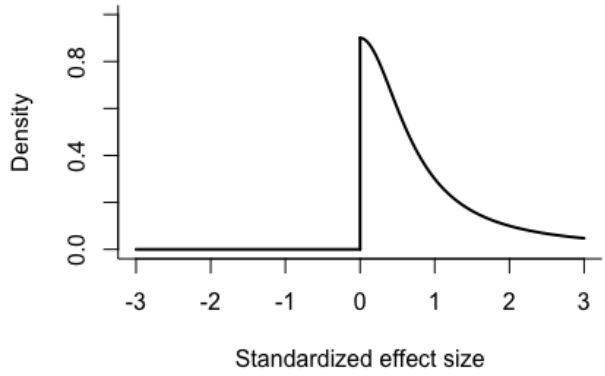


Figure 5.2: Cauchy prior under the alternative hypothesis that the effect size is positive.

the effect size is a Cauchy distribution (this distribution is equivalent to the t-distribution with one degree of freedom):

$$\mathcal{H}_1 : \delta \sim \text{Cauchy}(0, 1/\sqrt{2})$$

Since there is a clear expectation about the direction of the effect (i.e., novel sequences have *longer* looking times than familiar sequences) we will use a one-sided t-test. To achieve this, we truncate the distribution below zero, therefore assigning zero prior probability to effect sizes that are in the opposite direction than we expect. This truncation is the Bayesian equivalent of a one-sided test in the frequentist paradigm. Figure 5.2 shows the resulting prior distribution.

This default prior hence formulates the alternative hypothesis as being agnostic about the size of the effect and only stipulates that it is positive. Alternative specifications of the prior distribution are possible. Additional to the direction of the hypothesis, we could set a different scale of the Cauchy distribution. Typically, researchers would use other distributions to incorporate domain-specific knowledge about the size of the effect. Priors informed by domain-specific knowledge usually have the advantage that they make sharper predictions than the default priors, which can lead to stronger Bayes factors and thus

more efficient sequential testing designs. It is also advisable to provide a robustness analysis in the final report which shows how the conclusions would change if different prior distributions had been used (van Doorn, van den Bergh, Böhm, et al., 2021). However, for the purpose of BST, one needs to use a single prior specification to allow for unambiguous evaluation of the stopping criteria. Robustness analysis would be used only after the study has been completed.

In the case of rule learning, a meta-analysis of previous studies showed an average Hedges' g effect size of 0.32 for studies with "meaningful" stimuli (Rabagliati et al., 2019), so a new study in this area could assign a higher prior plausibility to values around 0.32 (see the Bayesian Updating box). However, Rabagliati et al. studied Hedges' g, which is an effect size used for between-subject comparisons. In the present example, the design is within-subjects, where the effect size is a standardized difference score. One should be careful about using those effect sizes interchangeably, because the two effect sizes correspond only when the correlation between the two paired measurements is exactly 0.5 (Dunlap, Cortina, Vaslow, & Burke, 1996; Morris & DeShon, 2002). We re-analyzed the meta-analysis data reported by Rabagliati et al. to find that the average correlation in previous studies is r=0.61, 95% CI = [0.47, 0.74]. This puts the average of the within-subject effect size to about d = 0.45, 95% CI = [0.01, 0.84] for studies in which meaningful stimuli were used. An informed prior distribution using these results is represented by a t-distribution with a degrees of freedom equal to 19.767, location equal to 0.443, and scale equal to 0.196[2]

For illustration, we conducted an additional BFDA analysis using this meta-analytic estimate on how one could use previous evidence for informing new studies. In the next section, we will be mainly focus on the BFDA results using the default Cauchy prior as the alternative hypothesis. The differences between the default and an informative prior is highlighted in Table 5.2.

Sequential testing design

After we established the main hypotheses that will be tested, the next step is to decide on the design of the study. That is, we need to specify how the data will

[2] For the sake of brevity, we do not include the reanalysis in the main text. Interested readers can refer to the Appendix A and our online materials at `osf.io/wak9e/`.

be collected, and what stopping criteria will determine the end of data collection.

The first decision that needs to be made is whether or not sequential testing will be done in the first place. Sequential testing brings some advantages compared to traditional study designs where data is collected until a pre-specified sample size has been reached. However, we will include traditional 'fixed n designs' in this example as well to demonstrate how the design works compared with the sequential testing designs.

The second decision is whether a maximum sample size should be specified. This allows the researcher to state in advance the sample size beyond which it is unfeasible or impractical to collect more data. In infant research, researchers are usually strictly limited with regard to what sample sizes they can realistically collect. Therefore, it is often reasonable to incorporate these resource constraints into the stopping rule of the sequential design. In this example, we will assume that the maximum sample size to be collected is 50 participants.

The third decision concerns specifying stopping criteria based on accumulated evidence for one or another hypothesis. In the current example, we will use the evidence thresholds $BF_{10} = 10$ and $BF_{10} = 1/10$. This means that under the sequential design, the data collection will be terminated if the data are 10 times more likely under one of the hypothesis compared to the other.

There may be additional aspects of the design to consider during planning of the study. As mentioned previously, it may not be always feasible to carry out the data analysis after collecting data from each individual participant. Instead, we may conduct the analysis in batches. In this manuscript, we will consider both an example where the Bayes factor is computed after each participant, and an example where the Bayes factor is computed after batches of five participants.

The criteria that we have described here result in three study designs that we will demonstrate in the next section. These three designs are summarised in Table 5.1. In the present article, we use the different designs to highlight the practical implications these design choices have on the data collection.

Design	Sequential	Max. sample size	BF thresholds	Step size
Fixed n	No	50	-	-
Sequential	Yes	50	$BF_{10} = 10$ or $1/10$	1
Sequential in batches	Yes	50	$BF_{10} = 10$ or $1/10$	5

Table 5.1: Summary of different study designs illustrated in this example.

5.3.3 Bayes Factor Design Analysis

Once we determined the study design, we conduct the BFDA which will give us information about the evidence we can expect under the selected design in case of the fixed sampling plan, as well as the sample sizes that we may expect to collect in case of the sequential sampling plan. This optional step helps us to assess whether the current study design is adequate to provide conclusive evidence to answer the research question, and whether the study is feasible to conduct.

BFDA is a simulation-based approach for design analysis that requires some knowledge in programming when using unconventional statistical models. Fortunately, many of the typical statistical tests used in Psychology, as well as variety of study designs, are implemented in the R-package `BFDA` (Schönbrodt & Stefan, 2018). This makes the analysis relatively straightforward, as the package does all necessary steps of the BFDA automatically. For example, the `BFDA.sim` function can be used to simulate the data, compute the corresponding Bayes factors, and decide whether stopping criteria have been met. The function repeats these steps many times so that we obtain the distribution of possible outcomes given the specified design. The simulations done for these typical designs are relatively simple and time efficient, and so conducting the analysis is a matter of only couple of minutes. However, we advise the reader to always run the simulation with only a few repetitions at the start to determine how long it would take to run the full analysis, and whether to run the full BFDA or simplify the approach.

Here is an example of using the `BFDA.sim` function from the `BFDA` package:

```
BFDA.sim(expected.ES = 0.45,
         type        = "t.paired",
         prior       = list(
                         "Cauchy",
                         list(
                           prior.location = 0,
                           prior.scale    = sqrt(2)/2)
                         ),
         alternative = "greater",
         design      = "sequential",
         boundary    = 10,
         n.min       = 5,
         n.max       = 50,
         stepsize    = 1,
         B           = 1000)
```

Here we see the following arguments:

`expected.ES` is the expected effect size which is used for generating the data. Under the null hypothesis, the effect size is 0. Here, the effect size under the alternative is set to the meta-analytic point estimate of the effect size under meaningful stimuli.

`type` specifies the type of test involved, in this case a paired t-test as we are comparing within participant averages over two types of trials — familiar and novel stimuli.

`prior` is the prior distribution on the effect size under the alternative hypothesis. In this example, it is the Cauchy$(0, \sqrt{2}/2)$ distribution defined earlier. Note that this is also the default prior that would be used in this case so the same analysis would if this argument was dropped.

`alternative` specifies whether the t-test is one-sided or not; in this case, we are expecting an effect larger than 0, Hence, the alternative is specified as `greater`. This results in the truncated prior distribution as shown in Figure 5.2.

`design` indicates whether sequential testing is used (`"sequential"`) or whether fixed n design is used (`"fixed.n"`).

`boundary` specifies the evidence threshold. When the obtained Bayes factor reaches this value, the data collection is stopped.

`n.min` specifies the minimum sample size.

`n.max` specifies the maximum sample size.

`stepsize` specifies how many participants are simulated between each computation of the Bayes factor. Here, it is set to 1, which means that the Bayes factor will be computed after every simulated participant. If batches of 5 participants are collected, the number needs to be increased to 5.

`B` specifies the number of repetitions of the simulation procedure.

Conceptually, this call of the `BFDA.sim` will repeat the following steps 1000 times:

1. Simulate data from one batch of participants (here, only one participant is in one batch), assuming an effect size $\delta = 0.45$ and add it to the data set.

2. Compute the Bayes factor comparing the null and alternative hypotheses on the current data set.

3. Determine whether stopping criteria have been met (i.e., is BF larger than 10 or smaller than $1/10$? Is the current sample size equal to 50?)

4. If criteria have not been met, repeat steps 1–3. If criteria have been met, record the obtained Bayes factor and the sample size, and run a new simulation.

Specifically, this code corresponds to the "sequential" design shown in Table 5.1. To get a complete picture of the design, we need to consider both hypotheses in the data simulation. Therefore, we run the same code twice with the only difference that the expected effect size argument `expected.ES` would

			Decision outcome			
Design	Simulation	Prior	Null	Undecided	Alternative	Mean N
Fixed n	Null	Default	25	74	1	50
		Informed	48	51	1	50
	Alternative	Default	≈ 0	33	67	50
		Informed	≈ 0	22	78	50
Sequential	Null	Default	38	58	4	42
		Informed	64	29	7	33
	Alternative	Default	≈ 0	21	79	30
		Informed	≈ 0	12	88	26
Sequential in batches	Null	Default	34	64	2	45
		Informed	60	35	5	36
	Alternative	Default	≈ 0	26	74	34
		Informed	≈ 0	14	86	29

Table 5.2: Overview of the BFDA results. The numbers represent the frequency (in percent) of adopting one of the three decisions (accepting null hypothesis, remaining undecided, accepting alternative hypothesis) under various scenarios. Design = design of the sequential testing (see Table 5.1). Simulation = whether the data were simulated under the assumption of the null hypothesis ($\delta = 0$) or an alternative hypothesis (instantiated as $\delta = 0.45$). Prior = Whether the default Cauchy prior was used, or the informative prior was used.

be specified as 0 in the second run of the function to generate data assuming that the null hypothesis is true.

Next, we present the results of the BFDA for the three alternative study designs presented in Table 5.1. The results for all designs are summarized in Table5.2 including the contrast between using informed versus defaults priors in each case. All code associated with the examples as well as the code for generating the figures can be accessed at `osf.io/wak9e/`.

Fixed n design

Let us first investigate the result of the BFDA in case one decided to use a "traditional" fixed sample size of fifty participants. Determining the distribution of evidence assuming a particular fixed design is akin to running a tradi-

tional sensitivity analysis, with the difference that the Bayesian hypothesis test can show evidence in favor of the alternative hypothesis, evidence in favor of the null, or inconclusive evidence. When using the default priors, if the data come from the simulation that assumes that the null hypothesis is true, about 25% of the simulated studies correctly show strong evidence for the null (i.e., $BF_{01} \leq 10$). Less than 1% of the studies incorrectly show evidence for the alternative ($BF_{10} \leq 10$). The rest of the studies (about 74%) remains inconclusive; see the first row in Table 5.2 for these numbers and the corresponding numbers when the informed is used. When using the default priors, if the data are generated under the assumption that the effect size is 0.45, about 67% of the studies correctly show support for the alternative hypothesis. About 33% of the studies remain inconclusive, and essentially zero studies incorrectly supports the null hypothesis (i.e., yield BF of 1/10 or less).

Figure 5.3 shows the distribution of the Bayes factors for a fixed sample size design of 50 participants. Clearly, in terms of error rates, or the number of simulations resulting in misleading conclusions, this design is effective at discriminating between the null and the alternative hypotheses. However, the plot reveals an issue with the design: Many studies would lead to extremely strong evidence (i.e., many Bayes factors would be larger than 30, with a large portion exceeding 100). Thus, although when designing this study it was determined that a Bayes factor of 10 would provide sufficient evidence, these simulations reveal that with 50 subjects, the evidence is often much stronger. Therefore, if one would be content with less conclusive results (e.g., $BF = 10$), the current design potentially wastes a lot of resources by collecting data from a larger sample than is necessary to reach the desired amount of evidence. As hinted previously, the solution to this inefficiency is sequential testing, which is discussed next.

Bayesian sample size specification

A BFDA for fixed-n designs can answer a similar question as a frequentist sensitivity analysis: What is the probability of finding positive evidence under the alternative hypothesis assuming a fixed number of participants? Hence, one may wonder whether there is a Bayesian coun-

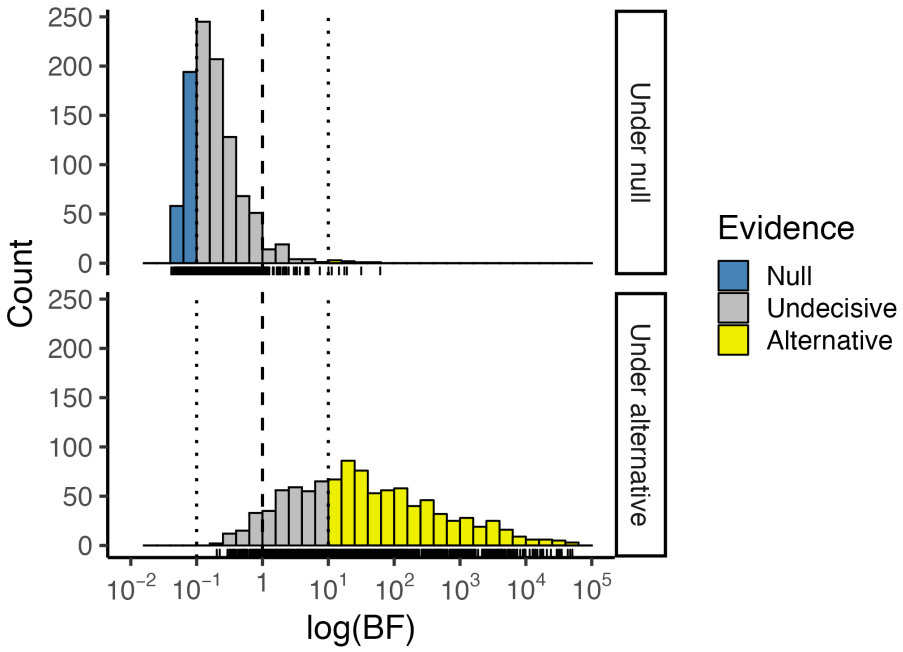


Figure 5.3: Distribution of Bayes factors under the fixed n design with 50 participants. Each data point represents the Bayes factor of one simulated study after collecting all 50 participants. The dotted lines show the Bayes factors of $1/10$ and 10, indicating "strong" evidence for the null and alternative hypothesis, respectively. The dashed line highlights a completely indifferent Bayes factor of 1. The bottom panel shows the distribution of Bayes factors assuming that the effect size is $\delta = 0.45$. The top panel shows the distribution of Bayes factors under the assumption that the null hypothesis is true (i.e., $\delta = 0$).

terpart to frequentist sample size specification based on power analyses: What is the required number of participants to have a pre-specified probability of finding positive evidence for an effect if the effect exists? BFDA also allows for this type of analysis.
Specifically, in a Bayesian setting, we may ask the question: What sample size is needed to get conclusive evidence for the alternative hypothesis (e.g., $BF_{10} > 10$) with a probability of at least 80%, if the true effect size is, say, 0.45? The answer cannot be obtained from a single fixed-n BFDA, but can be obtained by using a series of consecutive BFDAs where sample size is iteratively increased until the formulated requirement on design characteristics is met. In the BFDA R package (Schönbrodt & Stefan, 2018), this functionality can be found in the SSD function. Using this function with our rule learning example, we esablished that at least 63 participants would be needed to obtain a 'Bayesian power' of 80%.
One advantage of Bayesian statistics is that it allows researchers to quantify evidence in favor of the null hypothesis (Wagenmakers, Morey, & Lee, 2016). Thus we can plan not only for sufficient power to detect an existing effect, but also for a high probability of obtaining strong evidence for the null hypothesis, if the true effect is zero. For example, a researcher may aim to obtain strong evidence for the null hypothesis (i.e., $BF_{01} < 1/10$) with a probability of at least 60% if the rule learning effect is in fact zero. Using the SSD function reveals that the required number of participants in this scenario is 198 participants. It is typical that this number is much higher than for 'Bayesian power' under the alternative hypothesis because evidence for the alternative hypothesis generally accumulates faster than for the null hypothesis (V. E. Johnson & Rossell, 2010).

Sequential design

Instead of the fixed sample size design, researchers may be interested in performing a sequential design, and plan it using a BFDA for sequential designs. In our first sequential testing example, we consider a setup where a researcher performs a Bayesian hypothesis test after each participant (starting from a min-

imum of five participants), until a Bayes factor greater than 10 or smaller than 1/10 is encountered, or until the maximum number of participants ($n_{\text{max}} = 50$) is reached. Again, we run this simulation under the assumption that the null hypothesis is true ($\delta = 0$) and under the assumption that the effect size is $\delta = 0.45$. Figure 5.4 shows the results of the BFDA when the default priors were used. Specifically, each line represents a separate simulated study, and shows the accumulation of evidence (y-axis) as the data increases in size (x-axis). Straight lines drawn at $\text{BF}_{10} = 10$ and $\text{BF}_{10} = 1/10$ represent the evidence boundaries for the alternative and null hypothesis, respectively. The dots on these lines represent a study that terminated due to reaching the boundary, and is drawn at the sample size that was needed to obtain the required evidence. The distribution on the right side of each plot show how many simulated studies ended because the maximum sample size was reached, rather than one of the evidence thresholds. The distribution indicates the strength of evidence for these studies, the colors indicate the direction of the Bayes factor at $n_{\text{max}} = 50$.

Similar to the fixed design analysis discussed earlier, we can use the table 5.2 to determine the percentage of studies that lead to correct or incorrect conclusions, and the percentage of studies that remain undecided. Under the null hypothesis, about 58% of the simulated studies remain inconclusive even after collecting the maximum number of 50 participants. Although this number seems large, it is actually an improvement to the fixed designs, under which 74% of studies remained inconclusive. The reason for this phenomenon is that as we accrue more data, Bayes factors (on average) drift towards the more accurate hypothesis – but there is some variation due to sampling noise. This means that some studies that would not have reached the threshold at the maximum sample size may cross the threshold at an earlier point in time, which gives a slight advantage to the sequential design compared to the fixed design. A noticeable improvement also comes for the percentage of studies that correctly support the null hypothesis, which increased from 25% under the fixed design to about 38%. Under the assumption that the effect size is $\delta = 0.45$, the sequential design also performs better than the fixed n design in terms of the error rates. About 79% of the studies reached a Bayes factor of 10 in favour of the alternative hypothesis before the sample size of 50 was reached. About 21% of the studies terminated at the maximum sample size without reaching

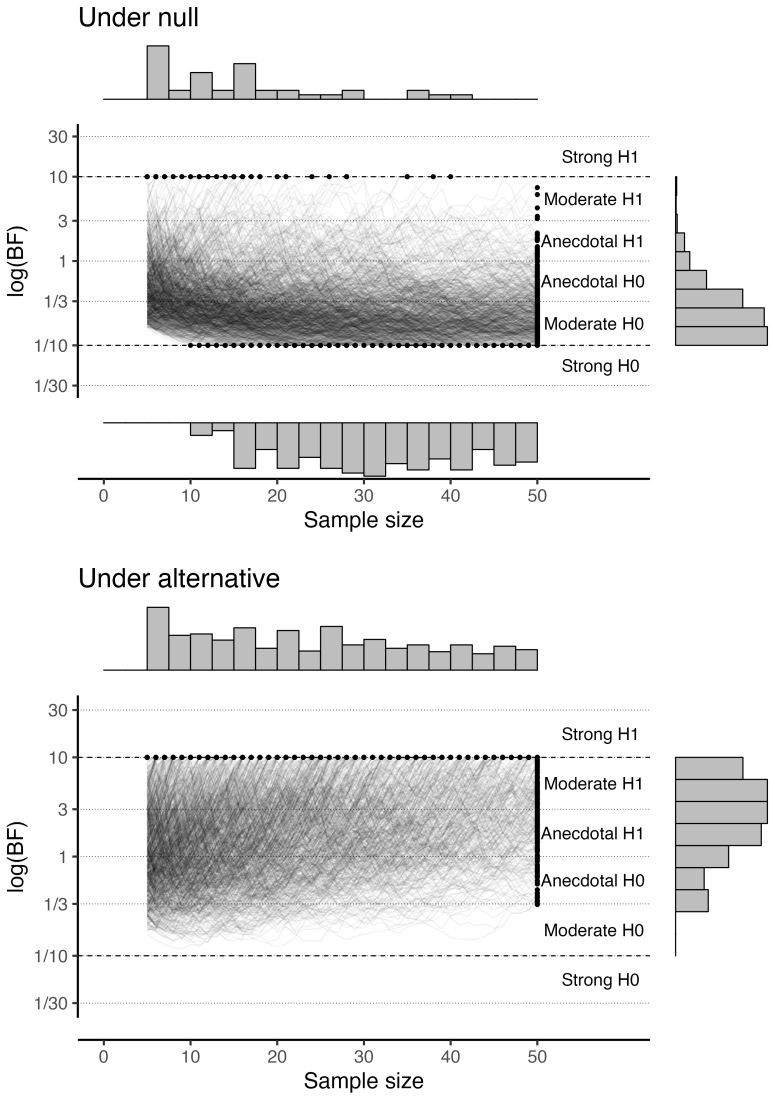


Figure 5.4: Individual simulated studies and the evidence obtained after each participant. Each line represents a single study and the Bayes factor is plotted on the y-axis after each added participant (on the x-axis). Histograms on the bottom and top margins display the distribution of sample sizes needed to reach evidence in favor of the null and alternative hypothesis, respectively. The histogram on the right side shows the distribution of final Bayes factors of studies that did not reach conclusive evidence for either hypothesis by the time the maximum sample size was reached. The top panel shows simulations under the assumption that the null hypothesis is true, and the bottom panel shows simulations assuming that the effect size is $\delta = 0.45$.

the evidence boundary.

Most importantly, we can also extract information about expected sample sizes from the BFA for sequential designs. Remember that in sequential testing, one has the opportunity to conclude the data collection early if enough evidence is accumulated, instead of waiting to collect the maximum sample size. From Figure 5.4, we can see that under the assumption that the true effect size is $\delta = 0.45$, the study can often be concluded much earlier than at the maximum sample size of 50 participants, which shows that one can improve the efficiency of the experiment by employing the sequential testing design. However, if the data come from the null hypothesis, the benefits of the sequential design in the current scenario are modest; the expected sample size to be collected is about 42, suggesting that under the null hypothesis one saves resources equivalent to only 8 participants on average. On the other hand, if the true effect size is 0.45, the average sample size at stopping point is about 30, meaning that the expectation is to save resources of about 20 participants compared to the fixed sample size design.

Sequential design in batches

The previous section showed that by allowing the experiment to end whenever a predefined amount of desired evidence is reached, the design becomes more efficient. Often, data collection can be terminated much earlier than the maximum sample size of fifty participants.

However, the previous analysis assumed that the Bayes factor will be computed after every single participant. In realistic scenarios, this is often not feasible; for example, participants are usually scheduled ahead of time, or are tested in groups in a larger laboratory so that it becomes impractical to calculate the Bayes factor sequentially after each participant. Additionally, the raw data may need to be processed before entering it into the main analysis, which can make it infeasible to run the analysis after every single experimental session.

However, these practical restrictions do not preclude the researcher from conducting a sequential design. For example, the researchers may decide to calculate the results after every week of collecting the data, and they estimate that every week, on average, they collect about 5 participants. The sequential analysis can then be used assuming that the Bayes factor will be not computed

after every participant, but once every week, which would mean computing the results approximately after every fifth participant.

In general, the efficiency of sequential designs with larger step sizes lies somewhere between the fixed designs and a sequential designs with a step size of one. The larger the step size, the smaller the benefits of the sequential analysis, as there are less and less opportunities to terminate the experiment before the maximum sample size is reached. However, even relatively large step sizes can bring some benefits to the researcher, especially in situations where the data is more diagnostic than expected.

Suppose that a researcher estimates that they will be able to run the analysis after about every fifth participant. Besides that, the sequential design is run again with a minimum sample size of five and a maximum of fifty participants.

As can be read from Table 5.2, the sequential design's performance lies somewhere in between the fixed n design and the sequential design presented previously. If the true effect size is zero, when using the deafult priors, the BFDA simulation results show that about 34% of studies correctly support the null hypothesis, about 2% support the alternative hypothesis, and about 64% run until the maximum sample size is reached. The average sample size at the stopping point is 45, suggesting that if we are lucky, we may be able to conclude the experiment earlier than depleting the whole pool of 50 participants.

Under the assumption that the effect size is 0.45, when using the default priors about 74% of the studies would correctly support the alternative hypothesis, virtually no studies would support the null hypothesis, and about 26% of the studies would run until the maximum sample size is reached. The average sample size at the stopping point is 34, suggesting that even without the possibility to test the effect after every subject, it is possible to retain some advantages of the sequential design over the fixed sample size.

Remarks on the BFDA

In this example, we conducted BFDA with three different study designs. This demonstration shows how BFDA can be used to gain insight into a planned study. One can run BFDA for various alternative designs to establish which design to choose for actual data collection. For example, the present examples demonstrated that one can increase the efficiency of the experiment by using

sequential testing, even if computing the Bayes factor after every single participant is unfeasible.

Importantly, we were able to run the variations of BFDA because the current example is relatively simple and the analysis can be comfortably run using the `BFDA` package in a matter of a few minutes. This is not always the case. However, it is important to keep in mind that running BFDA is an optional step in a sequential testing study. It can provide insights when planning the study and may help researchers to come up with better alternative designs, but it is not necessary for interpreting a sequential testing experiment in any way.

The efficiency of the design can be improved by incorporating knowledge about the studied phenomenon into the experiment by means of using informative priors. The main text focused on presenting the results using uninformative (default) prior to demonstrate how to use BFDA when no prior information is available. However, in the context of the present example, it would be possible to incorporate previous knowledge by using the meta-analytic results from Rabagliati et al. (2019). Table 5.2 highlights the differences between BFDA using default and informed priors. Informed priors make sharper predictions than the default priors, which leads to more decisive Bayes factors. This leads to decreased probability that the experiment ends up being inconclusive, and increases the chances of reaching correct conclusion earlier than when using default priors. The exception to this pattern is when the results from the current experiment for some reason deviate substantially from the past estimates. Informed priors guarantee sharper predictions, but if those predictions do not agree with current data, it could become harder to provide strong evidence for the correct hypothesis as the prior that makes confident but inaccurate predictions will suffer a penalty. In these cases, default priors may actually outperform informed priors.

5.3.4 Run the study and analyze the data

Once we planned the study, we may proceed with the data collection and data analysis. Here, we demonstrate an outcome of a sequential analysis of the data from Geambasu, van Renswoude, Visser, Raijmakers, and Levelt (2021), Study 2, a replication of a study reported by Marcus et al. (1999). Although this study was conducted with a fixed n design with 40 participants, we can analyze the

ppt	consistent	inconsistent
1	10.28616667	8.68
2	10.867	10.3805
3	13.22766667	13.35333333
4	10.86966667	10.44666667
5	7.096833333	8.205666667

Figure 5.5: Screenshot of the data from (Geambasu et al., 2021)

data sequentially to show how the evidence accrued over time and illustrate how the study might have unfolded if a sequential analysis had been adopted.

To provide a sense for the present illustration, Figure 5.5 shows the first five rows of the data set to be analysed here, if we open it in JASP (JASP Team, 2021). Each participant ("ppt") occupies a single row and the average looking times for consistent and inconsistent stimuli are recorded in separate columns.

Analyzing the entire data set of 40 participants yields a Bayes factor in favor of the alternative hypothesis of about 0.123, or equivalently, the data from all participants are about 8.15 times more likely under the null hypothesis compared to the alternative hypothesis.

To conduct the sequential analysis, the data from each participant are added in the order they were originally tested. Figure 5.6 shows the result of the sequential analysis. After each participant, the Bayes factor is computed, as specified earlier, for a one-sided test with a default Cauchy prior on effect size under the alternative hypothesis. Each point shows the Bayes factor obtained after each participant. In the figure, Bayes factors above 1 show evidence for the alternative hypothesis, and Bayes factors below 1 are evidence for the null hypothesis. It can be seen that the Bayes factor rapidly decreases, converging towards evidence for the null hypothesis. Although the evidence for the null is mostly moderate, there are two points (highlighted in red) where the Bayes factor crosses the bound of $1/10$ - first after including participant nr. 28, and again after including participant nr. 32. If the data were collected using sequential design with the same criteria as outlined in this previous section, the experiment would have been concluded earlier, thus saving resources of 12 participants that were tested after desired evidence threshold has been met. Note that although the Bayes factor is above the bound after participants number 28 and

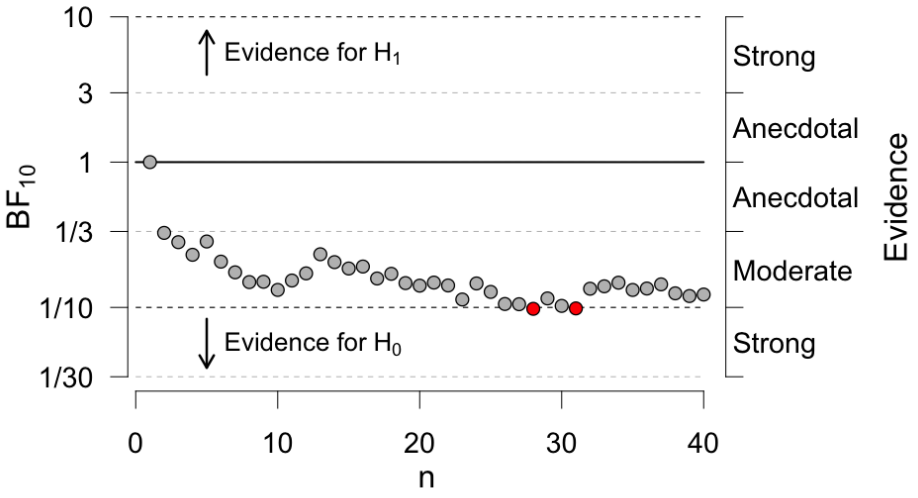


Figure 5.6: Output of a sequential analysis.

32, it moves slightly to the other side of that bound after adding participants 29 and 33. However, it is important remember that the Bayes factor summarizes the evidence contained in the data up to that point, and that at N=28, there is legitimate strong evidence for Ho, which is the same conclusion that we would make (albeit with less confidence) after sampling 40 participants. By using sequential sampling, we arrived at the same conclusion with fewer subjects. So, even if (due to randomness of the samples drawn), the Bayes factor slightly fluctuates around the stopping threshold in a certain region of the trajectory, the whole trajectory will tend towards the Ho threshold as the sample size grows. Therefore, by using sequential testing we don't need to wait until the "final" crossing of the threshold, and can simply make our decision after the first time it crosses.

If we analyze the data sequentially but in batches of five participants, the obtained Bayes factors are depicted in Figure 5.7. This results in the same pattern as in the previous sequential analysis, with the difference that some of the Bayes factors are not computed. Here, it would just so happen that the Bayes factor would not cross the evidence bounds at any point, which would mean that whereas the sequential testing design would be able to terminate the ex-

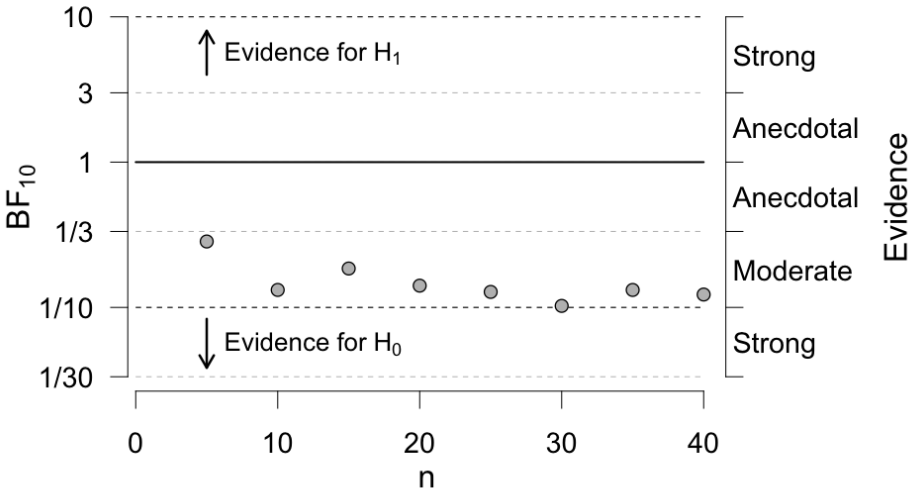


Figure 5.7: Output of a sequential analysis in batches of five participants.

periment earlier, the batched sequential analysis would not — eventually lead to the same result as the fixed n design.

Researchers may wonder what steps they should take if the Bayesian hypothesis test results in inconclusive evidence. From a statistical viewpoint, adding additional observations in a sequential manner is always possible, but resource limits might prohibit further data collection (Schönbrodt et al., 2017). In this case, we recommend that researchers acknowledge the uncertainty with respect to the hypotheses under test. This can be achieved by reporting the final Bayes factor together with a cautionary note that the amount of evidence obtained does not allow strong conclusions in favor of either hypothesis. Moreover, it is advisable to report an effect size estimate together with a credible interval so as to inform the planning of future studies. Importantly, similar to nonsignificant findings in frequentist testing, inconclusive Bayes factors should not be hidden in a file drawer, as this withholds information and introduces bias into the literature (Rothstein, Suton, & Borenstein, 2006).

5.4 Conclusion & Discussion

In this tutorial we have outlined the methods of BST and BFDA with the aim of providing developmental researchers with the appropriate tools to start experimenting with these methods themselves. The illustrations have shown that the use of sequential testing results in important gains in efficiency. In the presented scenario, researchers can expect an efficiency gain of more than 40% for a medium-sized effect size under the alternative hypothesis. Even if the effect turns out to be a null effect, there is still a gain in efficiency of 16%. In addition, making use of the Bayesian testing framework provides additional information because it allows researchers to distinguish between (strong) evidence in favor of the null hypothesis and inconclusive evidence (Altman & Bland, 1995). Mani et al. (2021) more elaborately discuss the advantages and disadvantages of using BST in developmental research. Here, in addition, we focused on planning study designs prior to data collection using BFDA.

In many typical developmental studies, common statistical tests are used such as t-tests, correlations, and regressions. For these cases, the BFDA R package can be used without further ado. Running a BFDA in those cases hence provides researchers with information about the reasonableness of their choices (e.g., the maximum sample size) and their chances of finding conclusive and strong evidence. This is similarly the case for running a sequential analysis. In many common cases, JASP suffices to run the sequential analysis repeatedly while data collection is ongoing. We have also shown an example of running a sequential analysis on a real data set, also highlighting the possibilities for additional robustness checks.

When researchers are planning for more complicated analyses and designs, e.g., multi-level models, sequential analysis during data collection is still possible and advisable. As stressed throughout, the Bayes factor can be validly interpreted as the strength of evidence at any point during the study and does not rely on any prior sampling plan. This is a major difference to frequentist null hypothesis significance testing. Running a full BFDA for these more complex sequential testing designs can be computationally infeasible. Our article discusses several potential solutions.[3]

[3]Note that for very complicated models it may be hard or impossible to compute consistent

The BFDA that we ran in the illustration shows that although we can likely save on running participants, there is also still some possibility that evidence will be inconclusive, even after running 50 participants. Assuming an effect size of 0.45, about 13% of studies would result in inconclusive evidence when the maximum sample size is 50. This number will naturally be higher when the effect size under study is smaller – as will often be the case in developmental studies. One conclusion one may draw from this is that researchers need to plan for the possibility of larger sample sizes if the desire is to arrive at conclusive results, and for the good of science this would certainly be desirable. It will decrease the rate of papers with inconclusive evidence and will likely contribute to larger generalizability (see Yarkoni, 2019, for discussion of generalizability).

Planning for the possibility of larger sample sizes may not be easy. One way to go is to aim for the type of large scale collaborative efforts that are being launched in many areas of psychology and other areas of science (Frank et al., 2017; Moshontz et al., 2018; Primates et al., 2019). However, that may not be practical or necessary for every study. It is important to note that the necessity for larger sample sizes for some studies is counterbalanced by the fact that on average smaller sample sizes will be required when using sequential testing. Even so, seeking (smaller scale) collaborations with other labs to plan for the eventuality of larger required sample size can pay off in a number of ways. Not only will such collaboration indeed provide the opportunity to run larger sample sizes. Such collaboration also forces the researchers involved to be very explicit about their testing procedures and lab practices. This in turn may have the beneficial side-effect that heterogeneity between studies is reduced and/or that it may reveal which procedural differences turn out to be more relevant than initially thought thereby gaining knowledge that can help increase reliability of procedures and measures (Byers-Heinlein, Bergmann, & Savalei, 2021, see discussion in).

A division of labor between labs and researchers can also be applied in cases where experimental researchers do not feel confident in running complicated analyses. Seeking collaborations in such cases can also benefit science at large by optimally using the expertise and experience from each of the partners that

Bayes Factors such that for example the BF is bounded; running a BFDA can help gaining insights into such cases if no other information about those designs is available.

are involved. Similarly, collaborations with between experimental scientists and computational modellers may be beneficial to both parties.

It may be thought that optional stopping rules such as applied in sequential testing as discussed here may lead to biased estimates of effect sizes. This concern is not without its grounds. Indeed, when stopping early on in a sequential testing design, the effect size may be slightly overestimated relative to the population value. Finding out under which conditions optional stopping may lead to problematic inferences and how severe those problems are, is an active area of research, studying the effect of different priors, model specification and research goals (de Heide & Grünwald, 2021; Hendriksen, de Heide, & Grünwald, 2021) and strategies to prevent potential negative side-effects of optional stopping (Sanborn & Hills, 2014). Regardless of this, across multiple sequential studies the overall evidence converges towards the population value. The alternative practice of fixed sample sizes, and frequentist null-hypothesis testing is also not without its problems. Sampling variation in that case can also lead to an overestimation of the effect, certainly when such findings are off-set against studies that end up in the file-drawer. The problem of biased findings is not unique to Bayesian practice, and rather calls for more widespread replications more generally. See (Yu, Sprenger, Thomas, & Dougherty, 2014) for a much broader discussion of the use and effects of decision heuristics in science independent of the statistical framework one is using.

The area of application for Bayesian sequential testing and BFDA is that of explicit hypotheses. These methods can then take care of efficiently gaining evidence for each of the hypotheses under study. Oftentimes, the hypotheses under study will represent a null hypothesis and an alternative hypothesis. Psychology in general, and developmental psychology also in particular, could benefit from having more informative hypotheses. Meaning that hypotheses under study are more tightly coupled to theoretical considerations, a call for stronger theories (Yarkoni, 2019) that improve generalizability. There has been much recent discussion about the 'theory crisis', the need for stronger theories, and how to develop them (Borsboom, van der Maas, Dalege, Kievit, & Haig, 2021). Stronger theories however can only be built on a stronger foundation of robust empirical findings (Eronen & Bringmann, 2021). The use of methods proposed in this tutorial can help build such strong foundations and thereby

build a stronger developmental psychology.

Open Practices Statement

All code associated with this article is available at osf.io/wak9e/

Appendix

5.A Determining informed priors

In the article, main focus was given to the BFDA using the default priors. However, informed priors were used as well, highlighting that using prior information can further improve characteristics of the design. This appendix provides details about how was the informative prior determined. All code associated with the analyses are available at the project's online repository.

Effect size of interest

Setting priors for Bayesian analyses requires an understanding of the underlying statistical model so that the probability distributions that represent different hypotheses are appropriate to the meaning of the parameters of interest.

The example discussed in this article focused on a within-subject design. In this design, each participant receives a score in two conditions. Given that the two measurements can be correlated, the data are assumed to be generated from the following distribution:

$$(X, Y) \sim \mathcal{N}_2\big((\mu_x, \mu_y), \Sigma\big), \tag{5.3}$$

where (X, Y) are the two measurements per participant, $\mathcal{N}_2$ is a bivariate normal distribution with means (μ_x, μ_y) and a variance covariance matrix Σ:

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}, \tag{5.4}$$

where σ is the standard deviation of the measures, and ρ the correlation

between the two measures.

To test whether the two measures have different means, the common approach it to use a paired-samples t-test. This test can be thought of as taking the difference score between the two measurements: $W = X - Y$, and conduct a one-sample t-test to test the mean of the difference score against zero.

Under the current assumptions, the distribution of the difference score is:

$$X - Y = W \sim \mathcal{N}(\mu_w, \sigma_w), \tag{5.5}$$

where the mean μ_w and the standard deviation σ_w of the difference score can be expressed in terms of the parameters of the distribution of X and Y:

$$\begin{aligned} \mu_w &= \mu_x - \mu_y, \\ \sigma_w &= \sigma\sqrt{2(1-\rho)}. \end{aligned} \tag{5.6}$$

The *within-subjects effect size* in the context of a paired-samples t-test is the *standardized mean difference score*:

$$\delta_w = \frac{\mu_w}{\sigma_w} = \frac{\mu_x - \mu_y}{\sigma\sqrt{2(1-\rho)}}. \tag{5.7}$$

An alternative to the within-subjects effect size is the *between-subjects effect size* (commonly empirically estimated using a Cohen's d or its alternatives like Hedges' g), which is applicable to both within-subjects and between-subjects designs. The population value of this effect size is defined as:

$$\delta_b = \frac{\mu_x - \mu_y}{\sigma}. \tag{5.8}$$

As can be seen by comparing Equations 5.7 and 5.8, the difference between the within-subjects (δ_w) and between-subjects (δ_b) effect sizes is in the denominator: Whereas the within-subjects effect size takes into account the correlation between the two measurements, the between-subjects effect size does not. Formally, the two effect sizes are related by the following equation:

$$\delta_w = \frac{\delta_b}{\sqrt{2(1-\rho)}}. \tag{5.9}$$

When the correlation is larger than 0.5, the within subject effect size is larger

than the between subject effect size.

In the Bayesian paired-samples t-test used in this article, the prior representing the alternative hypothesis is placed on the within-subjects effect size. We wish to determine a prior distribution that is informed by previous empirical evidence. Rabagliati et al. (2019) provided a meta-analysis of studies related to the current application. However, the effect size of interest in the meta-analysis was a between-subjects effect size, not a within subjects effect size. Therefore, we cannot use the estimate from the original meta-analysis directly because it is an estimate of a different effect size measure. However, all of the studies included in the meta-analyses were actually within-subjects. Thus, it is possible to reanalyze the data associated with the original meta-analysis to estimate the within-subjects effect size in question.

Estimating the within-subject effect size

Data provided by Rabagliati et al. (2019) were used to estimate the within-subjects effect size. No crucial modifications of the data were needed, as the within-subjects effect sizes and their standard errors are easily obtainable from test statistics (Borenstein, Hedges, Higgins, & Rothstein, 2021). The individual estimates are clustered within articles (some articles report multiple effect sizes from multiple studies) and labs (some labs published multiple articles). To account for this, we conduct a random effects meta-analysis.

Let d_{las} and se_{las} be the sample within-subjects effect size and the standard error of the estimate, respectively, for lab l, reported in article a, study s. The random effects meta-analysis of the correlation coefficient can be written down as follows:

$$\begin{aligned} d_{las} &\sim \mathcal{N}(\mu_{las}, se_{las}) \\ \mu_{las} &= \mu + \beta_l + \beta_a + \beta_s \\ \beta_l &\sim \mathcal{N}(0, \sigma_l) \\ \beta_a &\sim \mathcal{N}(0, \sigma_a) \\ \beta_s &\sim \mathcal{N}(0, \sigma_s) \end{aligned} \tag{5.10}$$

The average within-subject effect size is μ.

Only subset of the studies that were labelled as presenting "meaningful"

stimuli (Rabagliati et al., 2019), resulting in a set of 59 individual effect size estimates clustered under 10 labs. A frequentist analysis using the package `metafor` (Viechtbauer, 2010) yielded an estimate $\mu = 0.46$, 95% CI = [0.10,0.82]. We also ran a Bayesian meta-analysis using the JAGS software (Plummer, 2003) while using weakly informative priors for the parameters, and arrived at an estimate of $\mu = 0.45$, 95% CI = [0.01,0.84].

Determining the prior distribution using the empirical estimate

To conduct a BFDA analysis using an informed prior, one needs to represent the empirical estimate in terms of some probability distribution. The easiest way to do so would be using a Normal distribution with the mean set to the empirical point estimate, and a variance such that 95% of its mass lies within the confidence (or credible) intervals. In this article, we used a shifted and scaled t-distribution instead, as the t-distribution allows better flexibility at the tails of the distribution given the degrees of freedom parameter. The t-distribution was fitted using maximum likelihood approach to the posterior samples from the Bayesian analysis. The best fitting parameters were df $= 19.767$, location $= 0.443$, scale $= 0.196$. The resulting distribution is shown in Figure 5.8.

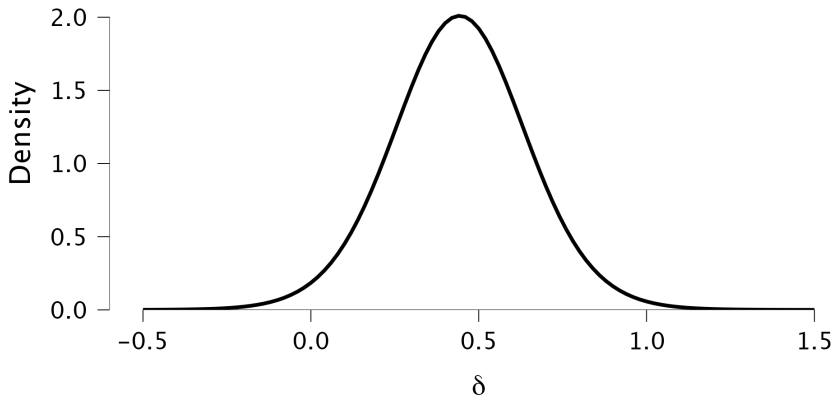


Figure 5.8: Distribution representing the estimate of the average within-subject effect size obtained by the reanalysis of data from Rabagliati et al. (2019).

The determined distribution was used in the article to contrast the results versus the default prior. The only difference was that as in the case of the default prior, we truncated the distribution below zero to represent a one-sided test.

One accusation you can't throw at me is that I've always done my best.

–Alan Shearer

Chapter 6

Habituation, Part I. Design Choices in the Infant Habituation Paradigm: A Pre-registered Crowd-Sourced Systematic Review and Meta-Analysis

This chapter is a Stage I. Registered Report accepted at *Infant & Child Development* and preprinted as Zaharieva, M., Kucharský, Š., Colonnesi, C., Gu, T., Jo, S., Luttenbacher, I., … Visser, I. (2022). Habituation, Part I. Design choices in the infant habituation paradigm: A pre-registered crowd-sourced systematic review and meta-analysis. *PsyArXiv*. doi: 10.31234/osf.io/bdtx9

Abstract

Methodological variations and inconsistency in reporting practices pose considerable challenges to the interpretation and generalizability of outcomes derived from the habituation paradigm - one of the most prominent methods for studying infant cognition. In a systematic review, we map out experimental design choices in habituation study samples aged 0-18 months using looking time measures. 2,853 records published in peer-reviewed journals between 2000- 2019 were extracted from PsycInfo and Web of Science. 781 (27.4%) papers were deemed eligible after screening (Fleiss kappa = .60, 95%, CI[.40 - .80], 6 blind raters). We adopt a collaborative, multi-lab approach for crowd-sourced data collection involving raters from the developmental research community. In a meta-analysis, we assess the impact of habituation detection criteria on the novelty effect size, moderated by age. Our results will inform a detailed evaluation of experimental designs and a set of recommendations to improve research and reporting practices in infant habituation research.

6.1 Introduction

The habituation paradigm is among the most prominent methods for studying infant cognition (Colombo & Mitchell, 2009; Kellman & Arterberry, 2000; Oakes, 2010) Experimental designs and protocols, as well as reporting practices, vary greatly between studies. These methodological variations and lack of consistency in reporting practices pose considerable challenges to the interpretation and generalizability of the outcomes derived from habituation studies. Precisely characterizing the habituation process is further hindered by procedural variations with unknown effects on the outcomes (Kucharský et al., 2022). As a result, some of the basic characteristics of the paradigm remain elusive, most notably concerning the factors that determine the novelty effect (Hunter & Ames, 1988), as well as the characteristics of the habituation process itself that affect optimal experimental design choices (Colombo & Mitchell, 2009). In this registered report, we perform a systematic review and meta-analysis of experimental design and reporting practices over the past 20 years of the infant habituation paradigm and weigh the impact of - potentially arbitrary - methodological design choices on the presence and strength of the novelty effect. Before discussing the current study in more detail, we briefly sketch out the workings of the habituation paradigm

and its relevance to the field of infant research.

6.1.1 The Habituation Paradigm

Investigating mental processes in infancy has advanced substantially as a consequence of the development of methods that indirectly quantify infants' interest in a stimulus, one of the most widely used being the habituation paradigm (Bornstein, 1985; Fantz, 1964; Saayman, Ames, & Moffett, 1964). This method allows researchers to infer whether infants can discriminate between two (classes of) stimuli, by capitalizing on the idea that infants are more readily interested and responsive towards stimuli that are novel to them, and less so towards stimuli that they have encountered repeatedly.

During a typical habituation procedure, infants' behavioral or neurophysiological responses towards a certain stimulus are monitored. Commonly measured responses are, for instance, sucking, looking time, head orientation, and neurophysiology (Fennell & Werker, 2003). In this paper, we focus on looking behavior because it is a widely measured response across infancy (Clohessy, Posner, & Rothbart, 2001; M. Johnson & Tucker, 1996; Plude, Enns, & Brodeur, 1994) that is specifically directed towards or away from a particular stimulus in the visual environment rather than representing a general arousal state. Our findings are nevertheless likely to generalize to other response modalities because the basic workings of the underlying habituation mechanism are supposedly the same (Rankin et al., 2009).

The fundamental assumption of the habituation method is that over repeated presentations, an infant gradually loses interest in what is being displayed, which is in turn reflected in the decreasing rate, duration, and intensity of the visual attention response (Aslin, 2007). Once response attenuation has occurred, a recovery of attention should be observed towards any new, unfamiliar stimulus - that is, if the infant perceived it as such. This recovery of interest is taken as evidence that the infant can discriminate between the novel and the familiar stimulus (Sokolov, 1963). In this study, we define habituation as the decreasing response to a stimulus (or a stimulus category) over repeated exposure. A dishabituation response (also known as a novelty preference), we define as the increasing response to an unfamiliar stimulus (or a stimulus category) following successful habituation. These definitions encompass studies that use

the habituation and dishabituation responses as a means to investigate other cognitive phenomena such as discrimination (Streri & Pêcheux, 1986) and category learning (Gureckis & Love, 2004; Mareschal, French, & Quinn, 2000), among others. Procedures that yield measurable (dis-)habituation processes thus include protocols that allow to quantify as well as directly compare habituation and dishabituation from looking responses to familiar versus novel stimuli (e.g., familiarization procedures).

The habituation procedure is infant-friendly, versatile and relatively easy to implement because it takes advantage of the natural tendency of infants - and to a great extent among other species - to direct attention towards meaningful stimuli in the environment (Fantz, 1957, 1958, 1964; Rankin et al., 2009; Saayman et al., 1964). In cognitive models, attention is thought to reveal steps in information processing such as selection among competing stimuli, stimulus representation, or comparison between the environment and contents of memory (Bornstein, Colombo, & Pauen, 2012). To this day, hundreds of habituation studies have been conducted to investigate the mental processes of infants. Streams of research such as language learning (e.g., Bijeljac-Babic, Höhle, & Nazzi, 2016; Byers-Heinlein et al., 2021; Fennell & Werker, 2003; Kajikawa, Fais, Mugitani, Werker, & Amano, 2006), face perception (e.g., Anzures, Quinn, Pascalis, Slater, & Lee, 2009; Damon, Quinn, Heron-Delaney, Lee, & Pascalis, 2016; Sakuta, Sato, Kanazawa, & Yamaguchi, 2014; Xiao et al., 2015), numerical cognition (e.g., Brannon, 2002; Lipton & Spelke, 2003, 2004; F. Xu, Spelke, & Goddard, 2005), rule learning (e.g., Bulf, Brenna, Valenza, Johnson, & Turati, 2015; Bulf, Johnson, & Valenza, 2011; Frank et al., 2009; Kirkham, Slemmer, Richardson, & Johnson, 2007), emotion processing (e.g., Addabbo, Longhi, Marchis, Tagliabue, & Turati, 2018; Brenna, Proietti, Montirosso, & Turati, 2013; Hock et al., 2017; Ichikawa, Kanazawa, & Yamaguchi, 2014), social development (e.g. A. Henderson, Wang, Matz, & Woodward, 2013; A. Henderson & Woodward, 2011; Kelly et al., 2007), intelligence (e.g., Kavšek, 2004; McCall & Carriger, 1993; Slater, 1997), memory (e.g., Dupierrix, Hillairet de Boisferon, Barbeau, & Pascalis, 2015; Jones, Pascalis, Eacott, & Herbert, 2011; Oakes & Kovack-Lesh, 2013; Zosh, Halberda, & Feigenson, 2011), among others, were able to make considerable progress in understanding sociocognitive development using the habituation paradigm. Yet, best practices in designing habitua-

tion studies are not firmly established and in concordance with methodological and technological developments in infant research, nor is our understanding of habituation as a cognitive process in infancy (Colombo & Mitchell, 2009; Kucharský et al., 2022; Sirois & Mareschal, 2002).

6.1.2 Understanding Experimental Design Factors in Relation to the Habituation Process

Despite a wealth of research into habituation, which factors contribute to producing either novelty or familiarity effects is still debated (e.g., Bergmann, Rabagliati, & Tsuji, 2019). Hunter and Ames (1988) offered a number of proposals about putative factors that could influence whether a familiarity or a novelty preference can be expected (these will be empirically assessed in the multi-lab project Many Babies 5: `manybabies.github.io/MB5`). Here we distinguish between three types of factors: 1) the design factors of the habituation phase, 2) the level of habituation, determined by the habituation detection criterion, and 3) age, which has shown to interact with the course and rate of the habituation process (Colombo, 2002; Colombo, Mitchell, O'Brien, & Horowitz, 1987; Colombo & Mitchell, 2009; Colombo, Shaddy, Richman, Maikranz, & Blaga, 2004; Hood et al., 1996; Slater, Brown, Mattock, & Bornstein, 1996). These factors are discussed in turn, after we briefly cover the goals researchers typically have with the habituation phase of the experiment.

Habituation/dishabituation procedures here are considered those involving 1) trials presenting the same stimulus/i sequentially during a training phase, and 2) test trials presenting a novel stimulus and the familiar. The goal of the habituation phase is that infants become less responsive towards a particular set of stimuli without disengaging from the task altogether due to fussiness or distress, for instance. The necessary balance between these two design demands makes optimal design choices important but very complex and context-specific. Typically, habituation studies make use of some criterion to determine whether infants have indeed arrived at a state of habituation - these criteria are discussed below. The application of any criterion suffers from false positives (i.e., reaching the criterion without being habituated) and (false) negatives (i.e., not reaching the criterion before getting fussy or distressed), and a well-performing criterion should reliably discriminate between the looking time distributions of

the infants who did and did not habituate. A habituation criterion that systematically yields small effect sizes (i.e., subtle differences in the looking times towards the novel versus the familiar stimuli) bears a higher risk of false positives and false negatives - especially in combination with other design factors that diminish the reliability of the measurement - because the looking time distributions of habituators and non-habituators in the sample will overlap more strongly. A meta-analytic approach is thus especially suitable for quantifying the relationship between habituation criteria and the effect sizes they tend to produce, while keeping the effect of other design factors fixed across infant habituation studies.

Whereas most studies anticipate a novelty preference during the post habituation (i.e„ test) phase, some report familiarity preferences (e.g., Fiser & Aslin, 2002; S. P. Johnson et al., 2009). The goal of the habituation phase design, however, is to maximize the likelihood of completing the habituation process - manifested as a novelty preference at the post habituation phase (Hunter, Ames, & Koopman, 1983; Hunter & Ames, 1988; Oakes, 2010). Thus, studies reporting on familiarity preferences likely employ design practices that tap into an earlier stage of the habituation process (Sirois & Mareschal, 2002). Below we discuss some of the design choices encountered in habituation studies.

6.1.3 Structural Design Factors

A typical habituation study involves the subsequent presentation of trials of one type - the to-be habituated stimulus/i, followed by a testing phase where looking times to novel and familiar stimuli are contrasted. Both novel and familiar trials need to be presented at post- habituation in order to assess the dishabituation effect (Oakes, 2010) because presenting only novel trials at post habituation may be, at least in part, affected by a regression to the mean (Ashmead & Davis, 1996; Dannemiller, 1984). Major design choices regarding the structure of the habituation task thus are:

1. the minimum and maximum duration of each trial,

2. the minimum and maximum number of trials that each infant is exposed to,

3. the number and order of test trials (familiar versus novel first, random versus blocked),

4. the number of trials to present before initiating the post-habituation phase (also covered in the discussion of habituation detection criteria below).

6.1.4 Stimulus Characteristics

A crucial determinant of looking times is the information value that the stimuli presented as novel and familiar carry (Kidd, Piantadosi, & Aslin, 2012; Richards, 2010). In this study, the focus is on structural factors of the habituation paradigm and we will hence leave factors such as stimulus type and complexity aside - also because these are hard to quantify and compare across a large set of studies. By abstracting over between-study variation in stimulus type and complexity, we are able to include a very large number of studies rather than constraining ourselves to a subset of studies using comparable stimuli. To provide a starting point for future investigations aiming to address the relation between stimulus characteristics and habituation performance, we gather superficial descriptives on the variability of stimulus characteristics such as stimulus modality and static versus dynamic presentation, as well as the page numbers on which images and schematics of the stimulus displays were reported in the original manuscripts so that these can be compiled into a database.

6.1.5 Looking Time Detection Methods

Another important factor pertaining to the precision and reproducibility of infant looking times obtained from habituation experiments is the looking time detection method. Earlier research using looking measures has relied on manual coding techniques (Aslin & McMurray, 2004; Oakes, 2012), in which at least one but preferably multiple raters record the infant's looking behavior either online, during the experiment, and/or offline from a video recording.

Recent advances in eye-tracking are becoming increasingly more prevalent in infant research, making looking time recordings automated, objective, and substantially more spatially and temporally precise (Aslin & McMurray, 2004;

Gredebäck, Johnson, & von Hofsten, 2009; Hunnius, 2007). Despite considerable perks, however, fully automated scripts are still not reliably implemented across habituation procedures (e.g., Cong et al., 2019; Oakes, 2012). A potential drawback of automated methods such as eye-tracking is that obtaining a signal of sufficient quality can be challenging, especially with young infants, which could lead to higher experimental attrition than what is typically observed with manual coding techniques.[1]

Habituation procedures that lead to higher experimental attrition due to fussiness can systematically bias samples (Slaughter & Suddendorf, 2007) and are therefore less advisable. Using manual looking detection methods, mean attrition rates (overall and due to fussiness) estimated from 143 habituation studies have been reported at 22.6% and 14.1%, respectively (Slaughter & Suddendorf, 2007). In contrast, a recent application of an automated face analysis tool to measure the looking times of 5 to 8-month-old infants from video footage yielded about 30% missing data relative to manually coded frames (Chouinard, Scott, & Cusack, 2019). To date, we are not aware of any systematic comparisons between eye-tracking and manual coding techniques in infant (habituation) samples; hence, we provide descriptives on current practices regarding the degree of automation in habituation experiments (such as looking time detection and stimulus presentation methods) and their relation to experimental attrition rate, as well as the strength of the novelty effect.

6.1.6 Habituation Detection Criteria

The single most important (and often discussed!) design factor is the use of a habituation criterion (Oakes, 2010). The logic of the habituation paradigm is that a decrease in interest by the infant signals that they have somehow processed the stimulus sufficiently to be able to discriminate it from other stimuli. The purpose of a habituation detection criterion is to then determine when a significant loss of interest has occurred in the infant such that the post- ha-

[1] Looking time measures are further sensitive to the specific data parsing algorithms and individual differences in data quality (R. Hessels, Andersson, Hooge, Nyström, & Kemner, 2015; R. Hessels & Hooge, 2019; Wass, Forssman, & Leppänen, 2014), though these issues are continuously being addressed (R. Hessels & Hooge, 2019; Leppänen, Forssman, Kaatiala, Yrttiaho, & Wass, 2015; van Renswoude et al., 2018; Wass, Smith, & Johnson, 2013).

bituation phase of the experiment can be initiated. A number of criteria have been applied throughout the literature that we broadly group under decrement, fixed, and model-based criteria.

Decrement Criteria

Probably the most commonly applied class of criteria are the decrement criteria (Horowitz, Paden, Bhana, Aitchison, & Self, 1972), which detect habituation when the 'final looking time' 'decreases' below a certain percentage of the 'baseline looking time'. A number of choices can be made in applying this criterion. First, the 'baseline looking time' can be defined as the looking time on the first trial or as an average of the first few trials; a common choice is to average across the 3 initial trials (Ashmead & Davis, 1996; Oakes, 2010) although other numbers certainly occur (Domsch, Lohaus, & Thomas, 2009). Similarly, the 'final looking time' is defined either as the looking time at the last trial or an average across a set of recent trials (Oakes, 2010). The 'decrease' in looking time is frequently set at a 50% decrease from the baseline to the final looking times (Ashmead & Davis, 1996; L. B. Cohen, 2004), but other percentages occur as well (e.g., Flom & Pick, 2012). Specifying these three variables already makes up for a large variability in the criteria being applied. One of our goals here is to report on how commonly different versions of the decrement criterion are encountered and what the consequences for the magnitude of the reported novelty effects are.

Yet another variant of the decrement criterion worth mentioning uses a similar definition of the looking time decrease and the final looking time but an alternative definition for the baseline looking time. Rather than the initial looking time, the maximum looking time (also referred to as the peak look) serves for calculating the percentage decrease from the final looking time. The reasoning here is that an infant's interest towards the stimulus can show an increase preceding the typical looking time decline pattern (Gilmore & Thomas, 2002; Hunter & Ames, 1988). Similarly to the initial and final looking times, the maximum looking time can be derived as an average across a single trial or a set of peak look trials (Colombo & Mitchell, 1990, discuss this variant). Another variation is to apply the criterion on the average looking time accumulated across all previous trials in the habituation phase.

Fixed Criteria

Under fixed criteria fall all procedures that require all infants to accrue a fixed number of trials or a fixed looking time before the testing phase can be initiated. Studies using fixed criteria are often referred to as familiarization studies, but for our purposes can still be considered habituation paradigms if a novelty effect is expected in the post habituation phase (e.g., Schlingloff, Csibra, & Tatone, 2020).

Model-Based Criteria

Other criteria for determining whether and when infants have habituated have been proposed too but are seldomly used. In particular, several authors proposed to substitute the classic decrement and fixed criteria with model-based criteria (Ashmead & Davis, 1996; Thomas & Gilmore, 2004). The authors defined mathematical models of the ideal habituation curve and proposed to use this to estimate when infants have habituated. Beyond reporting on the frequency of use, it is outside the scope of the current paper to discuss the application of these approaches (which is done in detail by Kucharský et al., 2022).

Relevance of Habituation Detection Criteria to the Novelty Effect and Attrition in Studies

Infants who have not fully habituated exhibit a familiarity preference (Hunter & Ames, 1988; Roder, Bushnell, & Sasseville, 2000; Rose, Gottfried, Melloy-Carminar, & Bridger, 1982; Sirois & Mareschal, 2002) and therefore contaminate the overall novelty effect if included in the analysis. In an often-cited paper, Oakes (2010) suggests several practices regarding the stringency of the habituation detection criterion, the use of sliding over fixed windows, and the maximum number of trials, among others, that are all thought to maximize the number of truly habituated infants in the analysis. Specifically, Oakes (2010) advises the use of a stringent criterion, operationalized as the 50% decrement criterion applied on a sliding window of three trials. She further proposes that 1) criteria requiring a larger decrement from the initial looking times are thought to yield larger effect sizes, though this relationship should be non-linear because criteria that are too strict would result in more infants who did habituate but

failed to reach the criterion (i.e., false negatives), 2) using sliding over fixed windows would result in lower attrition and false negatives because they pinpoint more precisely the timing of reaching the habituation criterion, a lower number of maximum trials would bear a lower chance of attrition due to fussiness overall and a lower risk of the infant reaching the habituation criterion by chance. One of our goals is to assess whether these recommendations are indeed borne out by evidence.

6.1.7 Crowd-Sourcing as a Data Extraction Approach for Review Papers in Infant Research

Over the past decades, the importance and utility of large-scale collaborative efforts have gained increasing recognition across scientific disciplines, including infant studies (Frank et al., 2017) - a field of research that is frequently confronted with the challenge of answering research questions in underpowered samples (Bergmann et al., 2018). This has inspired efforts such as ManyBabies (`manybabies.github.io`; Byers-Heinlein et al., 2020) and MetaLab (`langcog.github.io/metalab`; Gasparini et al., 2021; Tsuji, Bergmann, & Cristia, 2014; Tsuji et al., 2017) that aim to answer substantive questions with a greater degree of certainty. Whereas ManyBabies focuses on multi-lab replication studies, MetaLab offers repositories, templates, and tools for community-augmented systematic reviews and meta- analyses on central research themes in developmental linguistics and psychology.

Despite that a great body of empirical work using the habituation paradigm has been accumulated, very limited resources have been dedicated to help researchers make informed decisions about features of the experimental design that would allow them to assess the substantive research question at hand. Review papers typically aim at answering a substantive question and thus focus on a narrow, homogenous subset of literature. In contrast, a large body of literature needs to be taken into account to evaluate procedural and domain-specific variations across multiple disciplines that rely on the same phenomenon. Using crowd-sourcing as a general approach to review papers can greatly improve the replicability and generalizability of infant studies by allowing the systematic synthesis, meta-analysis, and updating of a much larger body of literature than what is typically feasible within individual research groups with localized

expertise. Here we outline and implement a crowd-sourcing data extraction workflow to a large dataset of infant habituation studies with the hope that others can adapt and build upon it. Similar crowd-sourced data extraction approaches for systematic reviews and meta-analyses have been applied to gene expression data in biomedical sciences (Mortensen, Adam, Trikalinos, Kraska, & Wallace, 2017; Strang & Simmons, 2018) and can be further applied to other comparable problems involving some sort of manual data extraction (e.g., corpus data).

6.2 The Current Study: Systematic Review & Meta-analysis

In a systematic review, we map out experimental design choices used in peer-reviewed habituation studies published in the period 2000-2019, including — but not limited to — the criteria applied to determine whether habituation has taken place. This part of the study can be considered a scoping review as it sets out to examine the range and nature of research activity to its fullest extent (Arksey & O'Malley, 2005). In a follow up meta-analysis, we assess how current experimental design practices in habituation studies relate to the research outcomes, with a specific focus on the association of these practices with 1) attrition rates and 2) the size of the novelty effect. Because the evaluation of study design choices is the central goal of this paper, certain features of the critical appraisal process (e.g., risk of bias assessment) common to systematic reviews and meta-analyses are postponed until after data extraction. To increase the reliability of the results, we use a crowd-sourced workflow in which raters willing to contribute to the data extraction process are recruited from the wider community of developmental researchers and each paper is randomly assigned for blind data extraction to at least two raters who subsequently discuss and resolve coding disagreements. Based on the descriptive (scoping), systematic review, and meta-analytic goals formulated below, we devise a set of recommendations for habituation studies that aim to provide researchers with an empirical base for designing their experiments, and to improve reporting standards.

6.2.1 Descriptive Goals & Hypotheses

In the descriptive analysis, we report on the distributions of current experimental design practices with regard to stimulus characteristics, the habituation task structure, presentation and response detection methods, and evaluate the extent to which the recommendations regarding habituation detection criteria and trial number proposed by Oakes (2010) are implemented.

Further, we provide an indication of the general data quality among the habituation studies included past the screening phase, such as descriptives on dropout rates and on inter-rater reliabilities and blind coding practices for the studies using manual coding of looking time.

6.2.2 Analyses of Attrition Rates: Goals & Hypotheses

Oakes (2010) suggests that more stringent habituation criteria will lead to fewer "habituators", but failing to exclude "non-habituators" may compromise the robustness of the novelty effects. Following the reasoning that habituation protocols aim to maximize the proportion of infants that reach habituation from the total number of infants included at the post- habituation phase (Hunter & Ames, 1988; Oakes, 2010), we evaluate empirically which experimental habituation design choices predict attrition caused by failure to meet habituation criteria. More stringent experimental designs, however, may also lead to higher experimental attrition (e.g., due to fussiness, etc.) which affects data quality. Thus, we need to evaluate which experimental design choices predict experimental attrition separately from habituation attrition - whereas the interpretation of habituation attrition is somewhat ambiguous because lower habituation attrition rates can presumably produce smaller novelty effects, minimizing experimental attrition is desirable in all measurement contexts. Figure 6.1 summarizes the relation among the types of attrition we analyze, highlighting that habituation attrition needs to be interpreted in tandem with the size of the novelty effect.

As discussed earlier, automatic response detection methods hold the promise to maximize measurement precision and the reproducibility of results, but whether their use introduces higher experimental attrition is unclear. We therefore compare the experimental attrition and attrition due to failed habitu-

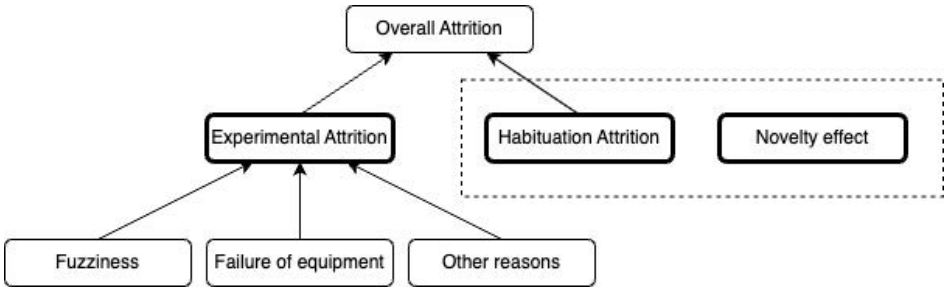


Figure 6.1: Breakdown of Overall Experimental Attrition Rate.

ation as a function of the response detection method: eye-tracking versus manual coding.

6.2.3 Meta-Analytic Goals & Hypotheses

In the meta-analysis, we focus on the relation between habituation criteria and the magnitude and direction of the effect size. Following the logic that if the habituation procedure is successfully implemented, novelty should be expected (Hunter et al., 1983), we ask three main questions:

1. What is the average size of the novelty effect during the post-habituation phase?
2. How are detection criteria related to the strength of the novelty effect?
3. How is age related to the strength of the novelty effect?

The specific hypotheses and statistical tests designed to answer these questions are summarized in Appendix A and under Modeling in the Results section. We compare across variants of percentage decrement versus fixed-type criteria used in habituation studies in predicting the presence and magnitude of the novelty effect as a function of age. and then zoom in on the evaluation of specific criteria. Variation in the sensory modality (visual, auditory, visual-auditory) in which the habituation stimuli were presented, as well as whether the habituation stimuli were presented statically versus dynamically will be incorporated as nuisance variables in order to control for between-study differences. Further, as our definition of habituation is relatively inclusive and

will result in a pool of studies that test other phenomena than a simple novelty/familiarity preference, we add a nuisance variable which codes whether or not the stimulus in the post-habituation phase is identical, or of the same type or category to the stimulus shown during the post-habituation phase. This allows us to distinguish between simple habituation studies and studies that also tap into further cognitive phenomena (e.g., categorization).

6.3 Method

6.3.1 Systematic Search

Search Strategy & Information Sources

The search strategy and syntax were devised in consultation with a librarian. We searched for peer-reviewed papers published in the period 2000-2019 on the topic of habituation in infancy (database-defined as 0-23 months), either in combination with a mention of a looking behavior response or experimental study design to reduce noise in our results. A recency bias is introduced by the time limits, which is acceptable for our study purposes to map out current research practices. Because the search strategy aimed to maximize the representativeness of the retrieved records, the systematic search was performed across two major databases in the field of psychology - PsycINFO and Web of Science, expecting that the majority of developmental science disciplines employing the habituation paradigm in human infants are to be found there.

The database searches were performed on May 21, 2020. The search syntax is available under Appendix B and online on the project's Open Science Framework (OSF) repository (`osf.io/cqvru`). Out of 3,858 results from the systematic search, 1,005 papers were manually removed in Zotero (v5.0.89; Ahmed & Al Dhubaib, 2011): 996 duplicates, 2 retracted papers, 7 "early access" papers. To update the search results with more recently published papers, we will rerun the same search once more immediately before the start of the data collection stage. The records of "early access" papers will be screened at this stage too in order to add the corresponding volume/issue details.

Study Selection & Eligibility Criteria

After de-duplication, the titles and abstracts of the remaining 2,853 papers were screened simultaneously for relevance by 6 blind raters (inter-rater reliability Fleiss' κ = .60, 95% CI = [.40, .80]) using the systematic review web application Rayyan (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016). To maximize consistency among raters during the screening process, articles were excluded in a hierarchical manner following the exclusion reason order provided in Figure 6.2.

We focused on including records describing original data on typical development. We further limit the scope of our study to 1) designs in which habituation is at least indirectly measurable or measured (e.g., measuring looking times on repeating trials), and 2) dishabituation is measurable (i.e., looking times toward the familiar stimulus/i in the post-habituation phase can be directly compared to looking times toward the novel stimulus/i in the post-habituation phase, as well as toward the familiar stimulus/i in the habituation phase; visual paired comparison and head turn preference study designs are thus excluded). Due to inconsistent use of terminology1, researchers may occasionally report using a familiarization paradigm whereas a habituation paradigm is actually being used. To increase the number of relevant habituation studies in our sample, records reporting on familiarization protocols were included as long as 1) more than one familiarization trials were presented during the habituation (pre-test) phase, and 2) novel stimuli were presented in the post-habituation (test) phase. During data extraction, the portion of papers in which researchers had an a priori hypothesis for a familiarity preference (and hence likely designed the procedure to target an earlier stage of the habituation process) will be excluded. Habituation protocols presenting tactile stimulation as the habituation stimuli were excluded because such protocols are very unlikely to elicit and measure changes in the infants' looking behavior responses that are comparable to those towards habituation stimuli in other modalities.

To maximize the amount of relevant search results, we added several synonymous key terms that are often used in the context of habituation/dishabituation procedures (e.g., discrimination, familiarization, visual preference, etc.). Nevertheless, studies that rely on habituation/dishabituation designs but do not explicitly mention "habituation" or any of the related key

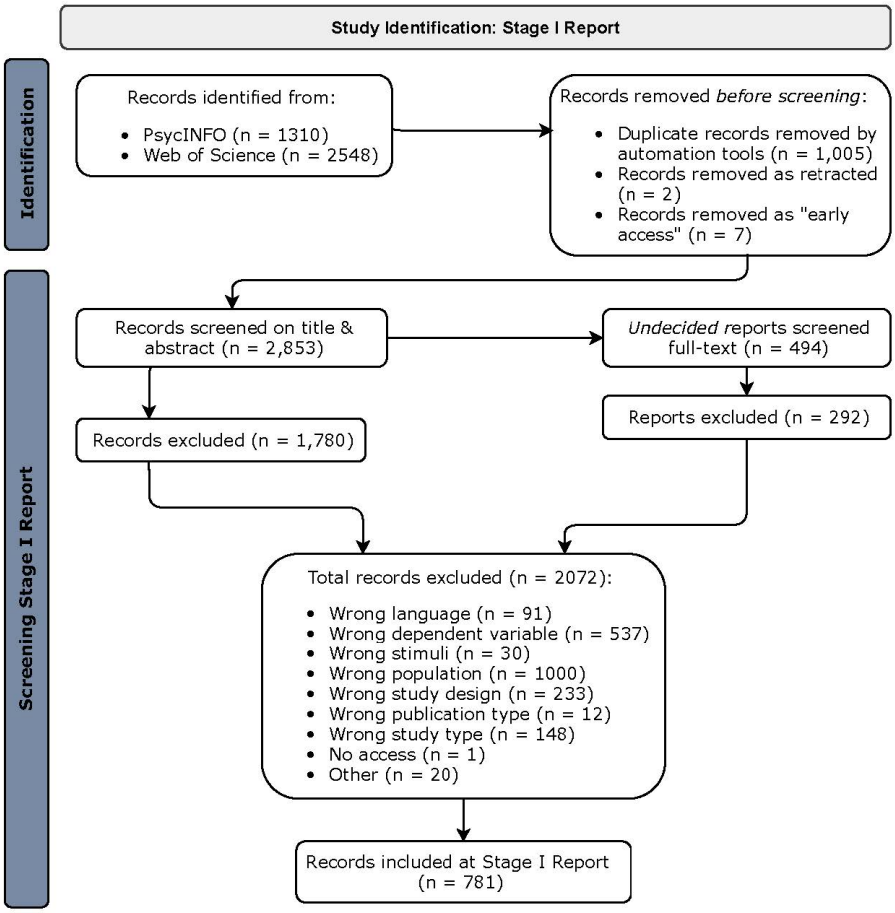


Figure 6.2: Study Identification Stages at Stage I Registered Report Outlined in the PRISMA Flow Diagram.

terms are omitted in our results. Studies using a head-preference to infer whether infants have habituated were also excluded. 781 (27.4%) papers describing original data from habituation studies using looking behavior measures in infants between 0 and 18 months were selected for data extraction. The PRISMA flow chart (Figure 6.2) outlines the number of papers at each step of the screening process.

6.3.2 Reporting Guidelines

Here we use the PRISMA 2020 (Page et al., 2021) and the NIRO (non-interventional, reproducible, and open systematic reviews; Topor et al., 2020) reporting guidelines for systematic reviews.

6.3.3 Developing the Coding Scheme

Prior to pre-registering this study, we conducted a feasibility pilot study that aimed to 1) select the set of variables on which data are extracted for our analyses, 2) define a variable coding scheme, and 3) devise a procedure for coding the data. Throughout, we were assessing the reliability of the coding to identify and resolve problems with the current coding. The initial feasibility study followed an iterative procedure. First, an initial list of variables that were coded and their definitions was implemented in a spreadsheet. Then, a random set of articles were selected and distributed among the authors of this study for initial coding. During this stage, the coders were able to communicate and exchange their thoughts on the coding and note down issues that were encountered (e.g., a missing level in a categorical variable). A new coding sheet was developed based on the initial experiences. Then, a set of 70 articles identified in the screening as relevant was distributed among seven raters such that each article was rated by at least three raters (this resulted in each rater being assigned 30 articles). The seven raters coded the articles independently to assess the reliability of the coding on all selected variables. Following the reliability results, additional changes were made to the coding scheme to standardize the data extraction of multiple experiments, experimental conditions, or age groups reported in the same paper, and the distinction between various habituation criteria, as well as introducing data input validation to minimize reporting errors. During this stage of

the project, we developed the coding manual following the PRISMA-P guidelines for systematic review and meta-analysis protocols (Moher et al., 2015) and prepared an online project for crowd-sourced data collection described below.

6.3.4 Crowd-sourced Data Collection

Data extraction from all papers included past the screening stage will be done using crowd-sourcing. Specifically, we will attract collaborators through developmental research mailing lists (e.g., the Cognitive Developmental Society (CDS), the International Congress of Infant Studies, and the ManyBabies mailing lists), inviting them to contribute to the data extraction as raters. Raters who subscribe to help with coding the articles will be invited to a Slack workspace for faster communication, and will receive further instructions with the data extraction workflow. We will employ data validation and provide coders with a coding manual to maintain the quality of the coding reliability, keeping the data organized and clean. Raters will be included as co-authors in the stage II report.

Use of `sysrev`

Sysrev (`sysrev.com`) is an online platform that facilitates the systematic reviews of

documents with multiple contributors. Articles identified in the screening stage will be uploaded to the sysrev project as `.pdf` files. Contributing raters will be required to create a (free) account on sysrev to log in. Sysrev will then automatically present the raters with articles that are not yet coded by more than one other rater. Raters will then proceed with coding the article directly on the sysrev platform. To limit human-error and ambiguity in the data extraction, we have implemented, where applicable: (1) multiple-choice items, (2) data validation checks that request the user to provide valid variable values whenever the input does not match the expected (e.g., checking whether user-entered values for numeric variables are indeed numeric), (3) the possibility to specify that data on the item is missing. Sysrev keeps the review in a database from which the data can be extracted in various standard formats.

Maintaining Reliability

To maximize inter-rater reliability, raters are provided with 1) a coding manual with instructions on how to extract information on the data items (including examples of edge-cases identified during the pilot stage) and how to resolve disagreements with other raters, 2) a separate sysrev project for training, containing five papers coded by consensus to help raters get acquainted with the data extraction process while using the coding manual. Upon joining the project, raters will be given access to the training and data collection sysrev projects along with the project's OSF repository containing the coding manual and the pdf's of the five training papers with the data extraction process annotated. The coding process will be performed incrementally to maximize the number of raters per article. The data will be subsampled by publication date in 5-year batches, prioritizing more recent reports because evaluating current practices is our primary focus. This results in the following record distribution: 188 records for 2015-2020; 230 for 2010-2015; 194 for 2005-2010; 169 for 2000-2005. Depending on the availability of contributing raters, we will determine whether each paper will be coded by at least two or at least three raters in the data collection sysrev project, and whether we let sysrev assign new raters with papers that have been already coded by at least one or two other raters respectively. This way, we hope to further maximize the number of raters per paper while making sure that raters coding the same papers are involved in the project at approximately the same time.

After completing the blind coding in the current batch, the data will be downloaded and analyzed for inter-rater agreement and specific disagreements. After the results are known, raters will be asked to resolve any disagreements with other raters by communicating via Slack or otherwise. For data items on which consensus is not reached, raters will inform the project coordinators, who will assign an additional coder to that article. The final value will be taken on a majority vote basis. If the disagreement could not be resolved even then (e.g., when the vote count is 1-1-1 for three alternative options), the project coordinators will then take the final decision on how to resolve the coding. At stage II, we will report on inter-rater reliability metrics, the variables on which disagreements arose, and the number of disagreements that required resolution by project coordinators. The analysis of inter-rater agreement after each batch will

be reported in the supplementary material, alongside an exploratory analysis studying whether the inter-rater reliability increased with increasing experience of the raters, providing some additional insights into the novel methodology of crowd-sourced review.

Recovering Missing Effect Size Information

The author(s) of studies for which essential data for computing the effect size is missing will be contacted. By failure to recover these data prior to initiating the meta-analysis, the effect size variables for these studies will be treated as missing.

6.3.5 Analysis Plan

The overview of the research questions, hypotheses, statistical analyses, and interpretation is provided in Appendix A.

Dependent and Independent Variables

An overview of the dependent and independent variables is provided separately for the descriptives (Table 1), and the meta-analysis and analyses of attrition rates (Table 2). The variables either overlap with or are composites of data items used for data collection in `sysrev`.

Descriptive Analyses

Before conducting any inferential statistics, the report will provide descriptive statistics and plots for all variables listed in Table 1. Categorical variables will be summarized in frequency tables, and continuous variables via summary statistics. These will be used to describe the typical trends and practices in habituation studies, with a focus on habituation criteria and their variations. Descriptive summaries will be also provided on the variability of experimental attrition rate (overall, due to fussiness, versus other causes lumped together), as well as the attrition rate due to a failed habituation phase (i.e., the infants excluded due to not meeting the habituation criterion). These summaries will be presented overall, as well as separately for distinct categories of presentation mode and looking time detection method.

Analyses of Attrition Rates

Some design choices may affect data quality. The current study focuses only on a small portion of possible data quality issues that may arise in infants research in general, and habituation experiments in particular (see Appendix A). Specifically, we will investigate whether there is any relationship between various design choices and the experimental attrition rate (i.e., whether there are any designs that are associated with fewer drop-outs). To this end, a binomial regression will be used with predictors corresponding to the degree of automation in the experiment: 1) presentation mode, 2) looking time detection method, and other variables such as: 3) stopping rule, 4) the maximum number of habituation trials) stimulus modality, and 6) average age, and a dependent variable "N drop-out total" out of "original sample N". Coefficients associated with a p-value less than 0.01 will be considered as statistically significant.

The attrition rate due to a failed habituation phase will be modeled with the same binomial regression, but with "N drop-out habituation" as the dependent variable. A Bayesian binomial regression will be used as a robustness check.

Whereas experimental attrition is generally undesirable, it is expected that a certain level of habituation attrition can lead to more robust results (Oakes, 2010). Therefore, while the desired outcome for the experimental attrition is to pinpoint which practices are related to lower attrition rates, factors related to the habituation attrition rates may need to be interpreted with the results of the meta-analytic results on the effect sizes (see below).

6.3.6 Meta-Analysis

Calculating Effect Sizes from the Extracted Statistics

The primary meta-analysis is based on the subset of studies that performed a comparison between novel and familiar stimuli in a within-subject design. The effect size of interest is the within-subject standardized mean difference (J. Cohen, 1988) - an effect size that reflects the standardized effect per participant, instead of the standardized difference between two measurements such as the more traditional Cohen's d (Dunlap et al., 1996; Morris & DeShon, 2002, see Section 6.3.6). This decision is based on the fact that the majority of studies use

a within-subject design to increase the power for detecting an effect, and that the novelty-familiarity preference is typically thought of as the average looking time difference between novel and familiar stimuli within the same individual, rather than the average difference between novel and familiar stimuli between participants.

The effect size of interest will be computed from the t-statistic and the associated degrees of freedom df as follows:

$$d_z = \frac{t}{df + 1} \tag{6.1}$$

The standard error of the effect size will be computed using the equation A1 from Morris and DeShon (2002):

$$se_d = \sqrt{\frac{{}^{df}/_{(df-2)}}{n}(1 + nd^2) - \frac{d^2}{c(df)^2}} \tag{6.2}$$

where $c(df) = 1 - {}^{3}/_{(4df-1)}$. The df in both equations is equal to $n - 1$, where n is the number of participants in the study. When an F-test is reported, the F-statistic will be converted to a t-statistic before converting it into the effect size using the formulas above.

Assessing Publication Bias

Publication bias will be assessed using standard procedures, as well as modeling approaches. Before analyzing the data using a meta-analysis, we will visualize the relationship between the effect size and standard error using a funnel plot. Further, we will compute the Bayesian Kendall's rank correlation test (van Doorn, Ly, Marsman, & Wagenmakers, 2018, 2019) to assess whether the standard errors and effect sizes correlate with each other. In case the Bayes factor favors the null hypothesis by at least a factor of three, we will conclude that there is low risk of publication bias. In case that the Bayes factor favors the alternative hypothesis by at least a factor of three, we will conclude that there is high risk of publication bias. We will also compute the Egger's test to assess the robustness of the correlation result. Additional analyses such as PET-PEESE meta- analysis selection models and robust meta-analysis (Maier, Bartoš, & Wagenmakers, 2020) will be run as exploratory analyses in case of doubt over the

robustness of the results against publication bias. In case of a high risk of publication bias, the meta-analytic models described in the Modeling section below will be limited. The models will be run anyway, but the results will be included only in the Supplementary materials and no conclusions will be drawn from these models. Instead, selection models and robust meta-analyses will be interpreted in the main text as exploratory analyses.

Modeling

Studies or conditions for which authors expected familiarity preference or a null result will be excluded from the primary analysis. First, fixed effects meta-analysis and random effect meta-analysis will be run to determine the average effect size of the novelty effect, the choice of fixed or random effects will be based on the omnibus test for heterogeneity. However, the primary meta-analysis will be a frequentist meta-regression. The variables used for predicting the size of the effect will be: 1) average age of the sample and 2) the stopping rule (dummy coded; fixed number of trials and fixed looking time are combined into one category, all three decrement criteria are combined into another category, and "other" with "model-based" criteria will be combined into the "rest" category). As nuisance variables, 1) stimulus modality (dummy coded), 2) presentation mode (dummy coded), and 3) identical stimulus (dummy coded) will be included in the model. The identical stimulus variable is included to distinguish studies whose focus is more constrained on the phenomenon of habituation by presenting exactly identical stimulus during the entire experiment from studies that involve ancillary phenomena (such as categorization) by letting the stimulus vary within a type or a category over the course of the experiment. For the main hypotheses, the coefficients of age and stopping rule will be tested individually at the $\alpha = 0.01$ level against the null hypothesis that the coefficients are zero. The effect of age is hypothesized as positive - therefore it will be tested with a one-sided test, and will directly answer the research hypothesis that age relates to the strength of the novelty effect. The effect of stopping rule is represented by multiple coefficients, each representing one category of the stopping rule criteria. We will conduct pairwise comparisons between the three combined categories ("fixed", "decrement", "rest") of the stopping rule. We do not have directional expectations regarding these pairwise comparisons.

If the p-value associated with the coefficient for the decrement criteria category is below 0.01, a follow-up analysis will be run to provide additional detail as to what makes the "% decrement criteria" more successful at giving rise to the effect sizes. The follow-up analysis will be run using the subset of studies that used one variant of the "% decrement criteria". Stimulus modality, presentation mode and identical stimulus will be modeled as nuisance variables. Additional variables included in the model will be 1) "% decrement N baseline trials", 2) "% decrement N criterion trials", 3) "% decrement percentage", 4) "maximum N habituation trials" "fixed vs. sliding window". Each of the variables is hypothesized to positively correlate with the effect size and therefore will be tested with a one-sided test at the $\alpha = 0.01$ level. The frequentist analysis will be carried out in the latest version of JASP (JASP Team, 2021). As a robustness analysis, a fixed effects meta-analysis, random effects meta-analysis, and another meta-regression will be performed in the Bayesian framework. A Bayesian analysis offers two main advantages here. First, the predictive performance of each model can be assessed, providing direct comparison between the fixed effects null model (effect size is zero across all studies), fixed effects alternative model (effect size is positive and fixed across studies), random effects null model (the average effect size is zero but varies from study to study), random effects alternative model (the average effect size is positive but varies across studies), and the meta-regression (explaining variability in the effect sizes as a function of the model predictors). Thus, the evidence in favor of the null model can be computed, which is otherwise impossible using the frequentist method. Second, a Bayesian meta-regression can be formulated to accommodate partially missing data per study. The primary analysis using the frequentist method to test the effect of individual characteristics of the decrement criteria requires to limit the data to the subset of studies using those criteria, therefore splitting the analysis in two steps. In the Bayesian analysis, those steps can be performed simultaneously within one model that also includes the variables for which a subset of the studies have missing data. For fitting the fixed and random effects meta-analysis, the metaBMA package in R will be used (Heck, Gronau, & Wagenmakers, 2019). For fitting the meta-regression model, a custom-coded Stan model will be used (B. Carpenter et al., 2017; Harrer, Cuijpers, Furukawa, & Ebert, 2021). Models will be compared using the bridge sampling technique

using the `bridgesampling` package (Gronau, Singmann, & Wagenmakers, 2020).

Additional Analyses

The primary analysis will be based on the standardized difference scores effect size across within-subject studies. Previous articles have warned against the use of this standardized effect size calculation for meta-analytic purposes (Dunlap et al., 1996). The difficulty with standardized difference scores is that it depends heavily on the correlation between the novel and familiar stimuli. A correlation higher than 0.5 results in larger effect size (Dunlap et al., 1996; Morris & DeShon, 2002) than that of the traditional between-subjects effect size (e.g., Cohen's d). Further, conducting a meta-analysis of the within-subject effect size requires the assumption that the correlation between novel and familiar stimuli is equal between all studies (Morris & DeShon, 2002). If that is not the case, one cannot interpret whether different effect sizes are due to the varying difference between the novel and familiar trials, or due to differences between their correlation. Similarly, an effect of a predictor in a meta-regression is difficult to interpret without this assumption; for example, an increase of the effect size with increasing age might mean that the correlation between novel and familiar stimuli increases with age, rather than that the difference between novel and familiar stimuli increases. However, it would be difficult to convert the standardized difference scores effect size into the classical Cohen's d or Hedges' g effect size, as that would either require studies reporting the means and standard deviations for novel and familiar stimuli, or the observed correlation between novel and familiar stimuli. The number of studies that do report these statistics will likely be very low.

Due to these difficulties, there will be a separate robustness meta-analysis that uses the Hedges' g effect size as the effect size of interest, from the group of studies that reported a between-subject comparison of novel and familiar stimuli, or those studies that reported a within- subject comparison, but did report the additional statistics required to calculate the between- subject effect size. The robustness analysis will be used to assess whether the conclusions from the primary meta-analysis hold with a different choice for the effect size. Because there will be missing data in the data set (see Section 6.3.6), we will

also provide exploratory analyses investigating whether reporting practices of habituation experiments improved over time.

The analysis of habituation attrition rates is ambiguous because it is not clear what level of attrition is needed to yield a valid sample of habituated participants at the post-habituation phase. A possible way to clarify this effect is to analyze the habituation attrition and novelty effect sizes in a joint meta-analytic model. In this model, habituation design factors can be used as predictors of both habituation attrition and the effect sizes at the same time. Further, a correlation between the two dependent variables will be estimated, which will suggest whether higher attrition rates correlate with higher effect sizes (Oakes, 2010).

Dealing with Missing Information

In descriptive analyses, the number of missing data entries for each variable of interest will be reported. In statistical analyses, only complete data (subject to the particular analysis) will be used for the main analyses. That is, we will use listwise deletion of rows in the data set based on only the variables that are entered into the specific analysis. Robustness analyses will be carried out by 1) coding missing information of categorical predictors as their own category, and by imputing the data using the mice package (van Buuren & Groothuis-Oudshoorn, 2011).

6.4 Results

[To be added after data collection]

6.5 Discussion

[To be added after data collection]

Open Practices Statement

Materials and code associated with this article are openly available at osf.io/cqvru. This article is a Stage 1 Registered Report accessible as a preprint at osf.io/bdtx9.

Appendix

For the sake of brevity, appendices are not included here. To access them, read the original preprint available at osf.io/bdtx9.

We are a little horse. A horse that still needs milk and to learn how to jump.

–José Mourinho

Chapter 7

Habituation, Part II. Rethinking the Habituation Paradigm

This chapter is published as Kucharský, Š., Zaharieva, M., Raijmakers, M., and Visser, I. (2022). Habituation, Part II. Rethinking the habituation paradigm. *Infant and Child Development*, e2383. doi: 10.1002/icd.2383

Abstract

The habituation paradigm has been applied to study the development of memory, perception, and other cognitive processes in preverbal infants, making it one of the most prominent experimental paradigms in infant research. However, there are many features of the process of habituation that remain elusive, which results in uncertainty about the best research practices.

This article first discusses current practices in habituation research (e.g., the use of habituation criteria) in relation to modeling the process of habituation, revealing several issues that impede progress in the field. To overcome these challenges, we propose to move towards a modeling framework to study critical features of the habituation process. To facilitate this transition, alternative experimental designs are proposed. The article encourages clearer thinking about the process of habituation, such that the theory, design, and analysis are all in line with each other.

The article concludes with concrete recommendations to improve current practices in infant habituation research.

7.1 Introduction

Infants show a gradual decrease of attention towards stimuli that are presented repeatedly — a phenomenon well known as habituation (Colombo & Mitchell, 2009; Fantz, 1964). Infant research uses the habituation phenomenon heavily as it allows researchers to unearth the development of capabilities such as learning and categorization, among others. Characteristics of the infants' looking behavior are often used as a proxy measure of habituation because eye movements are among the earliest overt behaviors to mature, making visual attention experiments especially feasible for studying the development of cognitive abilities from the first weeks of life to toddlerhood and beyond (Hunnius, 2007). Thus, the phenomenon of visual habituation is extensively used as a tool in infant research to study memory, perceptual, and cognitive abilities and their development, and became a prominent experimental paradigm in infant research (Colombo & Mitchell, 2009; Oakes, 2010). Although this article discusses visual habituation, it may be relevant for other dependent measures as well (e.g., Lloyd-Fox et al., 2019).

There is a great variety of implementations of the habituation paradigm (Colombo & Mitchell, 2009), which inspired the development of guidelines

for designing habituation studies (Oakes, 2010) and specialized software that fosters the adoption of these best practices (Oakes, Sperka, DeBolt, & Cantrell, 2019), following decades of theoretical, modeling, simulation, and empirical investigations into the topic (Ashmead & Davis, 1996; L. B. Cohen & Menten, 1981; Dannemiller, 1984; Gilmore & Thomas, 2002; Schöner & Thelen, 2006; Sirois & Mareschal, 2004; Thomas & Gilmore, 2004; Young & Hunter, 2015). However, there still appears to be a lot that is unknown about the habituation process itself as well as the consequences of experimental design choices used to make inferences about infants' abilities studied in the habituation paradigm. Given the past decade that led to the realization that psychological science may not be as credible as generally thought previously (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012), it is imperative to provide critical commentary of the current practices, and provide alternative ways of moving forward. The aim of this article is to contribute to this development.

The structure of this article is as follows. First, the article describes current practices in habituation research in terms of using infant-controlled designs in general, and habituation criteria in particular. This section identifies specific issues with these approaches and poses challenges that need to be resolved. Second, we present ideas that can lead the field to move towards resolving theses issues. In this section, we argue that more systematic investigation of the process of habituation is needed, identify three major theoretical questions, and present a modeling framework that enables us to resolve these questions. However, this proposal leads to changing the current habituation paradigm, and, as such, will bring about additional challenges. In the third section, we discuss some practical considerations for changing the habituation paradigm in anticipation of some issues that will arise as a consequence. The article ends with conclusions and recommendations.

7.2 Habituation as a tool

Historically, important work of Fantz (1964) is considered as one of the starting points in studying habituation in human infants (Tsang, 2012). The original methodology was as follows. An infant was repeatedly presented two stimuli at the same time, one novel, and one that remained unchanged between the trials.

The looking time to each stimulus was recorded. The general observation was that the amount of time the infant spent looking at the novel stimuli was longer relative to the amount of time spent looking at the familiar (constant) stimulus, thus demonstrating the effect of habituation towards the familiar stimulus. Although many studies targeted the habituation phenomenon as the center of the interest (e.g., Kagan & Lewis, 1965), the phenomenon appeared so reliably and in early stages of infant development, that it presented an opportunity to use habituation in experiments that do not necessarily aim to study habituation in and of itself, but as a research tool to study infants' abilities such as memory, learning and categorization (e.g., Fagan III, 1970; Horowitz, Paden, Bhana, Aitchison, & Self, 1972; Horowitz, Paden, Bhana, & Self, 1972; Wetherford & Cohen, 1973). The core logic is that if an infant displays a habituation effect, they display an ability to remember the familiar stimulus. Further, the fact that the infant shows increased looking times to a novel stimuli shows that the infant is able to *discriminate* between the two stimuli, or eventually, being able to distinguish between different categories of stimuli (Oakes, Madole, & Cohen, 1991) or being able to discriminate between different exemplars from the same stimulus category (Quinn, Eimas, & Tarr, 2001).

However, empirical data do not always suggest that infants prefer the novel stimulus, and occasionally produce effects in the opposite direction. The factors that lead to one or the other effects are to this day under active investigation. One influential explanation for these two phenomena is the model proposed by Hunter and Ames (1988). The authors proposed that the infant's attention towards the habituation stimulus initially increases and only then shows a monotonic decline pattern over repeated exposure to the stimulus. The speed at which attention rises and declines can be moderated by various factors such as age, familiarization time, stimulus complexity (relative to infant's cognitive ability), etc. Therefore, whether one observes preference for the novel or the familiar stimuli is entirely dependent on the interplay between these factors and the design of the study. This theoretical explanation is soon to be subjected to a systematic exploration in the ManyBabies 5 initiative (`manybabies.github.io/MB5/`).

The original methodology introduced by Fantz (1964) (in today's terms called *a visual paired comparison* paradigm) also had some drawbacks. One

of the more important concerns was that it is unclear whether longer looking times towards the novel stimulus relative to the familiar stimulus are caused by a novelty preference or, on contrary - by an avoidance of the familiar stimulus. Further, by allowing only a relative comparison, the visual paired comparison offers only indirect insight on how infants habituate over time in the absolute sense. As a result, alternative methodologies were developed (Aslin, 2007; Colombo & Mitchell, 2009; Turk-Browne, Scholl, & Chun, 2008). Although there exists a large variation between these methodologies, there is one aspect that is common to most of them: Separation of the experiment into a habituation (familiarisation) phase, and a test (novelty preference) phase. During the first (habituation) phase, infants are repeatedly exposed to a stimulus with the goal of habituating the infants to that stimulus. After this phase is completed, the novel stimulus is introduced which commences the test phase. The purpose of the test phase is to compare infants' reaction to the novel versus the old stimulus. A novelty effect occurs when the response toward the new stimuli exceeds that towards the old stimuli, and is interpreted as evidence that the infant was able to recognize the systematic differences between the old and the new stimuli. Instead of the term "novelty effect", researchers sometimes use the term "dishabituation". However, the term dishabituation can be sometimes used to refer to a phenomenon where the attention to the old stimulus spontaneously recovers to the initial level. Due to the ambiguity of the term dishabituation, we will use the term "novelty effect" throughout the article. The opposite effect is called a "familiarity effect".

Separation of the experiment into two phases gives an opportunity to assess two questions independently: 1) did habituation occur during the habituation phase, and 2) did a novelty (or a familiarity) effect occur. For example, Quinn, Eimas, and Rosenkrantz (1993) reported a study (Experiment 1 in the cited article) in which infants were shown pictures of cats (or dogs) in 6 habituation trials of 15 seconds each, and then presented side by-side with pictures with birds during the test phase – the main hypothesis being that the infants would prefer looking at the pictures of birds compared to cats (or dogs), which would demonstrate their ability to discriminate between the two categories. Quinn et al. (1993) found evidence for the novelty effect, despite not being able to demonstrate that the infants habituated (i.e., decreased their attention) to the familiar

stimuli during the habituation phase. This finding was somewhat surprising because the novelty effects would be expected only if the infants habituated to the old stimulus in the first place. There may be various plausible explanations of these findings, but the critical point of this example is that it is possible to make such observation simply by dividing the experiment in two phases.

In an ideal world, removing the influence of the moderating effects could be done by making sure that the habituation phase of the experiment is long enough for every participant to safely habituate. In reality, however, infants are likely to drop out from lengthy experiments (Slaughter & Suddendorf, 2007). This can lead to high attrition rates among the participants, which can be possibly related to the developmental stages of the infants (Hunnius, 2007). Habituation studies thus balance on a thin line, seeking to minimize the length of an experiment to maximise participants' comfort and decrease the attrition rate, all the while making sure that habituation was sufficient so as to adequately test the question of interest (Peterson, 2016).

To this end, the so called "infant-controlled habituation protocols" (Horowitz, Paden, Bhana, Aitchison, & Self, 1972; Horowitz, Paden, Bhana, & Self, 1972) became a common practice in the literature (Colombo & Mitchell, 2009, p. 228). In these protocols, infants' reaction to the stimulus determines the pace and course of the procedure, and allows researchers to stop the experiment early, once the infant is deemed to be habituated.

Over the years, researchers developed several simple decision rules (henceforth "habituation criteria") designed to detect that habituation was sufficient. Typically, the stimulus is repeatedly presented in a sequence of independent trials. Attention towards that stimulus is recorded either manually or using eye-tracking technology, usually by using the time it takes the infant to look away from the stimulus at each trial as a measurable proxy. During this procedure, attention towards that stimulus usually drops gradually. The mean of the most recent x number of trials ($\bar{x}$) is compared to the mean of the y number of criterion trials ($\bar{y}$). When the looking times drop below a threshold – when $\bar{x}$ is less than $z\%$ of $\bar{y}$ – researchers conclude that habituation was successfully induced and the habituation phase of the experiment can be concluded. Varieties of this procedure can by composed by varying 1) the number of trials x and y, where increasing these numbers is thought of as increasing the precision of the

procedure at the expense of making the experiment longer, 2) the set of trials used to define the "criterion". Commonly, the criterion trials are simply the first trials shown to the infant. However, alternatives exist such as taking the y trials with the highest looking time, inspired by the Hunter and Ames (1988) model, which suggests that the trials that attract the most attention may not be the first trials, nor the y trials that immediately preceded the x most recent trials. Lastly, 3) the threshold percentage drop in looking time z can be varied so to decrease or increase the sensitivity of the procedure, thus affecting the type-I (incorrectly declaring habituation when the infant has not habituated) and type-II errors (incorrectly missing out on an infant that has in fact habituated). The most commonly used value is 50%. Probably the most prevalent "initial 3-3-50%" criterion is one where 50% of the mean across the initial three trials is compared to the mean of the three most recent trials (Oakes, 2010), although variations of the settings of the criteria vary across the literature (e.g., Domsch et al., 2009; Flom & Pick, 2012). An ongoing systematic review of the existing literature will quantify the size of this variation (Zaharieva et al., 2022).

Most of the criteria that are commonly used today are some variations of heuristics developed decades ago (Ashmead & Davis, 1996; Dannemiller, 1984; Thomas & Gilmore, 2004), and are in spirit similar to performance criteria in learning experiments (Bogartz, 1965; Dannemiller, 1984) in that they are thought of as standardization which ensures that the level of habituation[1] is equated across participants (Peterson, 2016). Some of these criteria are currently readily available to the empirical researcher using dedicated software for creating habituation experiments (Oakes et al., 2019). Today, habituation criteria are considered an established part of habituation designs, and resulted in an extensive track record of studies.

7.2.1 Issues with the infant-controlled paradigm

Despite their popularity, habituation criteria have some drawbacks. Some of these issues were thoroughly debated and documented in previous literature (L. B. Cohen & Menten, 1981; Dannemiller, 1984; Thomas & Gilmore, 2004). Habituation guidelines have reflected on some of the issues advising how to

[1] In this article, we use the term "habituation level" to mean "the degree to which an infant is habituated at a certain point in time."

use them (L. B. Cohen, 2004; Oakes, 2010). However, we will argue in the following section that the issues with habituation criteria are deeper than what seems to be accepted in the applied literature, so much so that perhaps one ought to question *whether* to use habituation criteria in the first place rather than *how* to use habituation criteria.

Performance of habituation criteria

The use of habituation criteria in experiments assumes that these criteria successfully measure habituation levels[2], that these levels are possible to compare, control, and equate between infants (Dannemiller, 1984), and that habituation criteria have adequate error rates in order to "filter out" infants that do not habituate but to retain infants that do habituate (Oakes, 2010). Habituation criteria have a large intuitive appeal. However, it is prudent to verify the underlying assumptions. One of the options to assess performance of the habituation criteria is by running simulation studies. This section discusses findings of previous simulation work that has brought largely negative results, casting initial doubt on whether the habituation criteria serve their intended purpose.

Dannemiller (1984) conducted a simulation study to assess the performance of the initial 3-3-50% habituation criterion. The model that generated data in this simulation assumed that looking times at trial t can be decomposed into a signal and Gaussian noise component. The signal was modelled as an exponential decay function that starts at point g and decays with a rate r towards an asymptote k. By varying the three parameters of the signal component and the amount of noise (specified by the variance of the Gaussian noise), Dannemiller (1984) was able to find that the false alarm rate (i.e., the probability that the data reach the criterion given there is no decay in attention over time increases rapidly when the noise in the data increases. Using this criterion also leads to stopping the experiment too early (before reaching the desired 50% habituation level). The variance of the criterion is also high, suggesting that controlling or "equating" the level of habituation between participants is sub-optimal. Fur-

[2] We will set aside that habituation is putatively measured in terms of visual attention, which is further supposedly measured by looking times, despite the fact that there may not be one-to-one correspondence between covert and overt attention (Posner, Snyder, & Davidson, 1980; van der Stigchel & Theeuwes, 2007).

thermore, under certain conditions, the true positive rate of the criterion (i.e., the probability that the data reach the criterion given that there is a decay in attention over time) increases as noise increases; or in other words, data with low amount of noise can lead to difficulties to detect habituation. We will return to this issue in more detail in section 7.2.1. In sum, the criterion was shown to have a high Type-I error (declaring habituation when none occurred) and a theoretically possible high Type-II error (missing habituation when it occurred), and in the cases for which habituation is correctly detected, the habituation levels were not properly equated between participants.[3]

Ashmead and Davis (1996) conducted a similar simulation study to reach similar conclusions as Dannemiller (1984). Additionally, they found that the test-retest reliability of the habituation criterion is very low, under certain conditions bordering with zero (low reliability, albeit larger than zero, was also found empirically, McCall & Carriger, 1993). Because unreliable measures increase the noise in the data (Ashmead & Davis, 1996), low reliability further reduces the power to detect the experimental effect, such as the novelty effect. However, issues with low power are of course influenced by various design choices, other than the habituation criteria alone. These design choices that affect power are out of the scope of this article (DeBolt, Rhemtulla, & Oakes, 2020; Oakes, 2017; Visser et al., 2023).

In sum, the evidence from the few simulation studies evaluating the performance of habituation criteria does not suggest that the criteria have desirable properties. This does not necessarily mean that the criteria are not fit to their purpose, only that evidence in favour for the criteria from simulation studies is limited or inadequate. However, simulations come with their own limitations (e.g., selecting a data generative process, appropriate parameter settings), choices that can all influence the results with regards to the performance of the criteria, as well as the selection of the exact criterion that is studied. Thus, citing precise numbers obtained in the simulations are not of much importance as it is uncertain to what extent the patterns in the synthetic data resemble those encountered in empirical data, or that, perhaps, a slightly tweaked criterion could

[3]Ironically, the article by Dannemiller (1984) is sometimes cited as a canonical reference of the 50% decrement criterion, despite that even the abstract concludes with "the results preclude the use of trials to criterion as an index of rate of habituation" suggesting that the author himself was not convinced of the criterion's utility.

render better results. Further, the problems of habituation criteria and their performance to some extent rely on noise. Even if the simulation studies were accurately calibrated to reflect the amount of noise in the empirical data, it does not necessarily mean that when simulations show inadequate performance of the criteria, the criteria themselves are inadequate — it could only mean that it is necessary to reduce noise. Perhaps by reducing the measurement noise, by developing experimental protocols that reduce variance of infants' looking times, or by other means that can reduce the amount of noise in the experiment.

Conceptual issues

Simulation studies may have not provided strong evidence in favour of the habituation criteria. However, we can continue by critically examining their assumptions. Here, we discuss that the assumptions come with some underlying conceptual issues.

Arguably, the most problematic feature of habituation criteria relates to the finding that was first discussed by Dannemiller (1984): In certain scenarios, it is very difficult to successfully detect habituation *unless* the data contain a lot of noise, meaning that the data pass the criterion capitalizing on chance. To explain this phenomenon, Thomas and Gilmore (2004) noted the following observation. Typically, habituation criteria check whether the mean of the last x trials is less than $z\%$ of the mean of the first y trials (or the y successive trials with the largest mean). This at minimum requires the assumption that when a participant is "fully habituated", their looking times are lower than $z\%$ of the mean of the first y trials. Ideally, if habituation criteria are supposed to *equate* habituation levels between infants (Colombo & Mitchell, 2009; Dannemiller, 1984), this requires the assumption that when fully habituated, a person will either no longer look at the stimulus, or at least that all people drop down to the same level of attention (in proportion to their initial levels). However, this assumption is likely false — even a fully habituated person would still look at the stimulus for some period of time and there probably is substantial individual variation (Dannemiller, 1984; Gilmore & Thomas, 2002; Thomas & Gilmore, 2004). For this reason, most models assume that even after a point of habituation is reached, there is still some attention paid to the stimulus (L. B. Cohen & Menten, 1981; Dannemiller, 1984; Thomas & Gilmore, 2004). Performance

of the habituation criterion depends crucially on the degree to which residual attention is exhibited compared to the initial attention level and the noise in the data. In cases for which the residual attention is larger than the z% of the initial level, the criterion would rarely detect habituation without enough noise in the data, hence letting the data meet the criterion by chance. Figure 7.1 provides an illustration. All realisations of the habituation curve in the figure show a decline in attention levels. However, the criterion is unable to detect habituation under conditions of low noise and high residual attention level. The same level of residual attention passes the threshold only if the noise is high enough to bring the running mean below the threshold by chance, thus demonstrating the counter-intuitive effect that when residual attention is relatively high, decreasing noise in the observations increases the Type-II errors. In principle, the Type-II error rates of the habituation criteria can vary between $\approx 0\%$ and $\approx 100\%$ depending on this feature of the habituation process (Thomas & Gilmore, 2004).

The possibility of high Type-II error rate is important, as it is one of the crucial ingredients for evaluating the criterion's positive predictive value (PPV), i.e., the percentage of infants that truly habituated to the desired level in a pool of infants that were classified as habituators. The PPV is perhaps the desired metric for evaluating the classification properties of the criteria. PPV, in addition to estimating the sensitivity and specificity of the criteria, also requires an estimate of the true proportion of infants that do habituate. However, habituation criteria are usually intended for filtering out participants that did not habituate to the desired level (Oakes, 2010), and the discussion is mostly skewed by the concern of reducing false positives only. Minimizing false negatives (Type-II errors) and estimating the proportion of infants that habituate is sometimes of secondary concern in the habituation literature. Unfortunately, the risk of encountering high Type-II error due to high residual attention levels is rarely discussed or investigated in the habituation literature, beyond the mentioned theoretical and simulation articles, and the discussion of what proportion of infants are capable of habituating arguably receives even less attention. Thus, to our knowledge, PPV of the currently used habituation criteria under realistic scenarios have not yet been quantified or extensively studied.

Taken together, these results do not permit us to conclude whether or not

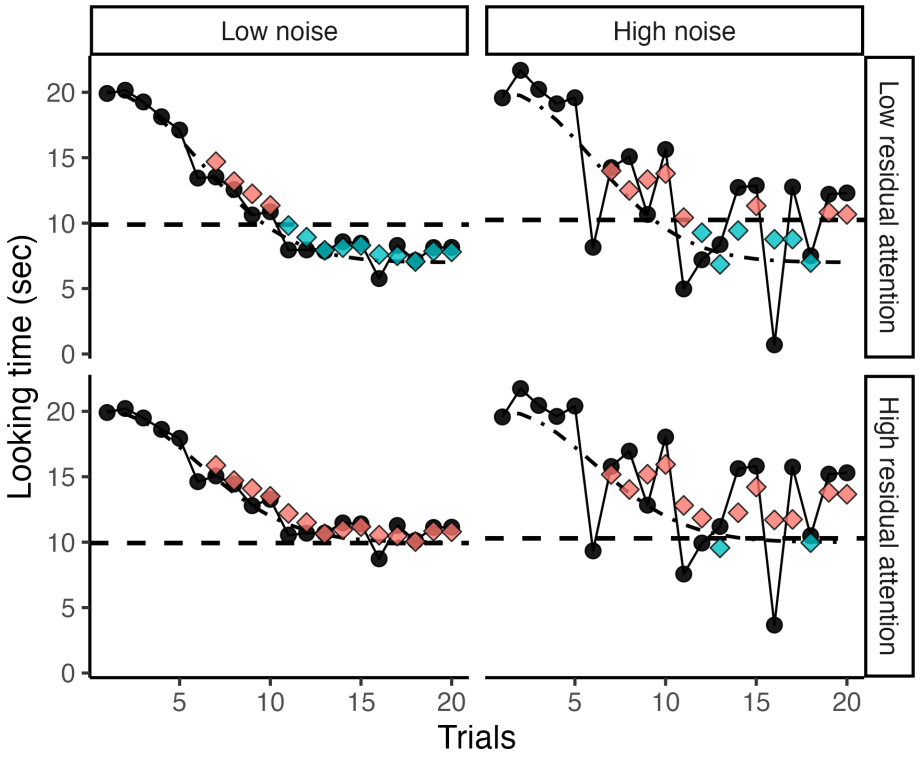


Figure 7.1: Four realisations of the model of habituation (Thomas & Gilmore, 2004) evaluated by the "initial 3-3-50%" criterion. Black dots connected with lines show the "observed" looking times. The dashed curve shows that the underlying habituation process across observations is the same but is manifested through different looking time errors per observation. Diamond shapes indicate the calculated mean of the most recent three trials (i.e. a diamond at trial t is the mean of the looking times at trial t, $t-1$, and $t-2$.). Red diamonds indicate that the running mean did not reach the threshold calculated as 50% of the mean of the first three trials (dashed line), blue diamonds indicate otherwise.

habituation criteria are suitable for their purpose. Previous simulation work suggests that under certain scenarios, they do not work very well. However, it may be that the vast majority of empirical applications of habituation criteria fall into a region of the parameter space where the criteria just about work. Unfortunately, the current empirical literature does not seem to provide much evidence to tackle that question.

Conditioning on the dependent variable

Aside from habituation criteria, other approaches to infant controlled designs were proposed using statistical models (Ashmead & Davis, 1996; L. B. Cohen & Menten, 1981; Dannemiller, 1984; Gilmore & Thomas, 2002; Thomas & Gilmore, 2004; Young & Hunter, 2015). These statistical models are based on the idea that infants' behavior can be described as a functional process where attention (measured by looking times) evolves over the course of the experiment. These methods are able to generate simulated data and as such were used in the simulation experiments that evaluated habituation criterias' performance. The other characteristic of these statistical models is that they can be fitted to empirical data to make inferences about the individual participants. These models allow more flexibility and control over the underlying assumptions, and can be used to make probabilistic instead of binary statements about whether or not infants habituate, and so could in principle overcome the conceptual limitation of habituation criteria discussed above (Gilmore & Thomas, 2002; Thomas & Gilmore, 2004). As such, these models were proposed as an alternative to habituation criteria and to make inferences about population (group) level effects as well.

Paradoxically, the general idea of using infant controlled designs hinders our ability to use these models to carry our inferences. As described previously, standard habituation designs aim to detect habituation as early as possible (Oakes, 2010). The consequence is that data sets from different infants have different number of administered habituation trials, depending on their behavior. Thus, it is impossible to know how the data would have evolved, had the experiment continued towards a fixed number of trials, or had a different habituation criterion been applied. Further, data from infants that do not show substantial decrease of attention are often discarded from the data

set (Oakes, 2010). If the method used for determining whether or not to administer additional trials works as intended, the amount of data from different participants is related to the characteristics of the infants' looking behavior that is being modeled. As a result, fast habituators are represented by less trials than slower habituators, and even slower habituators than that are omitted from the data set. If we wanted to use such data to make inferences from these statistical models, the analysis could be biased by such selection bias, where missing data are directly related to the dependent measure (i.e., missing data are not missing at random Steyvers & Benjamin, 2019). This issue will be further elaborated in Section 7.3.4.

Thus, if one wanted to use statistical models for making inferences, these inferences would be biased if any method for determining whether or not to display additional trial, or whether or not to exclude the participant from the data set, given infants' looking behavior, was used during the data collection.

7.2.2 Conclusions

The present section discussed challenges that occur with the practice of using infant-controlled paradigms. It seems that the validity and accuracy of the proposed methods are not currently being established empirically. This does not necessarily mean that the current methods are not fit to their purpose. After all, habituation criteria have been used in a variety of studies and they often do seem to produce desired effects. However, this does not give us much information about their performance in specific applications, and does not give us the ability to predict what kind of criteria to use in what context.

Simulations that would be useful for validating these methods are limited as one has to decide on the data generative process used in the simulation. To produce realistic simulations, one would have to decrease our epistemic uncertainty about the process of habituation and compare alternative models of habituation. However, the very use of infant-controlled methods in data collection possibly introduces a bias of data missing not at random, which can lead to misleading conclusions from the statistical models.

In light of these issues, it is worth questioning whether we should, as a field, reconsider the habituation paradigm. That means at least momentarily postponing the use of habituation criteria or other methods proposed for infant-

controlled studies, until we accumulate enough understanding of the process of habituation. In what follows next, we propose alternative ways to study and use the habituation phenomenon in empirical research while avoiding potential problems that we just discussed.

7.3 A way forward

The previous section described challenges with the current habituation paradigm. To propose a positive outlook, the current section proposes alternatives that hopefully help the field to move forward.

Specifically, first we turn attention to what features of the process of habituation need to be empirically established and verified. In this section, we provide an example that demonstrates the use of habituation models to answer questions about the process of habituation more directly, and how to evaluate individual differences between participants. This example also demonstrates some of the previously described issues with the infant-controlled paradigm.

Second, we discuss how the alternative paradigm we propose in this article could lead back to using the habituation phenomenon "as a tool", instead of studying habituation only in isolation.

As already suggested, one of the underlying themes of this article is that perhaps our understanding of the *process* of habituation has not advanced yet up to a point of confidence that we can measure, control and equate habituation levels between infants. A natural solution to this problem would be to turn the focus towards studying the characteristics of habituation itself, instead of relying on it as a tool with poorly understood workings. Such thinking is of course not novel and has a significant tradition in the literature, spanning several decades of theorising (see Colombo & Mitchell, 2009, for an overview) and formal modeling (Ashmead & Davis, 1996; L. B. Cohen & Menten, 1981; Dannemiller, 1984; Gilmore & Thomas, 2002; Schöner & Thelen, 2006; Sirois & Mareschal, 2004; Thomas & Gilmore, 2004; Young & Hunter, 2015).

7.3.1 Theoretical models of habituation

An important facet of understanding the process of habituation is to try to explain the phenomenon in terms of its cognitive or biological underpinnings.

Theoretical accounts of habituation have a long history. Here we briefly summarise some of the important theoretical explanations of the habituation phenomenon and how they can contribute to our understanding of the phenomenon. For more detailed overview, see Colombo and Mitchell (2009).

A popular theoretical account is one of Sokolov (1963), building upon the discovery of orienting reflex (Pavlov, 1927), and explains habituation in terms of a mismatch between the perceived stimulus and its internal representation in memory (Sokolov, 1977). Repeated presentations of the stimulus allow to form more accurate representations, leading to a lesser mismatch, thus decreasing the neuronal response to the stimulus over time. This theoretical account paved the way to link habituation to learning, as individual differences of the speed of habituation is thought to be associated with their ability to learn from the environment (form internal representation of the stimuli).

Additional theoretical model by (Jeffrey, 1968) proposed that the stimuli are not processed uniformly the same upon each presentation, but different features may be processed at different times (depending on their importance based on e.g., visual saliency), which can potentially explain additional individual variation. Such account calls for studying looking behavior in a greater visuo-spatial detail than only analyzing overall looking times. Despite that early evidence shown mixed results (Lasky, 1979; Leahy, 1976; Miller, 1972), the emergence of novel dynamic models of eye movement behavior (e.g., Kucharský, van Renswoude, Raijmakers, & Visser, 2021; Le Meur & Liu, 2015; Malem-Shinitski et al., 2020; Schwetlick et al., 2020) may prove to be useful to revisit such alternative explanations.

A dual process theory in turn posited that the habituation phenomenon is driven by two somewhat opposing processes (Groves & Thompson, 1970; Thompson & Spencer, 1966). In this model, one process is responsible for the decrease of the response to the stimulus over time, whereas the other is responsible for sensitisation, thus allowing potential increase (or spike) in the response, which is also reflected by the theoretical account of Hunter and Ames (1988) where response to the stimulus may initially increase before an eventual drop off. Model of Groves and Thompson (1970) also makes the prediction of "spontaneous" dishabituation effects where relatively novel variation of the old stimulus "resets" the habituation process, thereby restoring the attention

to initial levels.

An important class of theoretical process models are formal mathematical models that are drawn from broader theoretical framework. These formal models can generate data which can be compared with patterns in empirical observations to check plausibility of the models. Further, the models can be used to generate specific (perhaps not obvious) predictions. Empirical work can then aim to verify whether such predictions realize in a carefully designed experiment to elicit such patters. Notable works in this respect was published by Sirois and Mareschal (2004) which used connectionist approach and Schöner and Thelen (2006) who used dynamic field theory to link observable looking time patters to underlying neurological and cognitive mechanisms.

7.3.2 Statistical models of habituation

Another class of formal models are models that could be described as *statistical* models (Ashmead & Davis, 1996; L. B. Cohen & Menten, 1981; Dannemiller, 1984; Gilmore & Thomas, 2002; Thomas & Gilmore, 2004; Young & Hunter, 2015). These models are useful because not only because they can generate specific predictions that can be subsequently verified by empirical observations, similarly to the formal theoretical models discussed in the previous section, but can be also fit to the data to make inferences about the individual infants, or population (individual) effects. As such, they have been suggested as an alternative to habituation criteria, as was already discussed in Section 7.2.1, but also as an alternative to classical statistical approaches to carry out inferences (Ashmead & Davis, 1996; L. B. Cohen & Menten, 1981; Thomas & Gilmore, 2004; Young & Hunter, 2015).

Compared to theoretical models of habituation, statistical models of habituation often abstract away from some of the theoretical considerations, or in other words, do not usually attempt to explain the data in terms of the underlying biological or cognitive mechanisms. Nevertheless, these models are often carefully designed to reflect implications from theoretical models (e.g., Gilmore & Thomas, 2002; Thomas & Gilmore, 2004). By virtue of being statistical models, a variety model fit measures may be used to determine potential sources of misfit to empirical data, which can inform the initial theoretical assumptions.

In practice, statistical models have the potential to inform experimental design, data collection, analysis, and interpretation. However, to this date, modeling approaches to run and analyze habituation studies have not been widely adopted in empirical research. A potential reason why these models have not replaced habituation criteria in data collection is that the model estimation needs to take place online. While that became feasible with the increase of computational resources due to the advance of technology, it may have been a severe obstruction in the past. Further, the implementation of such procedures requires some statistical and programming expertise, and modeling approaches have not been made readily available to the empirical researchers as opposed to the habituation criteria (Oakes et al., 2019) and traditional statistics usually used to carry out the inferences.

These two issues would be soluble by implementing a modeling approach in an easy to use habituation software (e.g., Oakes et al., 2019). However, the question then becomes, which of the models should one choose to implement?

A promising approach would be to run simulation studies, similar to previously discussed simulation studies that verified performance of habituation criteria. However, simulation studies will be only helpful to the extent that the data generative process that is being implemented resembles the process induced in real-world labs.

Paradoxically, the abundance of proposed models of habituation does not make it easier to adopt this approach in practice. Similarly as in the case with the various habituation criteria, there is no clear guideline which model to use for empirical research and the validation of the models is largely limited. There is only little work dedicated to *assessing the performance* of different models of habituation. Typically, modeling articles present their own view of the problem, apply it to one or two data sets, and leave the problem be without much discussion and comparison to alternatives. Sometimes, longer term projects would have short series of articles that dedicate some effort in developing a single modeling paradigm (e.g. Dahlin, 2004; Gilmore & Thomas, 2002; Thomas & Gilmore, 2004). However, a recurring theme is that these models are seldomly compared to each other across varied data and applications, making it difficult to assess what model or theoretical standpoint is actually favorable when confronted with empirical data. Simply put, although there are plenty

of models, methods, and theoretical frameworks upon which to build a new experimental paradigm of habituation, the degree to which these approaches are corroborated by the data is relatively low, and so is the uncertainty high with regards to which framework can replace already established criteria.

Ideally, to come up with realistic models of the habituation process, one would first have to subject to analysis many habituation data sets coming from a plethora of tasks, designs, and populations. Using these data, process models of habituation proposed in the literature (Ashmead & Davis, 1996; L. B. Cohen & Menten, 1981; Dannemiller, 1984; Gilmore & Thomas, 2002; Schöner & Thelen, 2006; Sirois & Mareschal, 2004; Thomas & Gilmore, 2004; Young & Hunter, 2015) can then be compared. Those models that capture patterns in the data the best would be corroborated, and those that do not will either be updated or eventually dropped. Based on the estimated parameters (and their variability) of the final models, more realistic simulations can be carried out.

However, as we discussed in Section 7.2.1, data collected using some kind of infant-controlled paradigm induces data missing not at random, and as such may bias the analyses. Thus, it is important to use experimental designs that are not conditioned on the infants' behavior. An example of such design would be a fixed number of trials protocol, which is commonly referred to as familiarisation design.[4] The research goals of familiarization studies does not typically rotate around the question of whether individual infants have habituated however. Further, familiarisation studies often present the stimuli for very few trials such that any kind of modeling will be a difficult task. For the moment, let's set aside this issue and assume that we have enough diverse data sets that are not conditioned by the infants' behavior and that are fit to be used for modeling exercises. This section presents ideas about how to study habituation in fixed number of trials designs. We will discuss the practical implications for data collection in Section 7.4.

With the end goal of proposing alternative ways how to use habituation "as a tool", we first discuss how can the abundance of existing statistical modeling

[4]Sometimes, "familiarisation" studies can also employ infant-controlled aspects, such as when the experiment runs until the infant accumulates predetermined time spent looking at the stimulus, which can take more or less time depending on how much the infant pays attention to the stimulus. These experiments can be also considered infant-controlled and are not part of the current discussion.

approached be combined and compared. In order to foster model comparison, we identify three central features of the habituation process that are desirable to capture with formal statistical modeling what we discuss in the next three sections in terms of empirical questions. This allows us to propose general modeling framework that encompasses most of the proposed models to date, and follow with an example. Based on this modeling framework, we continue by discussing how to go beyond modeling habituation itself to using modeling to use habituation as a tool.

Question 1: Who habituates?

The habituation phase aims to ensure that the infant habituates to a stimulus (i.e., decreases the level of attention towards it) so that we can test for a novelty-familiarity effect during the post-habituation test trials. However, because not all infants show a conclusive habituation pattern, habituation criteria are used to "filter out" non-habituators under the premise that there is no reason to expect novelty-familiarity effect in infants who did not habituate in the first place (Oakes, 2010). Usually, participants that did not habituate according to some criteria would be excluded from the experiment and their data will often remain unreported. Thus, it is hard to judge how many infants habituate, how many infants do not habituate, or how many infants would be classified as "habituating" under one criterion but not under a different criterion.

It is crucial to point out that there is a large confusion about what it means if someone is said to "habituate". That is, habituation is sometimes referred to as a *state* of attention that is sufficiently low compared to the initial level of attention — a view consistent with the habituation criteria, where an infant is called a "habituator" if they pass the habituation criterion. However, habituation can be also thought of as a *process* where a person who is *habituating* shows a decrease of attention over time. This distinction is important because it is possible for someone to be habituating (i.e., show a decrease of attention over repeated exposure), but may not achieve a desired decrease of attention within a specific time. While this person would be classified as "non-habituator" under the view that habituation is simply a state of attention at a specific time, but under the lens of the process of attention, the person is habituating.

Within formal models of habituation, it is possible to sharply distinguish

between these habituation patterns. This has been already suggested some forty years ago by L. B. Cohen and Menten (1981):

> The experiment we would propose would run each subject for a fixed trial length (e.g., 20 trials), [...] However, one possible outcome of such an experiment is the result we have obtained: that neither model predicts the obtained curve. This would then necessitate a more elaborate model [...], or perhaps a multi-population model containing, for example, "habituators" and "non-habituators".

First, it is possible to encounter a sub-population of infants that does not show decline of attention over time, while other sub-population do. This alternative can be represented by "multi-population" (mixture) models, as L. B. Cohen and Menten (1981) mentioned in the quote above. This view has been applied in some of the modeling approaches (Young & Hunter, 2015); however, mixture models that can represent two distinct populations of "habituators" and "non-habituators" can be worked into the most of the other habituation models as well. Second, within the group of "habituators", it is possible to estimate the degree to which a person is habituated. Thus, instead of focusing on whether or not someone habituated, we may instead focuse on the exact level of attention at a certain point in time (Gilmore & Thomas, 2002; Thomas & Gilmore, 2004).

Thus, separating those two views of habituation enables us to shift focus towards different kinds of questions: 1) Who habituates, that is, which infants show a general decline in attention and which do not (if any), and 2) The speed of the habituation process, that is, how fast or to what extent an infant habituates, and what is the estimated attention level at a particular point in time. These two questions can be further related to other empirical questions, such as whether the proportion of infants that are habituators increases with age, etc.

Question 2: What is the shape of habituation?

Apart from the fact that it is uncertain who habituates and who does not, there is a considerable uncertainty and discussion about the shape of the habituation curve (Ashmead & Davis, 1996; L. B. Cohen & Menten, 1981; Dannemiller,

1984; Thomas & Gilmore, 2004; Young & Hunter, 2015). Specifically, there is a debate whether decline of attention during habituation experiment can be described by a parametric function, and if so, what shape of the function best describes the process, and to what extent there may be individual differences.

There is an alternative argument to be explored too of whether simple parametric functions can describe between and within person variability in the habituation process, or whether a more complex mechanistic model needs to be used to cover all corner cases of real-world behavior (Schöner & Thelen, 2006; Sirois & Mareschal, 2004).

The present modeling framework proposes to focus only on parametric functions, as those are easily combined and compared statistically, and will hopefully lead to sufficient description of patterns in the data. Example questions that can be answered by investigating the shape of the habituation curve are:

1. Is habituation a continuous process (Ashmead & Davis, 1996; Dannemiller, 1984; Gilmore & Thomas, 2002; Thomas & Gilmore, 2004; Young & Hunter, 2015) or a discrete phenomenon (L. B. Cohen & Menten, 1981)?

2. Is there a limit of "residual attention" levels or does attention level eventually decrease to zero (Thomas & Gilmore, 2004)?

3. Does attention always decrease over time, or is there a spike following the initial trials (Hunter & Ames, 1988; Thomas & Gilmore, 2004)?

4. What is the general speed of habituation process? Relatedly, how many trials does it take for an infant to habituate up to a certain habituation level (Thomas & Gilmore, 2004)?

As with the previous question of "who habituates?", the answers to such questions may be later explored in relation to other factors, such as age, the nature of the stimuli, etc.

Question 3: What is the distribution of looking times?

Lastly, there is some uncertainty around specifying correct distributions for looking times (Csibra, Hernik, Mascaro, Tatone, & Lengyel, 2016). Specifying a distribution that captures the main characteristics of the looking time

outcome can improve the performance of many models. Previous findings (L. B. Cohen & Menten, 1981; Csibra et al., 2016) have shown that looking times can be very skewed with a variance that increases with the mean, which is a typical phenomenon for positive-only random variables. Accordingly, it has been advocated to analyze looking times on a log scale using the logarithm as a transformation that stabilizes the variance and reduces skew (Csibra et al., 2016; Young & Hunter, 2015). On the other hand, reasoning about looking times in terms of their logarithms is more difficult than un-transformed looking times, and the habituation curves are usually defined on the natural scale. Thus, modelling approaches that aim to capture the shape of habituation curves usually resort to modelling the un-transformed looking times using a normal distribution for convenience. Other applications assumed Gamma distribution (L. B. Cohen & Menten, 1981).

The statistical models of habituation are relatively flexible, in a sense that various distributions can be applied. While the distribution of looking times is mostly a nuisance factor in the sense that it is not the primary focus of any analysis, determining which distributions capture the data best may improve the model performance and in the long term, improve our inferences.

7.3.3 Combining questions

As mentioned previously, the questions we outline here are not independent of each other. For example, to investigate whether there is a sub-group of infants that do not show a decline in attention, one needs to define the shape of the habituation curve for the group of infants that do habituate, and define the distribution of looking time so that a statistical model can be fit.

The idea behind a coherent modeling framework is as follows: One may define an ensemble of models that represent various combinations of the theoretical alternatives. These models may be fit to the empirical data simultaneously. Then, their predictive performance may be assessed in order to either perform model selection or model averaging (Hinne et al., 2020). Then, the inference may be carried out at the population level (e.g., what is the proportion of infants that are "non-habituators", or "what is the average speed of habituation") or at the individual level (e.g., "what is the level of attention of an infant following the habituation phase"). For the sake of brevity, the current proposal for

all models in question is not detailed in this article. Interested reader may find additional information at the project's open repository: osf.io/jr76y/.

7.3.4 Example: Modeling habituation

To demonstrate how one can study the characteristics of the habituation process without relying on habituation criteria, the present section shows a simulated example using an extension of the habituation model as proposed by Thomas and Gilmore (2004), which was already used as an illustration in Figure 7.1. This model may be further extended to include 1) some proportion of the infants that are "non-habituators", i.e., those that do not decline in attention at all, and 2) individual differences such that habituation trajectories may vary between participants, while adhering to the general habituation curve prescribed by the model. This model then allows inference at the population level (e.g., the average speed of habituation), and at individual level (e.g., the habituation level of a particular infant after a particular trial). More details about this simulation are provided in Appendix 7.A.

For the present illustration, we use parameter settings that are generally favorable for the use of habituation criteria - relatively low residual attention levels compared to the initial attention level, and relatively low noise in the observations. Further, in the current settings 90% of the population is set to be habituating (i.e., have their attention decrease according to the Thomas and Gilmore (2004) model), whereas 10% are non-habituators (i.e., their attention level remains at a constant). A realisation instance of this model, assuming 100 participants who undergo 20 habituation trials each, is displayed in Figure 7.2. The data were generated as follows: First, the infants were randomly assigned to the "habituator"/"non-habituator" sub-populations. In this instance, 89 infants were habituating and 11 non-habituating. For each of the habituating infants, individual parameters for initial attention levels, residual attention levels, the speed at which attention decreases over time, and the amount of noise were randomly drawn from their population distributions. Using these parameters, the data were simulated according to the Thomas and Gilmore (2004) model. For the non-habituators, only a constant attention level and noise were randomly drawn from the population distributions, i.e., the data was simulated as the constant attention level plus noise.

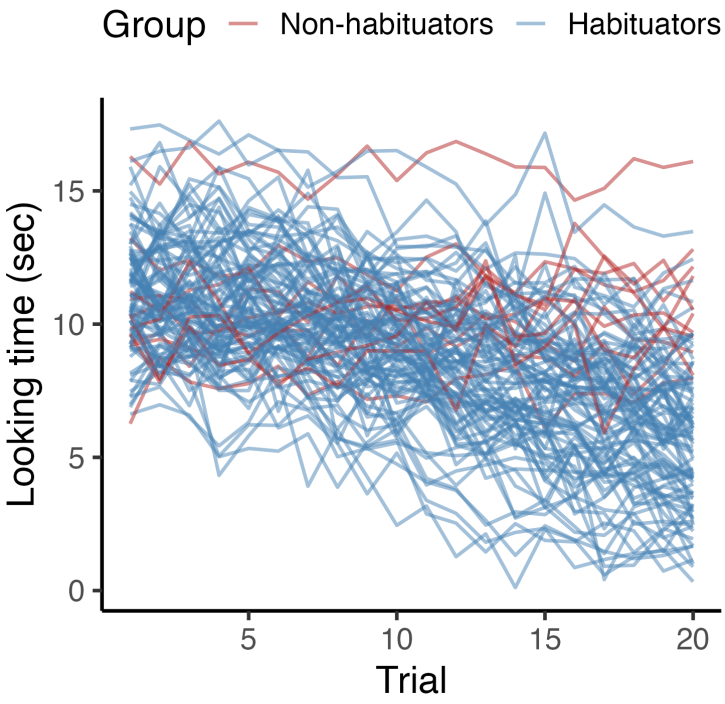


Figure 7.2: An example of simulated data from a mixture model of habituation. Each line presents looking times from an individual infant. Blue lines are "habituators" whose data are generated from the Thomas and Gilmore (2004) model, whereas red lines represent "non-habituators" whose data are generated as a constant plus noise.

Upon fitting this model back to the simulated data, we may discover that the estimates are relatively close to the true parameter values, suggesting that the parameters can be recovered accurately, see Table 7.1. In real situation, one would not commit to a single model before collecting the data, but would rather consider all models that are theoretically defensible – some alternatives were discussed in the previous section. Instead of fitting a single model, one would fit multiple models and perform model comparison or model averaging (Hinne et al., 2020). For the sake of brevity, we will not perform the model comparison here.

However, fitting habituation models to empirical data is difficult when the

				95% Credible interval	
Parameter	True Value	Mean	SD	Lower	Upper
μ_γ	2.30	2.32	0.07	2.17	2.46
σ_γ	0.25	0.23	0.05	0.15	0.34
μ_α	0.00	0.28	0.15	-0.05	0.55
σ_α	0.25	0.26	0.09	0.09	0.46
μ_β	2.30	2.26	0.03	2.20	2.32
σ_β	0.25	0.21	0.02	0.18	0.25
μ_δ	-6.00	-6.07	0.10	-6.27	-5.90
σ_δ	0.80	0.83	0.07	0.70	0.99
μ_σ	-0.25	-0.19	0.03	-0.26	-0.13
σ_σ	0.30	0.29	0.03	0.24	0.35
π	0.90	0.85	0.04	0.77	0.91

Table 7.1: True value and parameter estimates of the example model. See Appendix 7.A for explanation.

data were collected using habituation criterion. To illustrate how applying a habituation criterion hinders modeling approaches, let us imagine using the popular initial 3-3-50% criterion applied to the data shown in Figure 7.2. The data left after applying the criterion are shown in Figure 7.3. From the perspective of the recommendations for testing novelty effects using habituation criteria (Oakes, 2010), the criterion works quite well in that all of the infants that pass this criterion come from the population of "habituators", from which only one infant did not reach a 50% habituation on the latent level. Further, the Spearman's correlation between the number of trials it took to reach the habituation criterion and the speed of habituation δ is high (rho $= 0.873$), indicating that the criterion within this parameter setting corresponds quite well to the latent levels of attention. The only issue in the present example according to Oakes' (2010) recommendations is that the criterion is perhaps too strict, as it excludes an unnecessarily large number of participants that were actually habituating (54 out of 89).

These results already hint at the biases that are introduced when applying habituation criteria. First, as the number of trials correlates with the speed of habituation, fast habituators are represented with less trials in the data, which

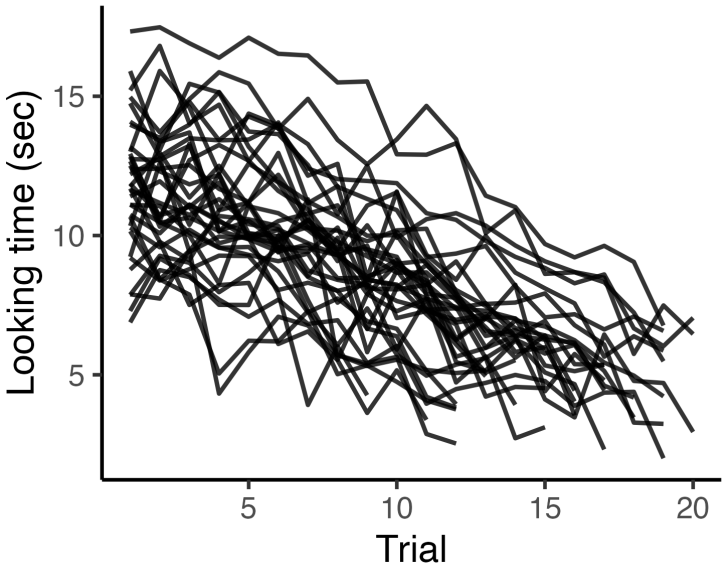


Figure 7.3: An example of simulated data from a mixture model of habituation after applying the 3-3-50% habituation criterion. Each line represents looking times from an individual infant.

will result in an underestimation of the habituation speed. However, infants who habituate too slow to be detected within the maximum number of trials are completely excluded from the data, which results in a bias in the opposite direction — the estimated speed of habituation will be pushed towards higher values than those generating the data. These biases act as a counterweight and the resulting bias entirely depends on the interaction between the experimental procedure (i.e., type of habituation criteria used and the maximum number of trials) and the parameters of the habituation process. In the present example, fitting the habituation model on the data set filtered through the habituation criterion results in overestimating the speed of habituation in the population. Whereas the true data generative process implies that the mean number of habituation trials that are required to reach 50% habituation is 17.72, the fitted model would claim that the number of trials is substantially less (12.44, 95%CI [11.15, 13.82]). However, it is in principle possible to construct a scenario where the bias goes in the opposite direction, which occurs when only negligible number of "slow habituators" are excluded entirely, but there is still

correlation between the number of trials administered and the speed of habituation (e.g., increasing the maximum number of trials would have that effect in the present example). The bias should decrease once the habituation criterion becomes unrelated to the habituation process, which essentially excludes data at random, though there would be little reason to use habituation criteria in such situations in the first place. All in all, if one wants to fit habituation models to empirical data, one cannot use data collected with habituation criteria, as those would result in biased estimates.

But let us revisit the scenario without using the habituation criterion. We already showed that when fitting the model back to the simulated data, we can recover the population parameters. In addition, we can interrogate the model to answer specific questions about individual differences. First, we may ask whether the model was able to distinguish the group of "non-habituators" and "habituators". Using the parameters estimated by the model, we can compute the posterior probability for each infant belonging to one sub-population or the other. Using maximum a posteriori probabilities, we may then classify the infants into the two groups. In the present simulation, there were 13 "non-habituators" of which 11 were correctly classified as such, and 2 were incorrectly classified as "habituators". All 87 "habituators" were correctly classified as such. Further, one difficulty with presenting a fixed number of trials to every infant is that we do not have control over the final habituation levels, which means that using this procedure to test for a novelty effect seems difficult, as it does not enable us to "equate" habituation levels between infants. However, instead of trying to equate infants in terms of their habituation levels, we may embrace the fact that there are individual differences and study them as such. To facilitate that, we can estimate the extent to which an individual infant has habituated during the habituation phase of the experiment directly from their data. Figure 7.4 shows that the true values are correctly recovered by the model in the simulation. This quantity then may be used in follow up analyses to account for the fact that the habituation levels are not equal across participants.

7.3.5 Beyond habituation

So far, the article focused only on what happens during the habituation phase of the experiment, and how to study the process of habituation more efficiently

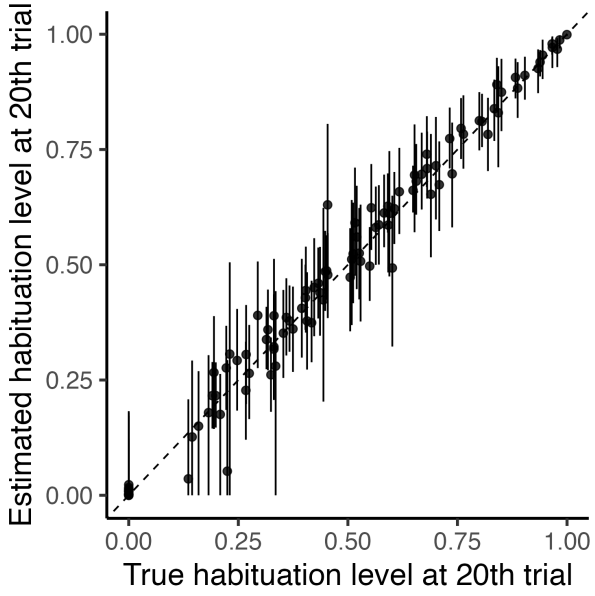


Figure 7.4: The true habituation levels at the 20th trial for each participant, plotted against the estimated habituation levels. Error bars display the 95% Credible intervals.

by reducing bias introduced by the infant-controlled paradigm. However, as was mentioned in the introduction, habituation is often not the primary phenomenon of interest, but is rather used as a tool to study other phenomena. A common design for habituation studies that seek to find a novelty effect, and present a set of novel and a set of familiar stimuli after the habituation phase (commonly known as the "testing" phase) to investigate whether or not the attention towards the novel stimulus is larger than that to the old stimulus. This section explores alternative designs to demonstrate how modeling approaches unlock additional possibilities for testing the novelty effect.

The first (and simplest) alternative is to run the habituation phase for a fixed number of trials, and then immediately follow up with a testing phase - a design commonly known as a familiarization study. However, instead of simply testing whether novel trials differ from the old trials, one can first estimate habituation levels per infant (presented in Figure 7.4) and use these as a covariate when testing for a novelty preference. Adding habituation level as a covariate

would enable researchers to take into account the natural variation in habituation levels and gain more power in the design by explaining away variability at the baseline. Thus, instead of trying to *equate* habituation levels experimentally, in this design we attempt to account for *in-equal* habituation levels that occur simply because infants differ between each other.

However, alternative designs can be used instead. Consider, for instance, that researchers want to establish whether a group of infants of certain age can discriminate between two types of stimuli. In the traditional habituation or familiarization designs, this question would be answered by testing for a novelty effect. However, instead of asking the question "is there a novelty effect", or possibly "what is the size of the novelty effect", we may test the question of distinguishing between the two stimuli more directly, by turning the test towards individual infants.

To illustrate this idea, consider an experiment where stimulus 1 is presented for a fixed number of trials, and is then replaced by a second stimulus for a fixed number of trials. If an infant distinguishes between the two stimuli and also habituates in general, we would be able to see two separate habituation curves: one for each stimulus. If an infant does not distinguish between the two stimuli, we would observe that the infant continues to habituate as if the first stimulus was unchanged. These two patterns assume that the infant does habituate (i.e., shows a decrease of attention over a repeated presentation of a stimulus). We already discussed the possibility that some infants may not show this pattern, and so it may be the case that there are infants that do not habituate but distinguish between the stimuli, and infants that do not habituate and do not distinguish between the stimuli. Figure 7.5 shows an example of the patterns that would arise from this experimental design under the four theoretical alternatives.

Thus, instead of asking about the average size of the novelty effect, we may first ask how many infants belong to each group (habituators vs. non-habituators crossed with discriminating vs. not discriminating). This would make the inference more person-centered, highlighting that, in addition to quantitative differences, there might be also qualitative differences between these sub-populations (Haaf & Rouder, 2019).

It may be difficult to detect which infants distinguish between the two stim-

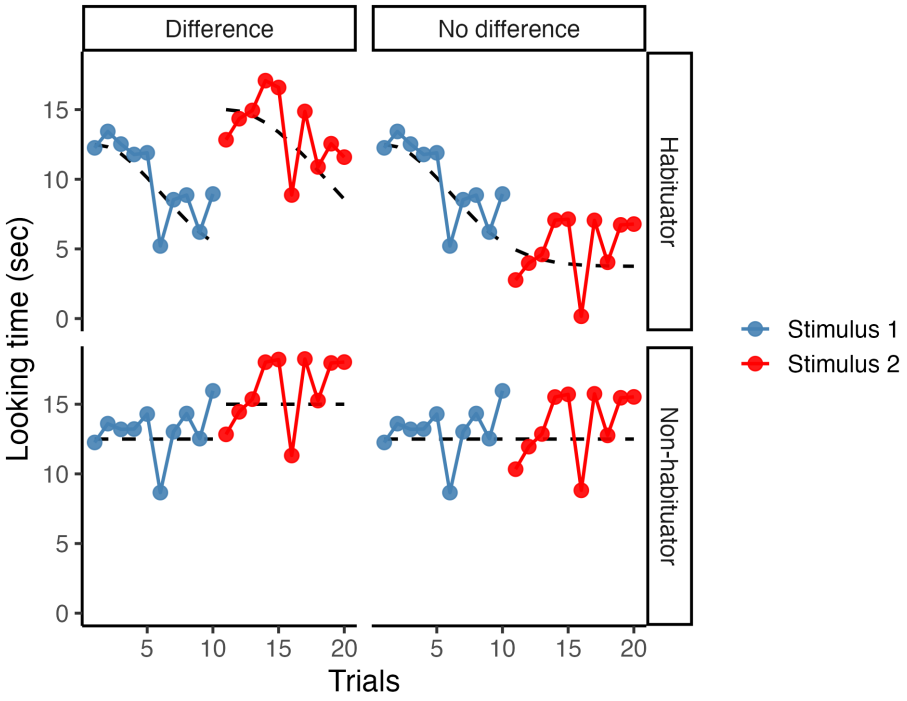


Figure 7.5: Four alternative outcomes of the sequential habituation phases design. In the first row, participant is habituating, i.e., their attention decreases with repeated exposure to the stimuli, whereas in the second row, the participant is not habituating. In the left column, the participant is able to distinguish between the two stimuli presented, whereas in the right column the participant does not. Colored dots joined by lines are the simulated looking times, and dashed lines show the evolution of the underlying attention levels.

uli among non-habituators because, in the absence of a habituation pattern, the differences between attention levels may be too small. Thus, the final model could focus only on the sample split as illustrated in Figure 7.6. First, we ask the question which infants show a habituation pattern and which infants do not. Then, we may ask which infants that do habituate also show the ability to discriminate between the stimuli and which do not. Additionally, we may estimate individual differences between infants and the variability of the stimuli, as well as practice and fatigue effects if the stimulus presentation order is counter-balanced or randomized across the sample.

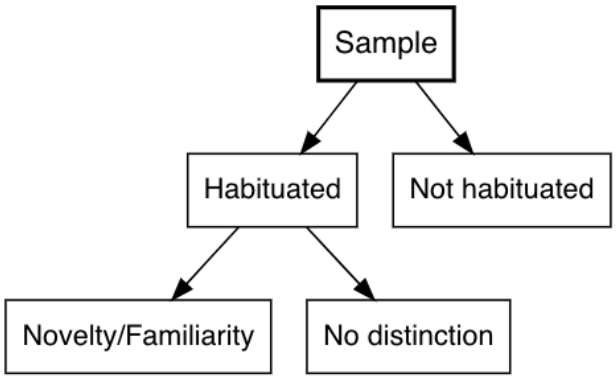


Figure 7.6: Conceptual illustration of types of events that can happen in habituation experiments.

The presented ideas need to be validated empirically, and it is possible that such designs would not be suitable for every research question that is currently tested using the habituation paradigm. However, we hope that those ideas provide inspiration to researchers to go beyond the established experimental designs, possibly by leaning on the power of formal models to construct more informative designs that tackle their research questions more directly.

7.4 Practical considerations

The previous section of this article presented experimental designs alternative to those that are currently being used in the habituation literature. Changing experimental paradigms, however, comes at a cost — whereas it may be relatively straightforward to replicate a study protocol that has been already implemented in the past, implementing novel approaches need additional time, resources, and effort to balance the design and solve unforeseen issues in order to collect data of sufficient quality.

In anticipation of such issues with the implementation of new designs, the current section provides several suggestions and answers to questions that may arise.

7.4.1 Fixed number of trials designs

One of the appealing features of the habituation criteria is that they enable researchers to terminate the experiment early, thus not requiring the infant to sit through the experiment in its maximum length. More broadly, this is the aim of any method that is considered "infant-controlled", and so includes modeling proposals such as that of Thomas and Gilmore (2004). From this view, presenting the experiment for a predefined number of stimuli may seem unethical, as abandoning the infant-controlled territory implies that some infants will be put under the laboratory settings for longer period of time than if some dynamic criterion was used.

However, as this article argued, our understanding of the infant-controlled methods is limited, and the very use of such methods precludes us from gaining more insight into the phenomenon itself. Thus, it is currently uncertain whether the criteria work as desired in the first place, and it is very hard to determine the benefits for individual infants from cutting the experiment shorter. It stands to question, then, whether minimizing the experiment duration for individual infants outweighs the fact that the data from those infants are hardly usable for any purpose other than simply assessing the presence of a novelty effect, especially given our uncertainty of the underlying mechanism itself.

Employing fixed number of trials designs will necessarily bring the question of determining an optimal number of trials to be presented to each participant. Models of habituation benefit from larger number of trials, but empirical research is of course constrained by the infants' capacity to undergo lengthy experimental procedures. An important distinction needs to be made here. The issue of infant-controlled studies is not necessarily that the number of trials varies between participants, but that the number of trials and the exclusion of the data is directly related to the habituation process itself. The models discussed in this article in fact *do not* require that every infant has the same number of observations, but merely that the number of data points for each infant is not related to the underlying phenomenon they intend to model.

This is an important distinction because the currently proposed approaches accommodate the optimization of trial number *while the data collection is running* as long as the reporting is fully transparent. For example, it is entirely possible to run the experiment in the following fashion. First, determine a range

of trials that are administered to the infants. For each individual, draw a random number from that predefined range. Then, data collection proceeds with a randomly drawn number of trials for each participant, and we collect information such the number of trials it took before the infant started to show signs of distress, fatigue, fuzziness, or simply dropped out. As the data accumulate (during a pilot testing but even during the data collection for the main experiment), perform analyses that predict the maximum number of trials before a critical proportion of participants tends to disengage from the task. Over time, we settle on a narrower range that is optimized for the current experimental setup at hand.

To sum up, fixed number of trials designs may be successfully replaced with a dynamic number of trials designs, if one is careful about not introducing the biases discussed throughout this article.

7.4.2 Data quality and efficiency in the laboratory setting

Changes in experimental paradigms bring challenges in determining the fine details of the data collection, which may affect data quality. However, those challenges may instead prove to be opportunities to improve the efficiency of the designs, and in turn — improve data quality in the long run.

For example, an important feature of infant-controlled studies is that recording of the looking time needs to be done "online" (in real time), so that the habituation criteria (or habituation models) may be applied between trials to determine whether additional trial need to be presented. To meet this need, researchers often record looking times manually during the experiment, a method which is error-prone, complicates the data collection process, and requires reliability among coders involving training and precise criteria of what constitutes "looking away" that can be biased if coders are not blind to conditions or stimuli. The fact that there is a dedicated software developed to aid researchers in running habituation studies (Oakes et al., 2019) is a case in point that tackling these issues is not trivial.

In fixed number of trials designs, on the other hand, the coding of looking times can be entirely done offline after the experiment has been concluded. By removing online coding, the data collection can become much smoother and less complex. Further, automated methods for looking times coding may

be used, such as eye-tracking. Eye-tracking is more objective, time efficient, and provides higher spatial and temporal resolution than manual coding (Dalrymple, Manner, Harmelink, Teska, & Elison, 2018; Hunnius, 2007; Oakes, 2017). Using eye-tracking also opens new possibilities to analyze specific gaze patterns within the experiments (e.g., how are infants scanning the presented stimuli), unlocking novel ways in which the data from habituation experiments could be used, which is impossible with manual coding (Aslin, 2007). On the other hand, early evidence shows that eye-tracking (and alternative automatic methods) comes with higher data loss than manual coding (Byers-Heinlein et al., 2021; Chouinard et al., 2019; ManyBabies Consortium, 2020; Venker et al., 2020). A possible solution could be combining both approaches.

Another benefit of relying on modeling approaches is that it presents the opportunity to remove arbitrary criteria for excluding or modifying data for the sake of cleanliness. In typical habituation studies, infants are excluded from the analysis for not meeting the habituation criterion. The previous section already demonstrated that this is not necessary because the data can be analyzed in its entirety by modelling the possibility that certain infants do not habituate. But there are other practical issues that relate to excluding data. For example, many studies define a minimum looking time for a trial to be considered valid. If the looking time falls below this threshold, the trial is simply discarded and another trial is run instead. However, discarding trials based on arbitrary thresholds may mean that we are systematically disregarding an important piece of information about the process of habituation. Extremely short looking times could indicate some sort of dual process, for instance, where the infant first needs to establish an interest in the stimulus, and if that happens, proceeds to gather information with the typical process of habituation. It may also be the case that establishing interest in the stimulus becomes less likely as the infant becomes habituated to the stimulus. Analyzing short trials may reveal that our understanding of the habituation process is incomplete, which would in turn inform model adaptations to accommodate additional phenomena. In short, modeling approaches allow us to consider alternatives to our apriori understanding of how the data unfolds in reality, rather than coercing the data into a shape that we expect based on our current knowledge.

While modeling allows us to accommodate many phenomena, it cannot

be considered as a remedy to any problem encountered during data collection. Data quality is paramount. We already discussed that using eye-tracking may improve measurement precision but there are other design factors that need to be considered. For example, another common practice in habituation studies is defining a maximum trial duration. If an infant did not look away (sufficiently) within the maximum trial duration, the trial is ended nevertheless and the maximum time is recorded as the infant's looking time, which essentially censors the data. While censoring can be modelled, there are issues with that approach. First, if the proportion of trials that ended at the maximum is large, modeling may become difficult as the data misses a lot of information about the exact shape of the habituation curve. Second, it is questionable whether ending the trial by design does not interfere with the habituation process, as the to-be habituated stimulus disappearing in front of the infant may elicit other, unmeasured processes.

Similarly as for determining an optimal number of trials, it is recommendable to adjust the stimuli and criteria during pilot studies or data collection until the maximum trial length and other design factors are optimized. The goal of setting sensible experimental parameters is to reduce artefacts. A good example is a study reported by Bergmann and Cristia (2018), whose data from the habituation phase is displayed in Figure 7.7. The number of trials that ended on the maximum trial duration of 20 sec is relatively small (around 4.4% of all the data points). Defining a looser maximum trial duration rather than no limit at all probably resulted in much smoother data collection and, at the same time, does not severely impede the data analysis, provided that the models are adjusted for censoring rarely occurring data points. However, the authors also encountered other data sets, in which censoring of the data was much more prevalent and hence problematic. Although this is just an anecdotal observation, it demonstrates that reducing artefacts introduced by the study design should not be overlooked, regardless whether one uses novel modeling approaches or sticks to the established experimental paradigms.

7.4.3 Sample size

An important design choice in any study is the sample size necessary to answer specific research questions with a sufficient degree of certainty. Determining

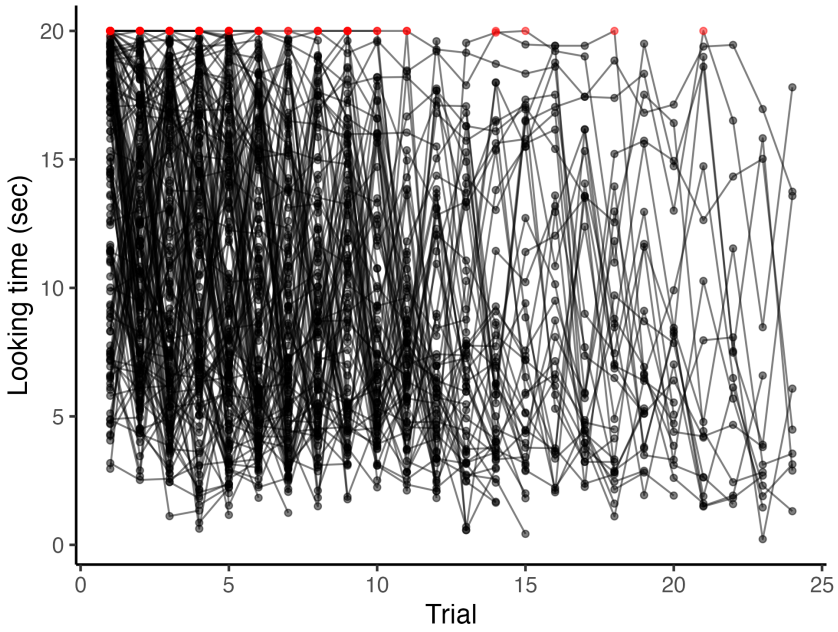


Figure 7.7: Plot of real data from Bergmann and Cristia (2018). Points highlighted red are trials that are close to the maximum trial length.

the sample size might be relatively straightforward when testing for a novelty effect in established paradigms, as previous research can provide enough detail about the expected effect sizes necessary for calculating traditional power analysis (J. Cohen, 1992) or its Bayesian analogue — Bayes factor design analysis (Schönbrodt & Wagenmakers, 2018).

However, with novel paradigms and redefined questions of interest, such calculations might be more difficult. This can lead to uncertainty as to what sample sizes to plan for with new designs. Further, with more complex designs and statistical models, the required sample sizes may become too large. A common approach to this problem would be to first conduct studies of descriptive nature using rules of thumb for planning sample sizes, with the hope that such previous studies would provide enough information to plan future studies with more care, or to inform the statistical models with prior knowledge (Visser et al., 2023). An alternative approach would be to conduct sequential designs: Data collection is continued until a sufficient amount of evidence about the

question of interest accumulates, or the researcher runs out of resources or patience (Schönbrodt et al., 2017; Stefan et al., 2019). Sequential designs hold promising potential to improve the efficiency of experiments, and as such are valuable especially in contexts where achieving large sample sizes is challenging or unlikely (Visser et al., 2023). Evidence from different studies may be integrated to plan new experiments, which would further improve the study design efficiency over the long run (Stefan et al., 2019). Thus, as our predictions become more specific, we will be able to target increasingly more specific questions with increasingly efficient experiments (Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019).

7.5 Conclusion, Discussion, & Recommendations

This article summarised the current state of research in the field of habituation experiments. We have concluded that the performance of methods currently used to detect habituation is uncertain and needs more investigation. However, the very use of infant-controlled designs means that extensive validation of such methods is difficult.

The current proposal argues for re-directing the focus of infant habituation research to first establish the basic, universal characteristics of the habituation process rather than using the verbal explanation of the phenomenon as a tool. Once we can be reasonably confident that we understand, control, and predict the course of habituation, the process will be easier to use as an *experimental paradigm* to study other processes and phenomena.

Based on these challenges, we proposed several remedies. These remedies will themselves bring additional challenges, and as such it seems appropriate to group those based on the difficulty of their implementation.

Minor adaptations

First, we propose to shift focus towards experimental designs that do not depend on the infants' behavior. These may be akin to familiarisation studies with fixed number of trials, or more elaborate designs that aim to optimize the number of trials to administer during the data collection.

Experimental designs that are not conditioned on the infants' behavior will lead to us not being able to control or equate attention levels of participants. Instead, the natural variability in the habituation levels can be accounted for in the analysis, after the data have been collected. If possible, do not set constraints on minimum or maximum number of seconds per trial. If necessary, set the floor or ceiling levels such that the majority of the data points fit in between these constraints; run pilot sessions to check that the floor and ceiling is set appropriately. Further, do not exclude participants that meet some criteria from the data set; instead, involve all participants in the analysis, acknowledging the heterogeneity in the data to its fullest extent.

We also recommend considering alternative measurement methods for recording looking times of the participants, such as eye-tracking instead of manual coding.

It is highly advisable to be as transparent as possible; make every detail of the experiment known to the reader. When presenting the analysis, plot the raw data so that the source of the study's conclusions are traceable.

Major adaptations

Experimental designs and analyses should be informed by (formal) theories, and should be tailored to answer specific research questions.

This adds new challenges in the type of conducting simulation studies to validate novel statistical models to investigate whether it is even feasible to answer specific questions, but also possibly changing the experimental designs in more radical way — for example, running two habituation sessions after one another, instead of running a habituation session followed by a testing session.

These changes should in general lead to better understanding of the habituation phenomenon, and ideally new experimental paradigms emerge that are more efficient and clear as they target specific theoretical questions with methods designed to answer them.

Transparency, collaboration, and mixing expertise

The new habituation designs proposed in this article need empirical validation and further methodological work. As mentioned earlier, one of the challenges

of fully leaning on formal habituation models to design experiments is the large uncertainty about which models would fare the best after model comparison on various data sets. An ideal scenario for reducing this uncertainty would be to fit the models on existing data sets from various age groups, populations, and stimuli. Such activity is likely impossible for individual researchers or even whole research groups.

A way to improve current practices is to introduce more transparency and foster collaboration between research teams. Transparency can be achieved by various means: plotting the raw data in the manuscript and complying more closely to existing reporting standards, sharing the data (including raw data), code, and materials alongside articles that report substantive findings. In accordance with being fully transparent, this article is accompanied by a repository with the code that produced all reported results reported, as well as implementations of various habituation models that were discussed but not reported. The repository can be found at: `osf.io/jr76y/`, where updates about future developments will be posted to guide researchers interested in modelling work for habituation.

Despite our hopes for experimental designs with greater information value, infant research still suffers from limiting factors, such as the relative sparsity of participants. It may be the case that individual laboratories will not be able to feasibly collect enough data to provide meaningful evidence about specific research questions. Collaborations between research teams — such as the MetaBabies, ManyLabs, PsychScience Accelerator, EEGManyPipelines, Many analysis, and variety of other initiatives — will be further gaining prominence and importance for habituation studies as well.

Lastly, it may seem that the current proposal is needlessly complicated and putting an unnecessary burden on the empirical researcher to not only be an expert on the topic of habituation and infant development, but also an expert in experimental designs, statistical modeling, eye-tracking, and other highly complex tasks that may take years of practice to master. While it is true that the issues raised in this article add complexity, tackling these issues can be embraced as an opportunity to explore even more interesting ideas. Further, there is no obligation for individual researchers to master all details of a research program themselves. Instead, the breadth of the topic of habituation calls for even more

active collaborative efforts between experts from different fields.

Open Practices Statement

The code and data used in this article are publicly available at `osf.io/jr76y`.

Appendix

7.A Extended habituation model of Thomas and Gilmore (2004)

Thomas and Gilmore (2004) developed a model which describes the observed looking times y_t for participant at trial t in the following way:

$$\begin{aligned} y_t &\sim \text{Normal}(\mu_t, \sigma) \\ \mu_t &= \alpha + \beta \exp(-\delta(t-1)^2), \end{aligned} \tag{7.1}$$

where δ is a parameter that determines the speed of the habituation process, α controls the residual attention level, and $\alpha + \beta$ together specify the initial attention level. Examples of data that the model generates is presented in Figure 7.1. The model is extended in two ways. First, it is assumed that individual differences exist and so each participant gets their own value of the parameters α, β, δ, and σ. These parameters are modelled using a hierarchical structure, therefore estimating the group means and standard deviations for all four parameters. Second, a "multi-population model" as suggested by L. B. Cohen and Menten (1981) is specified; that is, the model assumes that the data come from a mixture of two populations. One population is described by the habituation process defined above, the other is a group of "non-habituators", that is infants who do not, and will not show a decrease of attention. The second group is modeled by replacing the expression for μ_t in Equation 7.1 with an intercept parameter that does not change with t. Further, the original model by Thomas and Gilmore (2004) uses a Normal distribution for the looking times; in the present application, a truncated Normal distribution is used to honor the fact that looking times cannot be smaller than zero. The full model specification

used in this example is as follows:

$$
\begin{aligned}
y_{it} &\sim \text{Normal}(\mu_{it}, \sigma)_{\text{T}(0,\infty)} \\
\mu_{it} &= \begin{cases} \gamma_i & \text{if } z_i = 0 \\ \alpha_i + \beta_i \exp\left[-\delta_i (t-1)^2\right] & \text{if } z_i = 1 \end{cases} \\
z &\sim \text{Bernoulli}(\pi) \\
\log(\gamma) &\sim \text{Normal}(\mu_\gamma, \sigma_\gamma) \\
\log(\alpha) &\sim \text{Normal}(\mu_\alpha, \sigma_\alpha) \\
\log(\beta) &\sim \text{Normal}(\mu_\beta, \sigma_\beta) \\
\log(\delta) &\sim \text{Normal}(\mu_\delta, \sigma_\delta) \\
\log(\sigma) &\sim \text{Normal}(\mu_\sigma, \sigma_\sigma),
\end{aligned}
\tag{7.2}
$$

a looking time y_{it} of participant i and trial t is distributed according to the truncated normal distribution. The mean parameter of the distribution is either an intercept γ_i that does not depend on the trial, or a habituation curve that is conditional on whether or not the infant belongs to the group of "non-habituators" or "habituators" (indicated by $z_i = 1$ which is driven by the proportion of infants in the population that are "habituators", π). All parameters are modelled at the individual level in log space to ensure the positivity of the parameters, assuming a Normally distribution with population means and standard deviations that are estimated from the data. This allows to model individual differences (e.g., allowing the model to accommodate slow and fast habituators) with hierarchical pooling of information across the participants to improve the individual parameter estimates.

Part III

Learning under Uncertainty

I don't want preconceptions. I want to learn as much as possible.

–Pep Guardiola

Chapter 8

Analytic Posterior Distribution and Bayes Factor for Pearson Partial Correlations

This chapter is preprinted as Kucharský, Š., Wagenmakers, E.-J., van den Bergh, D., and Ly, A. (2023). Analytic posterior distribution and Bayes factor for Pearson partial correlations. *PsyArXiv*. doi: 10.31234/osf.io/6muwy

Abstract

This article outlines a novel Bayesian approach to the testing and estimation of Pearson partial correlations. By generalizing a Bayesian inference procedure for Pearson's correlation coefficient we obtain analytic expressions for the Bayes factor and for the (marginal) posterior distribution of a partial correlation coefficient. Full Bayesian inference can be achieved using only the sample size, the number of controlling variables and the relevant summary statistics, that is, the sample partial correlation. The present approach is illustrated with two empirical examples.

8.1 Introduction

The Pearson partial correlation coefficient quantifies the linear relationship between two continuous variables while taking into account the effects of other (confounding) variables. Under certain distributional assumptions (e.g., multivariate normality), the partial correlation coincides with a conditional correlation and can therefore be used to assess conditional (in)dependence between a set of variables (Baba, Shibata, & Sibuya, 2004; Baba & Sibuya, 2005; Lawrance, 1976). This fact is central in Gaussian Graphical Models (Lauritzen, 1996), where partial correlations are used to map out unique relationships between a number of variables (i.e., partial correlation networks; Costantini et al., 2015). For these reasons, inference for partial correlation is included both in popular introductory statistical textbooks (Agresti & Finlay, 2009; Field, 2017; Lomax & Hahs-Vaughn, 2012; Moore, McCabe, & Craig, 2012) and statistical software packages (e.g., IBM Corp., 2017; JASP Team, 2021; Kim, 2015) as a basic statistical tool.

In practical applications, researchers often wish to test whether a population partial correlation is zero. The dominant approach is to use frequentist null hypothesis significance testing, which is already well developed for the case of a partial correlation (Weatherburn, 1961, pp. 242–263). However, frequentist hypothesis testing comes with several limitations (e.g., Amrhein, Greenland, & McShane, 2019; Nuzzo, 2014; Wagenmakers, 2007; Wasserstein & Lazar, 2016), one of them being the inability to quantify evidence in favor of the null hypothesis. In other words, the frequentist test does not discriminate between 'evidence of absence' and 'absence of evidence' (e.g., Keysers et al., 2020). This

is particularly problematic for partial correlations, because researchers often wish to claim evidence for the null hypothesis of conditional independence. In general, it is desirable that a method of testing can quantify evidence in favor of either conditional independence or conditional dependence (Epskamp, 2017, pp. 240–241).

This frequentist limitation can be overcome within the framework of Bayesian statistics, namely with *Bayes factors* (Jeffreys, 1961; Kass & Raftery, 1995). Bayes factor tests for partial correlations have already been proposed by Wetzels and Wagenmakers (2012), by M. Wang, Chen, Lu, and Dong (2019), and by Williams and Mulder (2020). First, Wetzels and Wagenmakers (2012) proposed to test the coefficient indirectly by testing an increase of explained variance in a linear regression model. However, this procedure yields results that are sensitive to the direction of the effect (i.e., which of the two variables of interest is used as the predictor), which is undesirable as partial correlation is an undirected coefficient. Second, M. Wang et al. (2019) also use a setup from linear regression with the popular Zellner's g-prior (Zellner, 1986), but proposed to test the regression coefficient instead, which effectively leads to converting the coefficient to a t-statistic. One downside of this approach is that the resulting Bayes factor inherits an issue of the Zellner's g-prior in that it is not information consistent (Liang, Paulo, Molina, Clyde, & Berger, 2008). A drawback common to both approaches is that they do not test the partial correlation directly, but represent it with a proxy statistic, be it a change in R^2 or a coefficient in a linear regression. This makes it difficult to reason about the implied prior distribution on the partial correlation coefficient, frustrating the use of informed priors. Neither approach provides the posterior distribution for the population partial correlation coefficient, which is of course the central target for Bayesian parameter estimation.

A third Bayes factor test for partial correlations was introduced by Williams and Mulder (2020) and Williams (2021), who proposed to fit a full multivariate normal model to the data with either Wishart or Matrix-F prior distributions (J. Mulder & Pericchi, 2018) on the covariance matrix. Based on the full sample covariance matrix it is then possible to obtain Bayes factors for individual partial correlations through the Savage-Dickey density ratio (Dickey & Lientz, 1970). To obtain the Bayes factor, Williams and Mulder (2020) proposed to use

either analytic approximations to the posterior distributions, or apply MCMC sampling. This method is available to the practical researcher with the R package `BGGM` (Williams & Mulder, 2019).

Here we introduce a new Bayesian approach to test and estimate partial correlation coefficients. Our work generalizes the Bayesian development for Pearson's correlation coefficient (Ly et al., 2018, 2016b), and inherits many of its desirable properties. For instance, the complete Bayesian inference can be conducted with only the relevant summary statistics and is computationally cheap. Our main results are twofold and address both Bayes factor testing and Bayesian parameter estimation. First, we provide an expression for the Bayes factor of a nullity of a partial correlation coefficient. We elaborate how the proposed Bayes factor fulfils certain desiderata that allow intuitive inferences, making it an attractive option for a default Bayesian testing method (Bayarri, Berger, Forte, & García-Donato, 2012; Jeffreys, 1961; Ly et al., 2016b). Second, we derive an analytic marginal posterior distribution for the partial correlation coefficient. This posterior distribution facilitates a Bayesian estimation effort. In general, the inference is carried out on the partial correlation coefficient itself which also encourages the use of informed prior distributions (e.g., Gronau, Ly, & Wagenmakers, 2020).

The paper continues as follows: Section 8.2 presents the proposed Bayes factor and (marginal) posterior for the partial correlation coefficient explicitly. Section 8.3 highlights two properties of the proposed Bayes factor. Section 8.4 illustrates the method with two examples, and the paper is concluded in Section 8.5. All derivations are provided in the appendix.

8.2 Bayesian inference for partial correlation

Interest centers on the population partial correlation $\rho_{xy.z}$ that measures the degree of association between X and Y after the effects of k number of controlling variables Z on X and Y are removed. We focus on two somewhat different aspects of Bayesian inference: testing (i.e., "is there evidence for the presence of an effect?") and estimation (i.e., "assuming the effect exists, how strong is it?"). For testing, the relevant question is whether or not the partial correlation equals zero – more specifically, whether and to what extent the data provide

support for (or against) the hypothesis that the partial correlation coefficient equals zero. To address this question we may report the Bayes factor that compares the predictive performance of the null model $\mathcal{M}_0$ that operationalizes the null hypothesis $\rho_{xy.z} = 0$ to the predictive performance of the alternative model $\mathcal{M}_1$ that operationalizes the alternative hypothesis $\rho_{xy.z} \in (-1, 1)$. For estimation, the alternative model is assumed to hold true, wherein an unknown (population) partial correlation is free to vary between -1 and 1. The goal is then to infer this unknown parameter based on the available data. To address this question we report the posterior distribution of the partial correlation.

In Appendix 8.A we provide the detailed derivations of the Bayes factor and posterior distribution presented in this section. It is assumed that the variables X, Y and the k number of controlling variables Z are jointly multivariate-normally distributed. The resulting $p = k + 2$ multivariate normal distribution has $p(p + 3)/2$ number of parameters, but the focus of inference is only on one of them, namely, $\rho_{xy.z}$. For instance, for $k = 0, 1, 2, 3, 4$ controlling variables there are $5, 9, 14, 20$ and 27 parameters respectively, thus, $4, 8, 13, 19$ and 26 so-called nuisance parameters. Analogously, the data from a p multivariate normal distribution can be (sufficiently) summarized by $p(p+3)/2$ values. Our derivations show that the use of specific priors result in analytic Bayes factor and marginal posterior for $\rho_{xy.z}$ that solely depend on the sample size, the number of controlling variables, and the corresponding sample partial correlation $r_{xy.z}$ from the total of $p(p + 3)/2$ sufficient statistics. This is achieved in Appendix 8.A by (1) isolating that part of the likelihood that only involves $\rho_{xy,z}$ and its sampled counterpart $r_{xy.z}$, and by (2) choosing appropriate priors on the nuisance parameters. For ease of exposition the nuisance parameters are collected in the symbol θ_0, an explicit account is given in Appendix 8.A.

The alternative model includes one additional parameter, namely, the population partial correlation. Hence, the parameters of the alternative model can be denoted as $\theta = (\theta_0, \rho_{xy.z})$, and the prior on $\rho_{xy.z}$ is set independently from the nuisance parameters as $\pi(\theta) = \pi(\theta_0) \times \pi(\theta_{xy.z})$.

In accordance with the previous work on the Pearson's correlation (Ly et al., 2018), the default prior used for the partial correlation here is a symmetric beta distribution stretched to the interval -1 to 1. This stretched beta prior has a single hyperparameter α that governs the concentration of the distribution

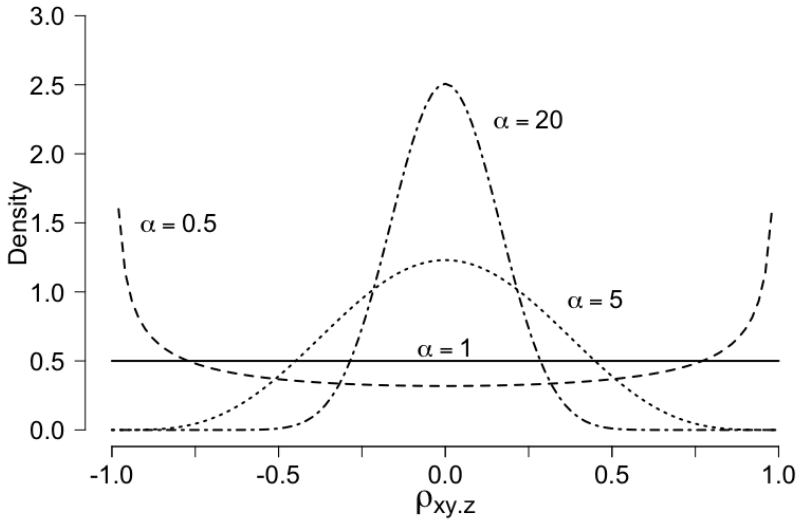


Figure 8.1: Examples of the symmetric stretched beta prior distribution for the partial correlation $\rho_{xy.z}$ defined by four different values for the hyperparameter α.

around zero: if $\alpha = 1$, the prior distribution is uniform; when α is large, the distribution is peaked and centered on the value $\rho_{xy.z} = 0$. Figure 8.1 shows four examples of the stretched beta prior distribution for $\rho_{xy.z}$ using different values for the α hyperparameter.

With the appropriate priors, as detailed in Equation 8.15 in Appendix 8.A, on the nuisance parameter θ_0 in both models, and the stretched beta prior on $\rho_{xy.z}$ in the alternative model, the Bayes factor for the alternative model over the null model is given by

$$\text{BF}_{10} = \frac{\mathcal{B}\left(\frac{1}{2}, \alpha + \frac{n-k-\gamma-\delta-1}{2}\right)}{\mathcal{B}\left(\frac{1}{2}, \alpha\right)} \times {}_2F_1\left(\frac{n-k-\gamma-1}{2}, \frac{n-k-\delta-1}{2}; \alpha + \frac{n-k-\gamma-\delta}{2}; r^2_{xy.z}\right). \tag{8.1}$$

The marginal posterior distribution of $\rho_{xy.z}$ under the alternative model is

given by:

$$\pi(\rho_{xy.z} \mid n, k, r_{xy.z}) = \frac{(1-\rho_{xy.z}^2)^{(2\alpha+n-k-\gamma-\delta-3)/2}}{\mathcal{B}\left(\frac{1}{2}, \alpha + \frac{n-k-\gamma-\delta-1}{2}\right)\ {}_2F_1\left(\frac{n-k-\gamma-1}{2}, \frac{n-k-\delta-1}{2}; \alpha + \frac{n-k-\gamma-\delta}{2}; r_{xy.z}^2\right)} \times \left[{}_2F_1\left(\frac{n-k-\gamma-1}{2}, \frac{n-k-\delta-1}{2}; \frac{1}{2}; r_{xy.z}^2\rho_{xy.z}^2\right) + 2r_{xy.z}\rho_{xy.z}W_{\gamma,\delta}(n-k)\ {}_2F_1\left(\frac{n-k-\gamma}{2}, \frac{n-k-\delta}{2}; \frac{3}{2}; r_{xy.z}^2\rho_{xy.z}^2\right)\right]. \tag{8.2}$$

Note that the expressions only depend on the data via the relevant summary statistic $r_{xy.z}$, the sample size n, and the number of conditioning variables k as promised. The hyperparameters δ and γ can be tuned by the statistician, but are typically set to zero (see Section 8.3.1). $\mathcal{B}(a, b)$ is the beta function, ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function, and $W_{\gamma,\delta}(\tilde{n})$ is defined in Equation 8.20 in Appendix 8.A.

By substituting $\tilde{n} = n-k$, the number of samples exceeding the number of controlling variables, we see that the results generalize the Pearson's correlation of Ly et al. (2018) in the sense that the Pearson's correlation is a special case of the partial correlation when the number of controlling variables is zero ($k = 0$).

8.3 Properties of the Bayes factor

Bayesian model comparison and selection methods, such as the Bayes factor, are sensitive to the choice of priors, and this sensitivity does not vanish as the sample size increases (Bayarri et al., 2012; Kass & Raftery, 1995).

Because of this sensitivity, a considerable effort has been exerted to develop "objective", or "default" methods that would provide standard inferences for typical testing scenarios (Bayarri et al., 2012; Berger, 2006). A pioneer in this field was Jeffreys (1961) who not only proposed various default Bayesian tests for common statistical problems, but also provided a set of *desiderata* for newly developed tests, such that they provide an intuitive framework for inference (Bayarri et al., 2012; Ly, Verhagen, & Wagenmakers, 2016a; Ly et al., 2016b).

Here we show that the Bayes factor presented in this article meets Jeffreys's

desiderata for Bayes factors: predictive matching and information consistency.

8.3.1 Predictive matching

A Bayes factor is predictively matched if it equals 1 when the Bayes factor is presented with completely uninformative data, that is, when the data bear no evidence for either hypothesis.

This occurs when the data are of insufficient size, thus, less than a minimal sample size needed to distinguish between the null and alternative models. For these data sets the the Bayes factor should remain indifferent (i.e., $\text{BF} = 1$).

Note that we cannot infer the partial correlation when $n \leq k + 1$ as the sample partial correlation is then undefined. When $n = k + 2$, we automatically get $r_{xy.z} = \pm 1$, regardless of the value of the population coefficient. Hence, the minimum sample size is $n_{\min} = k + 3$. For data sets of size $n < k + 1$, we define the Bayes factor to be one. For $n = k + 2$, we enter the values $r_{xy.z} = 1$, into Equation 8.1 and obtain with $\tilde{\alpha} = \alpha + \frac{n-k-\gamma-\delta-1}{2}$

$$\begin{aligned}
\text{BF}_{10} &= \frac{\mathcal{B}\left(\frac{1}{2}, \tilde{\alpha}\right)}{\mathcal{B}\left(\frac{1}{2}, \alpha\right)} \times {}_2F_1\left(\frac{1-\gamma}{2}, \frac{1-\delta}{2}; \tilde{\alpha} + \frac{1}{2}; 1\right) \\
&= \frac{\mathcal{B}\left(\frac{1}{2}, \tilde{\alpha}\right)}{\mathcal{B}\left(\frac{1}{2}, \alpha\right)} \times \frac{\Gamma(\tilde{\alpha} + \frac{1}{2})\Gamma(\tilde{\alpha} + \frac{1}{2} - \frac{1-\gamma}{2} - \frac{1-\delta}{2})}{\Gamma(\tilde{\alpha} + \frac{1}{2} - \frac{1-\gamma}{2})\Gamma(\tilde{\alpha} + \frac{1}{2} - \frac{1-\delta}{2})} \\
&= \frac{\Gamma(\alpha + \frac{1-\gamma-\delta}{2})\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha + \frac{1-\gamma}{2})\Gamma(\alpha + \frac{1-\delta}{2})}.
\end{aligned} \tag{8.3}$$

Thus, for the Bayes factor to be predictively matched, we require $\gamma = \delta = 0$.

8.3.2 Information consistency

A Bayes factor is information consistent if it diverges to infinity, thus, falsifies the null, when the Bayes factor is presented with overwhelmingly informative data.

Overwhelmingly informative data sets are of sufficient size, thus, $n \geq n_{\min} = k + 3$ with $r_{xy.z} = 1$ or $r_{xy.z} = -1$, because if the null were true, then such an event occurs with chance zero. Equation 8.21 and the surrounding discussion in the appendix imply that the Bayes factor diverges whenever

$2\alpha + 2 + k \leq n$. For a Bayes factor to already diverge at the minimum sample size, we plugin $n = k + 3$ and conclude that information consistency requires $\alpha \leq 1/2$. On the other hand, when $r_{xy.z} = \pm 1$ and $\alpha > 0$ is given, we conclude that the Bayes factor diverges whenever $n - k \geq 2\alpha + 2$. For instance, for $\alpha = 1$ the Bayes factor diverges only when $n \geq k + 4$, thus, missing the information consistency desideratum by one observation. In other words, for $\alpha = 1$, we are one observation more reluctant to falsify the null. Analogously, for $\alpha = 20$ we require an additional 39 perfectly (partially) correlated observations before we are willing to falsify the null.

8.4 Two examples

Our first example features a data set from experimental psychology. Lleras, Porporino, Burack, and Enns (2011) had $n = 40$ participants complete a visual search task. The data showed a relatively high correlation ($r = .51$) between 'successful search time' and 'rapid resumption'. In a Bayesian reanalysis, we assigned the correlation coefficient a stretched beta prior with $\alpha = 0.5$ (according to the information consistency requirement explained in Section 8.3.2). Figure 8.2 shows the resulting inference. In terms of testing, the Bayes factor $\text{BF}_{10} \approx 33.85$ indicates 'very strong' evidence in favor of the alternative model over the null model (cf. Jeffreys, 1961, Appendix B; Lee & Wagenmakers, 2013, Table 7.1). In terms of estimation, the posterior distribution for Pearson's ρ (under the alternative model) is relatively symmetric around 0.49, with a central 95% credible that ranges from 0.23 to 0.70.

However, after controlling for participant's age the sample partial correlation coefficient decreases to almost zero ($r_{xy.z} = .01$). In a Bayesian reanalysis, we again assigned the partial correlation coefficient a stretched beta distribution with $\alpha = 0.5$. Figure 8.3 shows the resulting inference. In terms of testing, the Bayes factor $\text{BF}_{01} \approx 7.76$ indicates 'moderate' evidence in favor of the null model over the alternative model (cf. Jeffreys, 1961, Appendix B; Lee & Wagenmakers, 2013, Table 7.1). For comparison, Wetzels and Wagenmakers (2012) reported $\text{BF}_{01} = 7.70$ and M. Wang et al. (2019) reported $\text{BF}_{01} = 2.02$. The method by Williams and Mulder (2020) requires the full sample covariance matrix of the three variables ('successful search time', 'rapid resumption', and

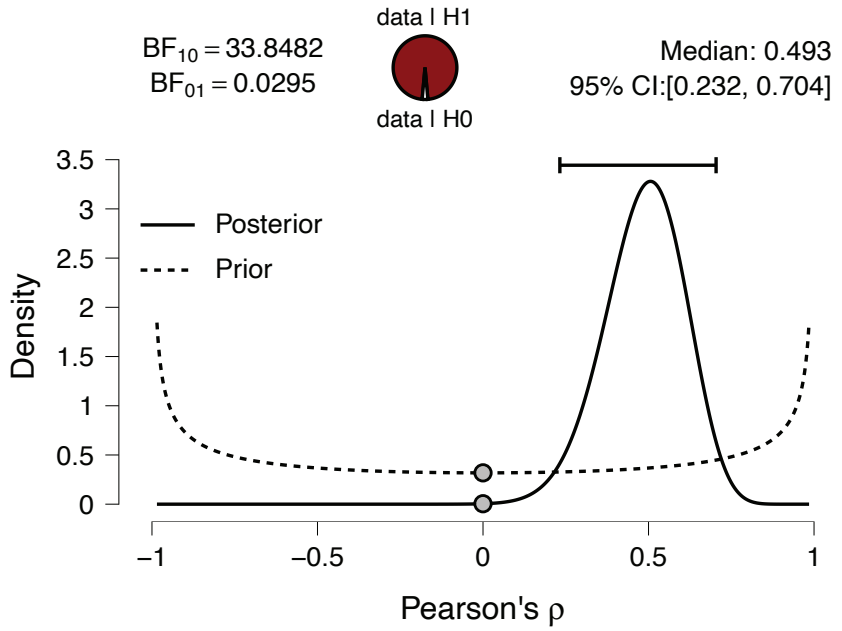


Figure 8.2: Bayesian inference on the Pearson's correlation coefficient (Ly et al., 2018) for the data reported in Lleras et al. (2011) (i.e., $r = 51$, $n = 40$). The Bayes factor indicates very strong support for the alternative model over the null model. The grey dots at $\rho = 0$ visualize the Savage-Dickey density ratio (Dickey & Lientz, 1970). Figure concept from JASP (e.g., van Doorn, van den Bergh, Böhm, et al., 2021).

'age') to obtain the Bayes factor for the partial correlation. To our knowledge, this information is unfortunately not openly available.

In terms of estimation, the posterior distribution for $\rho_{xy.z}$ (under the alternative model) is relatively symmetric around 0.01, with a central 95% credible that ranges from -0.30 to 0.32.

Our second example concerns the relation between COVID-19 infections and air pollution across 55 Italian cities in the period from March 17$^{\text{th}}$ to April 7$^{\text{th}}$, 2020. Specifically, Coccia (2021) reported a partial correlation between the logarithm of COVID-19 infections and the logarithm of the number of days with increased air pollution, controlling for population density. The summary statistics that allow a complete Bayesian reanalysis are the sample partial correlation $r_{xy.z} = .479$, sample size $n = 55$, and the number of controlling

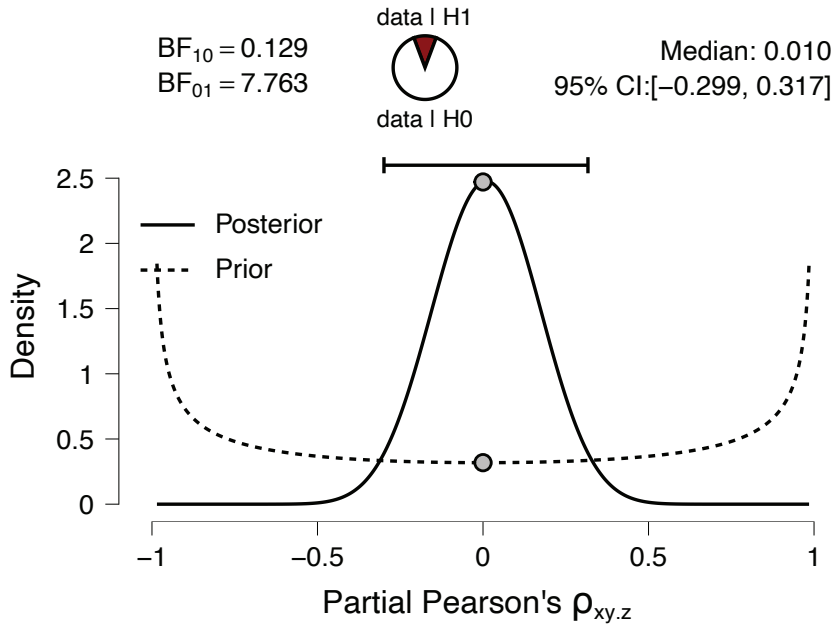


Figure 8.3: Bayesian inference on the partial correlation coefficient for the data reported in Lleras et al. (2011) (i.e., $r_{xy.z} = .01$, $n = 40$). After controlling for participant age, the Bayes factor indicates moderate support for the null model over the alternative model. The grey dots at $\rho = 0$ visualize the Savage-Dickey density ratio (Dickey & Lientz, 1970). Figure concept from JASP (e.g., van Doorn, van den Bergh, Böhm, et al., 2021).

variables (population density) $k = 1$. We again assigned the partial correlation coefficient a stretched beta prior with $\alpha = 0.5$. Figure 8.4 shows the resulting inference. In terms of testing, the Bayes factor $\text{BF}_{10} \approx 84.70$ indicates 'extreme' evidence in favor of the alternative model over the null model (cf. Jeffreys, 1961, Appendix B; Lee & Wagenmakers, 2013, Table 7.1). In terms of estimation, the posterior distribution for $\rho_{xy.z}$ (under the alternative model) is relatively symmetric around 0.47, with a central 95% credible that ranges from 0.24 to 0.66. The data therefore appear to provide compelling statistical support for an association between the intensity of air pollution and the susceptibility to COVID-19 infections.

To explore the robustness of this conclusion we may reanalyze the data from Coccia (2021) using an informed prior that assigns more mass to values of

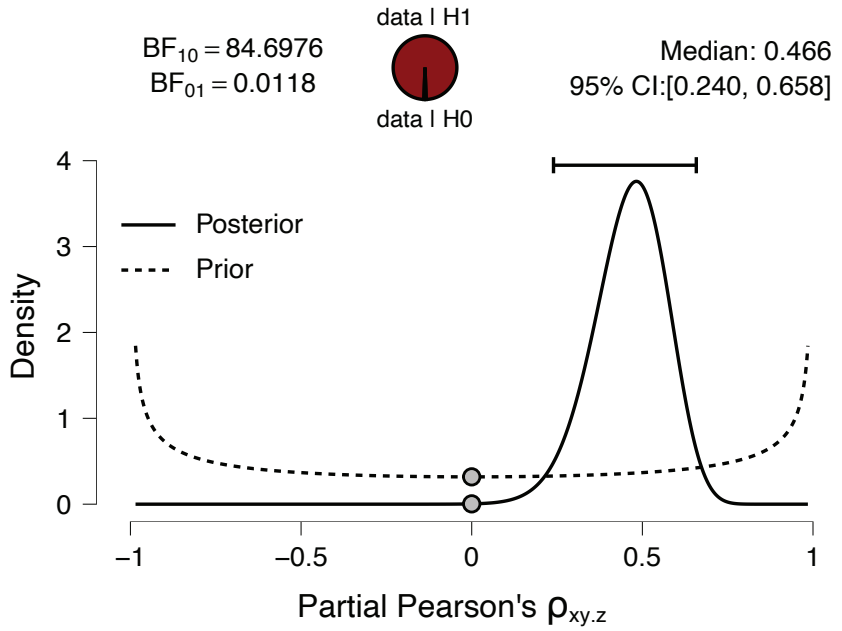


Figure 8.4: Bayesian inference on the partial correlation coefficient for the data reported in Coccia (2021). The Bayes factor indicates extreme support for the alternative model over the null model. The grey dots at $\rho = 0$ visualize the Savage-Dickey density ratio (Dickey & Lientz, 1970). Figure concept from JASP (e.g., van Doorn, van den Bergh, Böhm, et al., 2021).

$\rho_{xy.z}$ near zero. Specifically, we assign $\rho_{xy.z}$ a stretched beta distribution with the α hyperparameter set to 20. The result is shown in Figure 8.5. Compared to the result with a uniform prior, the posterior distribution is now closer to zero, and Bayes factor is noticeably smaller: $\text{BF}_{10} \approx 29.21$. However, the data still provide strong evidence in favor of the alternative model over the null model.

Instead of picking a single value of the hyperparameter α to explore the robustness of the result, we may report the Bayes factor for a range of possible prior specifications (see van Doorn, van den Bergh, Böhm, et al., 2021). Robustness plot in Figure 8.6 shows Bayes factors for the data from Coccia (2021) as a function of $\kappa = \frac{1}{\alpha}$, which can be interpreted as a prior width. For values of κ close to zero, the predictions from the alternative model become indistinguishable from the null model and the Bayes factor decreases to 1 accordingly. However, for a wide range of priors (starting from $\kappa \approx 0.25$), the Bayes fac-

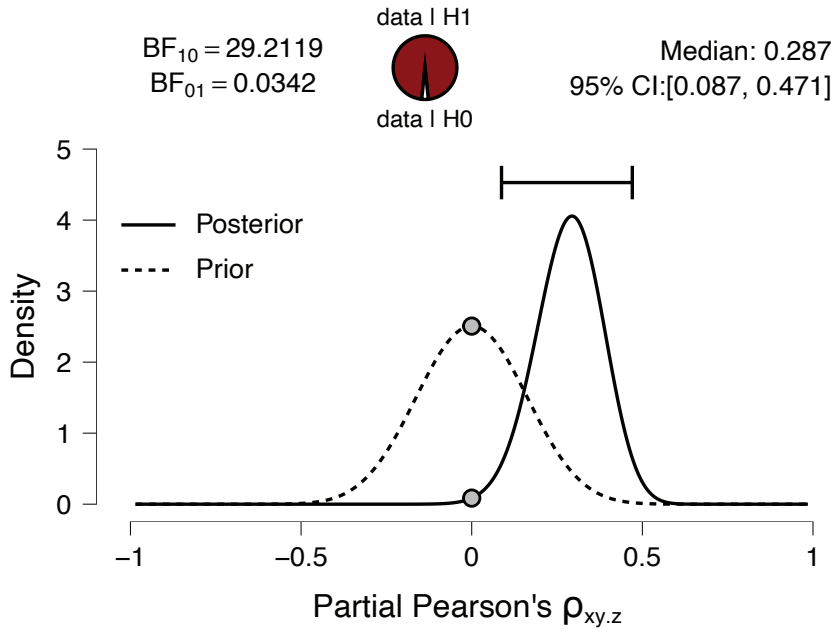


Figure 8.5: Bayesian inference on the partial correlation coefficient for the data reported in Coccia (2021), using the informed prior with $\alpha = 20$. Compared to the results based on the uniform prior, the posterior distribution has shifted toward zero, and the Bayes factor is less compelling; however, the support for the alternative model over the null model is still strong. The grey dots at $\rho = 0$ visualize the Savage-Dickey density ratio (Dickey & Lientz, 1970). Figure concept from JASP (e.g., van Doorn, van den Bergh, Böhm, et al., 2021).

tors is relatively stable around 100, indicating that the evidence against the null hypothesis is robust.

8.5 Concluding remarks

We presented a new approach for Bayesian testing and estimation of a partial correlation coefficient. The framework generalizes previous work on Pearson's correlation coefficient (Ly et al., 2018, 2016b) and inherits several desirable properties; for instance, when $\delta = \gamma = 0$ the Bayes factor is predictively matched, and when $\alpha \leq 1/2$ also information consistent. The inference is carried out on the partial correlation coefficient itself, as opposed to two previ-

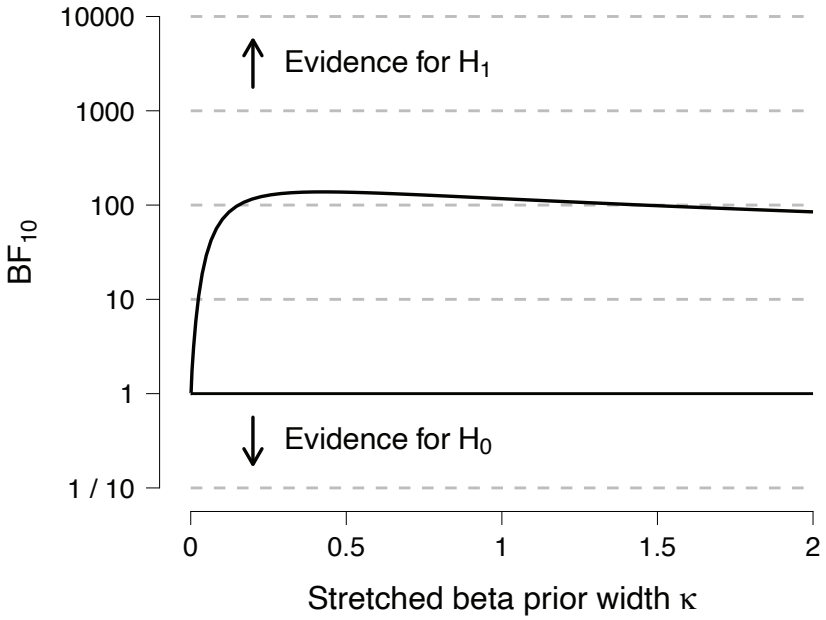


Figure 8.6: Bayes factor robustness plot (see van Doorn, van den Bergh, Böhm, et al., 2021) for the partial correlation reported by Coccia (2021): Bayes factor is plotted as a function of the prior width $\kappa = \frac{1}{\alpha}$. The evidence in favour of the alternative model is strong for a wide range of prior specifications.

ous proposals of Bayesian tests of partial correlations. As a result, we obtained an analytic expression for the Bayes factor and for the posterior distribution of the partial correlation coefficient. Furthermore, the full inference can be carried out using only the sample size, the number of controlling variables, and the relevant summary statistics, that is, only the sample partial correlation $r_{xy.z}$ corresponding to the target of inference $\rho_{xy.z}$. For these reasons, the methodology developed here is arguably an attractive option as a default Bayesian inference procedure for partial correlations.

It is important to note that the Bayesian and frequentist inference for a partial correlation share many key assumptions: the vector of observations must be independent, and the variables must be (approximately) multivariate-normally distributed. Although small deviations from these assumptions may not completely invalidate the results, the coefficient may be especially sensitive to distortion in case of nonlinear relationships between variables, in the presence of

outliers, or with significant measurement error (K. Liu, 1988; Osborne & Waters, 2002; Quade, 2017; Vargha, Bergman, & Delaney, 2013). Whenever researchers have access to the raw data, we recommend that they carefully check these assumptions; when researchers report original work we encourage them to publicly archive the (properly anonymized) data, or –at a minimum– plot the data so that readers may confirm that the analysis is appropriate and informative (e.g., Wagenmakers et al., 2021).

Appendix

8.A Derivation of the main results

8.A.1 Reparametrising the multivariate normal model

Let $X \in \mathbb{R}$, $Y \in \mathbb{R}$ be two random variables and $Z \in \mathbb{R}^k$ a random vector ($k \in \mathbb{Z}_{\geq 0}$). We consider the random vector $W \in \mathbb{R}^p$ collecting X, Y, Z, where $p = k + 2$, to follow the multivariate normal distribution. Further, let n be the number of observations of the random vector W, and summarize the observed data as follows: $d = (n, \bar{w}, S)$, where $\bar{w}$ is the vector of sample means, $\bar{w}^{(j)} = n^{-1} \sum_i^n w_i^{(j)}$ for $j \in 1, ..., p$, and $S = n^{-1}(w - \bar{w})'(w - \bar{w})$ is an average sum of squares and cross-product matrix.

The likelihood of the model is

$$\begin{aligned} f(d \mid \mu, \Sigma) =& (2\pi)^{-\frac{pn}{2}} |\Sigma|^{-\frac{n}{2}} \times \\ & \exp\left(-\frac{n}{2}[\text{tr}(\Sigma^{-1}S) + (\bar{w} - \mu)'\Sigma^{-1}(\bar{w} - \mu)]\right), \end{aligned} \tag{8.4}$$

where tr(.) denotes the trace operator and $|\ .\ |$ denotes the determinant, μ is the vector of population means, and Σ is the population variance-covariance matrix.

As there are p population means, p variances, and $\binom{p}{2}$ correlations, the model contains $p(p+3)/2$ parameters. We will remove the nuisance parameters by carefully setting suitable prior distributions and marginalizing them out of the likelihood. Specifically, we choose priors such that the nuisance parameters do not influence the inference about $\rho_{xy.z}$. This can be accomplished by allowing the structure of the parameters (and their priors) to reflect the structure of the likelihood.

Our approach is to reparametrize the multivariate normal model of the random vector W in order to isolate the partial correlation $\rho_{xy.z}$ as a separate parameter. It is then possible to fix the value of $\rho_{xy.z}$ to zero (under the null model $\mathcal{M}_0$), or assign it a prior probability distribution (under the alternative model $\mathcal{M}_1$). The desired reparametrization is achieved using the Schur decomposition of the variance-covariance matrix, which replaces Σ with three sets of parameters: $\Sigma_{11.2}$, Σ_{22}, and B. First, $\Sigma_{11.2}$ is the conditional variance-covariance matrix of the variables X and Y after controlling for variables Z. $\Sigma_{11.2}$ can be decomposed in terms of three parameters: the focal parameter $\rho_{xy.z}$ that represents the partial correlation, and the two standard deviations of the residuals, $\sigma_{x.z}$ and $\sigma_{y.z}$, after regressing X and Y on Z, respectively. Second, Σ_{22} is the variance-covariance matrix of the controlling variables Z; third, B is a matrix of regression coefficients that represent the relation of X and Y to Z.

We achieve the reparametrization by first decomposing Σ into a block matrix,

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}, \tag{8.5}$$

where Σ_{11} is a 2×2 variance-covariance matrix of the variables X and Y, Σ_{22} is a $k \times k$ variance-covariance matrix of the controlling variables Z, and $\Sigma_{12} = \Sigma_{21}'$ is a $2 \times k$ matrix of cross-covariances of X with Z, and Y with Z.

To obtain the partial correlation instead of the classic correlation, we use the Schur complement of Σ_{11} in Σ defined as $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, which for the multivariate normal model can be interpreted as the conditional variance-covariance matrix of X and Y given Z. This can be done analogously with the sample counterpart S provided that S_{22} is positive definite.

We now need to rewrite the trace in the exponent of the likelihood in Equation 8.10 in terms of this decomposition. The Schur decomposition allows us to invert Σ easily as:

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11.2}^{-1} & -\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix}.$$

To let the sample partial covariance appear in the likelihood, we write S as a

block matrix:

$$S = \begin{bmatrix} S_{11.2} + S_{12}S_{22}^{-1}S_{21} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

This leads to

$$\begin{aligned} \text{tr}(\Sigma^{-1}S) = &\text{tr}(\Sigma_{11.2}^{-1}S_{11.2}) + \text{tr}(\Sigma_{22}S_{22}) \\ &+ \text{tr}(\Sigma_{11.2}^{-1}S_{12}S_{22}^{-1}S_{21}) - \text{tr}(\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}S_{21}) \\ &- \text{tr}(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}S_{12}) + \text{tr}(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}S_{22}). \end{aligned} \tag{8.6}$$

The first trace involves the block of interest (i.e., the partial correlation scaled by the partial variance) and the second trace involves the nuisance covariance of Z. The last four terms involve the nuisance blocks with respect to the cross-correlation Σ_{12}, and can be written (by factoring out S_{22} and $\Sigma_{11.2}$) as a quadratic term in the form

$$\text{tr}\left[S_{22}(S_{12}S_{22}^{-1} - \Sigma_{12}\Sigma_{22}^{-1})'\Sigma_{11.2}^{-1}(S_{12}S_{22}^{-1} - \Sigma_{12}\Sigma_{22}^{-1})\right],$$

where we used the cyclic property and linearity of traces, and multiplied by the identity $I = S_{22}S_{22}^{-1}$ where needed in order to collect the four traces inside of a single quadratic term, as detailed in the Appendix 8.B. Substituting the four traces with the quadratic terms leads to

$$\begin{aligned} \text{tr}(S\Sigma^{-1}) = &\text{tr}(S_{11.2}\Sigma_{11.2}^{-1}) + \text{tr}(S_{22}\Sigma_{22}^{-1}) \\ &+\text{tr}\left[S_{22}(S_{12}S_{22}^{-1} - \Sigma_{12}\Sigma_{22}^{-1})'\Sigma_{11.2}^{-1}(S_{12}S_{22}^{-1} - \Sigma_{12}\Sigma_{22}^{-1})\right]. \end{aligned} \tag{8.7}$$

In the above equation, the nuisance block including the terms $\Sigma_{12}\Sigma_{22}^{-1}$ can be interpreted as the linear regression coefficients of X and Y on Z. We write the matrix as $\text{B} = \Sigma_{12}\Sigma_{22}^{-1}$ and analogously $\hat{\text{B}} = S_{12}S_{22}^{-1}$. Using the fact that

$|\Sigma| = |\Sigma_{11.2}| \times |\Sigma_{22}|$, we arrive at

$$
\begin{aligned}
f(d \mid \mu, \Sigma) =& (2\pi)^{-\frac{pn}{2}} |\Sigma_{11.2}|^{-\frac{n}{2}} |\Sigma_{22}|^{-\frac{n}{2}} \times \\
& \exp\Big(-\frac{n}{2} (\bar{w} - \mu)' \Sigma^{-1} (\bar{w} - \mu) \Big) \times \\
& \exp\Big(-\frac{n}{2} \text{tr}(S_{11.2} \Sigma_{11.2}^{-1}) \Big) \times \\
& \exp\Big(-\frac{n}{2} \text{tr}(S_{22} \Sigma_{22}^{-1}) \Big) \times \\
& \exp\Big(-\frac{n}{2} \text{tr}\Big[S_{22} (\hat{\text{B}} - \text{B})' \Sigma_{11.2}^{-1} (\hat{\text{B}} - \text{B}) \Big] \Big).
\end{aligned}
\tag{8.8}
$$

Equation 8.8 shows that the multivariate normal model can be reparametrized such that the sets of parameters μ, $\Sigma_{11.2}$, Σ_{22}, and B can be integrated out independently of one another, if the prior factorizes similarly.

8.A.2 Integrating out the nuisance parameters

Equation 8.8 shows that the multivariate normal likelihood factorizes in four parts, one that involves the parameter of interest, one that involves the nuisance block regarding the variance-covariance of the conditioning variables Z, one that involves the regression of X and Y on Z, and the vector of means. This decomposition suggests prior distributions of the form $\pi(\mu, \Sigma) = \pi(\mu)\pi(\text{B})\pi(\Sigma_{22})\pi(\Sigma_{11.2})$. Because μ and B act as location parameters and Σ_{22} represents a scaling factor, we choose the following priors: $\mu \propto 1$, $\text{B} \propto 1$, and $\Sigma_{22} \propto |\Sigma_{22}|^{-(k+1)/2}$. Note that improper priors generally do not cause complications for the computation of Bayes factors, with the provisos that the joint posterior distribution is proper, and that the improper priors are assigned exclusively to nuisance parameters and emphatically not to the test-relevant parameter (cf. Ly et al., 2016a; Wagenmakers & Ly, 2023).

Integrating out μ

With improper priors $\mu_i \propto 1$ for $i \in 1, \ldots, p$ we can integrate μ out of the likelihood. This leads to a simple Gaussian integral and costs one degree of

freedom:

$$\begin{aligned}\int \exp\left(-\frac{1}{2}(\bar{w}-\mu)'(n^{-1}\Sigma)^{-1}(\bar{w}-\mu)\right)\, d\mu &= |2\pi n^{-1}\Sigma|^{1/2} \\ &= (2\pi n^{-1})^{p/2}|\Sigma|^{1/2}.\end{aligned} \tag{8.9}$$

The resulting reduced likelihood is

$$\begin{aligned}f(d \mid \Sigma_{11.2}, \Sigma, \mathrm{B}) =& (2\pi)^{p(1-n)/2} n^{-p/2} |\Sigma_{11.2}|^{(1-n)/2} |\Sigma_{22}|^{(1-n)/2} \times \\ & \exp\Big(-\frac{n}{2}\mathrm{tr}(S_{11.2}\Sigma_{11.2}^{-1})\Big)\times \\ & \exp\Big(-\frac{n}{2}\mathrm{tr}(S_{22}\Sigma_{22}^{-1})\Big)\times \\ & \exp\Big(-\frac{n}{2}\mathrm{tr}\left[S_{22}(\hat{\mathrm{B}}-\mathrm{B})'\Sigma_{11.2}^{-1}(\hat{\mathrm{B}}-\mathrm{B})\right]\Big),\end{aligned} \tag{8.10}$$

leaving us with $p(p+1)/2$ parameters.

Integrating out B

To integrate out B, observe that the relevant part of the likelihood is a kernel of a $2 \times k$ matrix normal distribution with a location matrix $\mathrm{B} \in \mathbb{R}^{2\times k}$, one 2×2 scale matrix $\Sigma_{11.2}$, and a second, $k \times k$, scale matrix S_{22}. By assigning priors on $\mathrm{B}_{ij} \propto 1$ for $i \in 1, 2$, and $j \in 1, \ldots, k$, we get:

$$\begin{aligned}\int \exp\Big(-\frac{n}{2}\mathrm{tr}\left[S_{22}(\hat{\mathrm{B}}-\mathrm{B})'\Sigma_{11.2}^{-1}(\hat{\mathrm{B}}-\mathrm{B})\right]\Big)\, d\mathrm{B} =& (2\pi)^k n^{-k} \times \\ & |S_{22}|^{-1}|\Sigma_{11.2}|^{k/2},\end{aligned} \tag{8.11}$$

provided that $|S_{22}| > 0$ and $|\Sigma_{11.2}| > 0$. The resulting reduced likelihood is

$$\begin{aligned}f(d \mid \Sigma_{11.2}, \Sigma) =& |S_{22}|^{-1}(2\pi)^{p(1-n)/2+k} n^{-p/2-k} \times \\ & |\Sigma_{11.2}|^{(1-n+k)/2} |\Sigma_{22}|^{(1-n)/2} \times \\ & \exp\Big(-\frac{n}{2}\mathrm{tr}(S_{11.2}\Sigma_{11.2}^{-1})\Big)\times \\ & \exp\Big(-\frac{n}{2}\mathrm{tr}(S_{22}\Sigma_{22}^{-1})\Big),\end{aligned} \tag{8.12}$$

leaving us with $3 + k(k+1)/2$ parameters.

Integrating out Σ_{22}

To integrate out Σ_{22}, we collect the relevant part of the likelihood and the prior and observe that it results in a Wishart integral:

$$\begin{aligned}
&\int |\Sigma_{22}|^{-(k+1)/2} |\Sigma_{22}|^{(1-n)/2} \exp\left(-\frac{n}{2}\text{tr}(S_{22}\Sigma_{22}^{-1})\right) d\Sigma_{22} = \\
&\int |\Sigma_{22}|^{-(n+k)/2} \exp\left(-\frac{n}{2}\text{tr}(S_{22}\Sigma_{22}^{-1})\right) d\Sigma_{22} = \\
&2^{(n+k)k/2} \times \Gamma_k\left(\frac{n+k}{2}\right) \times |S_{22}|^{-(n-1)/2},
\end{aligned} \tag{8.13}$$

where $\Gamma_k(.)$ is the k-variate gamma function. The resulting reduced likelihood is

$$\begin{aligned}
f(d \mid \Sigma_{11.2}) = |\Sigma_{11.2}|^{(1-n+k)/2} \exp\Big(-\frac{n}{2}\text{tr}(S_{11.2}\Sigma_{11.2}^{-1})\Big) \times \\
\underbrace{2^{(n+k)k/2} \times \Gamma_k\left(\frac{n+k}{2}\right) |S_{22}|^{-(1+n)/2} (2\pi)^{p(1-n)/2+k} n^{-p/2-k}}_{C(n,k,S_{22})},
\end{aligned} \tag{8.14}$$

leaving us with just three parameters.

8.A.3 Final derivation of the Bayes factor and the marginal posterior distribution

The resulting reduced likelihood contains only three parameters: the partial correlation $\rho_{xy.z}$ and the residual variances $\sigma^2_{x.z}$ and $\sigma^2_{y.z}$. Expressing the reduced likelihood Equation 8.14 in terms of these quantities yields

$$
\begin{aligned}
&f(s_{x.z}, s_{y.z}, r_{xy.z} \mid \sigma_{x.z}, \sigma_{y.z}, \rho_{xy.z}) = \\
&C(n, k, S_{22}) \times \\
&\left(\sqrt{1-\rho_{xy.z}^2}\sigma_{x.z}\sigma_{y.z}\right)^{1-n+k} \times \\
&\exp\left(-\frac{n}{2(1-\rho_{xy.z}^2)}\left[\frac{s_{x.z}^2}{\sigma_{x.z}^2} - 2\rho_{xy.z}\frac{r_{xy.z}s_{x.z}s_{y.z}}{\sigma_{x.z}\sigma_{y.z}} + \frac{s_{y.z}^2}{\sigma_{y.z}^2}\right]\right).
\end{aligned}
$$

which is similar to the likelihood for the Pearson's correlation coefficient described in (Ly et al., 2018, Eq. 8), the only difference being the additional k term in the exponent of the determinant. This allows us to use of Ly et al. (2018). In sum, if we use the prior

$$\mu \propto 1, \mathrm{B} \propto 1, \Sigma_{22} \propto |\Sigma_{22}|^{-(k+1)/2}, \tag{8.15}$$

$$\sigma_{x.z} \propto \sigma_{x.z}^{\gamma-1}, \sigma_{y.z} \propto \sigma_{y.z}^{\delta-1}, \text{ and} \tag{8.16}$$

$$\rho \sim \text{stretched-beta}(\alpha, \alpha) \tag{8.17}$$

we then obtain the Bayes factor for the alternative model over the null model is given by

$$\mathrm{BF}_{10} = \frac{\mathcal{B}\left(\frac{1}{2}, \tilde{\alpha}\right)}{\mathcal{B}\left(\frac{1}{2}, \alpha\right)} \times {}_2F_1\left(\frac{n-k-\gamma-1}{2}, \frac{n-k-\delta-1}{2}; \tilde{\alpha} + \frac{1}{2}; r_{xy.z}^2\right), \tag{8.18}$$

where $\tilde{\alpha} = \alpha + \frac{n-k-\gamma-\delta-1}{2}$. For $\gamma = \delta = 0$, the marginal posterior distribution of $\rho_{xy.z}$ is

$$\pi(\rho_{xy.z} \mid n, k, r_{xy.z}) = \frac{(1-\rho^2_{xy.z})^{(2\alpha+n-k-\gamma-\delta-3)/2}}{\mathcal{B}\left(\frac{1}{2}, \alpha + \frac{n-k-\gamma-\delta-1}{2}\right)\ {}_2F_1\left(\frac{n-k-\gamma-1}{2}, \frac{n-k-\delta-1}{2}; \alpha + \frac{n-k-\gamma-\delta}{2}; r^2_{xy.z}\right)} \times \left[{}_2F_1\left(\frac{n-k-\gamma-1}{2}, \frac{n-k-\delta-1}{2}; \frac{1}{2}; r^2_{xy.z}\rho^2_{xy.z}\right) + 2r_{xy.z}\rho_{xy.z}W_{\gamma,\delta}(n-k)\ {}_2F_1\left(\frac{n-k-\gamma}{2}, \frac{n-k-\delta}{2}; \frac{3}{2}; r^2_{xy.z}\rho^2_{xy.z}\right)\right], \tag{8.19}$$

where $W_{\gamma,\delta}(\tilde{n})$ is defined in Ly et al. (2018, p. 7) as

$$W_{\gamma,\delta}(\tilde{n}) = \frac{\Gamma\left(\frac{\tilde{n}-\gamma}{2}\right)\Gamma\left(\frac{\tilde{n}-\delta}{2}\right)}{\Gamma\left(\frac{\tilde{n}-\gamma-1}{2}\right)\Gamma\left(\frac{\tilde{n}-\delta-1}{2}\right)}. \tag{8.20}$$

Special case of overwhelmingly informative data

With $n \geq n_{\text{min}}$, thus, $\tilde{n} \geq 3$, where $\tilde{n} = n - k$ and with $r = \pm 1$ a straightforward computation shows that

$$\text{BF}_{10} = \frac{2^{1-2\alpha}\sqrt{\pi}}{\mathcal{B}(\alpha,\alpha)} \frac{\Gamma(\alpha + \frac{\tilde{n}-1}{2})\Gamma(\alpha + \frac{\tilde{n}}{2})\Gamma(\alpha + 1 - \frac{\tilde{n}}{2})}{\Gamma(\alpha + \frac{\tilde{n}}{2})\Gamma(\alpha + \frac{1}{2})\Gamma(\alpha + \frac{1}{2})}. \tag{8.21}$$

This Bayes factor diverges, when one of the gamma functions in the numerator has a non-positive argument, thus, when $\alpha + 1 - \frac{\tilde{n}}{2} \leq 0$, hence, when $\alpha + 1 \leq \frac{\tilde{n}}{2}$. For this to already occur at $n = n_{\text{min}}$ we require $\alpha \leq 1/2$.

8.B Rewriting the trace

Using the inversion of a block matrix, the matrix Σ^{-1} can be rewritten in a block form as:

$$\Sigma^{-1} = \begin{bmatrix} \Sigma^{-1}_{11.2} & -\Sigma^{-1}_{11.2}\Sigma_{12}\Sigma^{-1}_{22} \\ -\Sigma^{-1}_{22}\Sigma_{21}\Sigma^{-1}_{11.2} & \Sigma^{-1}_{22} + \Sigma^{-1}_{22}\Sigma_{21}\Sigma^{-1}_{11.2}\Sigma_{12}\Sigma^{-1}_{22} \end{bmatrix}.$$

Similarly, we can write S as a block matrix:

$$S = \begin{bmatrix} S_{11.2} + S_{12}S_{22}^{-1}S_{21} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

Assuming $|S_{22}| > 0$, we can expand the trace inside of the kernel of the multivariate normal distribution, and rearrange the terms using the cyclic property of traces:

$$\begin{aligned} \text{tr}(S\Sigma^{-1}) =& \text{tr}\left(S_{11.2}\Sigma_{11.2}^{-1}\right) + \\ & \text{tr}\left(S_{22}\Sigma_{22}^{-1}\right) + \\ & \text{tr}\left(S_{12}S_{22}^{-1}S_{21}\Sigma_{11.2}^{-1}\right) + \\ & \text{tr}\left(-S_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\right) + \\ & \text{tr}\left(-S_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\right) + \\ & \text{tr}\left(S_{22}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\right) \\ =& \text{tr}\left(S_{11.2}\Sigma_{11.2}^{-1}\right) + \\ & \text{tr}\left(S_{22}\Sigma_{22}^{-1}\right) + \\ & \text{tr}\left(S_{22}^{-1}S_{21}\Sigma_{11.2}^{-1}S_{12}\right) + \\ & \text{tr}\left(-S_{22}^{-1}S_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}S_{22}\right) + \\ & \text{tr}\left(-\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}S_{12}\right) + \\ & \text{tr}\left(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}S_{22}\right), \end{aligned} \tag{8.22}$$

There are now six traces in the expression. The first two traces involve the conditional variance-covariance matrix $\Sigma_{11.2}$ and the variance-covariance matrix Σ_{22} of the controlling variables, and their sample counterparts. The latter four traces involve the cross correlations between the controlled and controlling variables. After rearranging the terms in Equation 8.22, the matrices inside of the four traces have the same dimensions, and so can be collected in a single trace, using the linearity property of traces. We can rearrange the terms further

and factor out common terms to obtain the final quadratic expression:

$$
\begin{aligned}
&\text{tr}(S\Sigma^{-1}) - \text{tr}\left(S_{11.2}\Sigma_{11.2}^{-1}\right) - \text{tr}\left(S_{22}\Sigma_{22}^{-1}\right) = \\
&\text{tr}\big(S_{22}^{-1}S_{21}\Sigma_{11.2}^{-1}S_{12} - S_{22}^{-1}S_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}S_{22} - \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}S_{12} + \\
&\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}S_{22}\big) = \\
&\text{tr}\left((S_{22}^{-1}S_{21}\Sigma_{11.2}^{-1} - \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1})(S_{12} - \Sigma_{12}\Sigma_{22}^{-1}S_{22})\right) = \\
&\text{tr}\big((S_{22}^{-1}S_{21} - \Sigma_{22}^{-1}\Sigma_{21})\Sigma_{11.2}^{-1}(S_{12}S_{22}^{-1} - \Sigma_{12}\Sigma_{22}^{-1})S_{22}\big) = \\
&\text{tr}\big(S_{22}(S_{12}S_{22}^{-1} - \Sigma_{12}\Sigma_{22}^{-1})'\Sigma_{11.2}^{-1}(S_{12}S_{22}^{-1} - \Sigma_{12}\Sigma_{22}^{-1})\big).
\end{aligned}
$$

This quadratic expression replaces the four terms involving the cross correlations:

$$
\begin{aligned}
\text{tr}(S\Sigma^{-1}) = &\text{tr}(S_{11.2}\Sigma_{11.2}^{-1}) + \\
&\text{tr}(S_{22}\Sigma_{22}^{-1}) + \\
&\text{tr}\big[S_{22}(S_{12}S_{22}^{-1} - \Sigma_{12}\Sigma_{22}^{-1})'\Sigma_{11.2}^{-1}(S_{12}S_{22}^{-1} - \Sigma_{12}\Sigma_{22}^{-1})\big].
\end{aligned}
$$

I'm going to make a prediction–it could go either way.

–Ron Atkinson

Chapter 9

Correct Conclusions from Fallible Medical Tests: A Tutorial with JASP

This chapter is preprinted as Kucharský, Š., and Wagenmakers, E.-J. (2023). Correct conclustions from fallible medical tests: A tutorial with JASP. *OSF Preprints*. doi: 10.31219/osf.io/jksz6

Abstract

Medical professionals, patients, students, and the public at large regularly need to interpret the outcome of medical tests. These tests are error-prone, however, and the fact that the outcome is positive (or negative) does not establish with certainty that the disease is present (or absent). The correct interpretation of the test outcome demands that Bayes' rule is used to combine the quality of the test (i.e., *sensitivity* and *specificity*) with available background information (i.e., disease *prevalence*). It is well known that people find it difficult to understand and apply Bayes' rule, and that the correct outcome is often at odds with intuition, especially when a test with good operating characteristics yields a positive result, but the disease is relatively rare. Less well known is that in practical application, the values of sensitivity, specificity, and prevalence are usually associated with considerable uncertainty. The correct interpretation of a medical test demands that this uncertainty is explicitly acknowledged and properly taken into account.

To facilitate the correct interpretation of fallible medical tests we introduce the `Binary classification` module in the open-source software JASP. This module explains medical testing through a series of informative visualizations. The module also allows users to propagate uncertainty in sensitivity, specificity, and prevalence to derived measure of interest, such as positive predictive value. The module can be used both in teaching and in medical practice.

9.1 Introduction

Binarius, a 23-year old student at NYU, experiences fatigue and headache. Binarius fears having contracted malaria after a recent visit to Central Africa. They decide to take a rapid diagnostic test (RDT) for malaria and it comes back positive. What is the probability that Binarius has the disease? In technical terms, what is the *positive predictive value*? This scenario typifies a ubiquitous problem in medical testing that confronts all medical students at some point during their studies. Despite the importance of correctly interpreting the test outcome, and despite the apparent simplicity of the inference, it is well documented that people (including medical professionals) often get it wrong (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 2004; Marewski & Gigerenzer, 2022; Villejoubert & Mandel, 2002).

Back to Binarius. Suppose you learn that the relevant RDT tends to return

a positive result for 84 out of every 100 positive cases (i.e., the test's *sensitivity* equals 84%). A common mistake is to conclude that Binarius's probability of having the disease is 84%, an error of interpretation known as 'confusion of the inverse', 'transposing the conditional', or simply the 'inverse fallacy' (Villejoubert & Mandel, 2002). In the case of Binarius, the error involves confusing the conditional probability of $p(\text{positive test} \mid \text{malaria})$ (which is known to be 84%) with the desired positive predictive value $p(\text{malaria} \mid \text{positive test})$. The transposition is incorrect for two reasons. Firstly, the transposition ignores the test's *specificity*, that is, the probability of returning a negative result given that the disease is absent – and its complement, the probability of returning a positive result given that the disease is absent (i.e., the false alarm rate or false positive rate). For the sake of the argument, suppose that Binarius used an RDT with a specificity of only 16%; this means that the RDT tends to return a false positive result for 84 out of 100 negative cases – in other words, the probability of a positive outcome is 84% if Binarius does *not* have malaria. But the probability of a positive outcome was also 84% if he does have malaria; hence, the positive outcome is completely uninformative about Binarius' disease status. Secondly, the transposition also ignores the *prevalence* of the disease. In order to compute the positive predictive value, we need to take into account not only the characteristics of the test (i.e., sensitivity and specificity), but also the relevant background information (Bianchi, Alexander, & Cash, 2009); the outcome of an otherwise excellent test cannot be interpreted at face value when prevalence is low (Lau & Aw, 2021). The tendency for people to disregard prevalence (or not take it into account fully) is known as 'base-rate neglect' or the 'prosecutor's fallacy' (Kahneman & Tversky, 1973; Welsh & Navarro, 2012). A concrete example will be presented shortly.

These errors of reasoning can be eliminated if one is trained in Bayesian inference and one is able to recognize that the problem at hand calls for the application of Bayes' theorem. However, this is often not the case and both medical professionals and general audiences often need guidance on how to interpret the outcome of diagnostic medical tests correctly (Bianchi et al., 2009; Klement & Bandyopadhyay, 2021; Lau & Aw, 2021; Pepe, 2003; Watson, Richter, & Deeks, 2020; Watson, Whiting, & Brush, 2020). It is noteworthy that even researchers, supposedly trained to avoid such mistakes, sometimes fail to apply

the correct reasoning in their scientific work (Suojanen, 1999).

Several remedial teaching programs have been suggested to improve the interpretation of test results (Galesic, Garcia-Retamero, & Gigerenzer, 2009; Gigerenzer, 1996; Gigerenzer et al., 2007; Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 2004; Marewski & Gigerenzer, 2022). The key ideas behind these proposals are (1) to discuss concrete examples instead of relying on general mathematical formulae; (2) to show frequencies and counts of events to supplement or supplant the traditional representation in terms of proportions and probabilities; and (3) to explain the same reasoning process from different perspectives.

These teaching programs generally do not address another fundamental problem in the interpretation of medical test results, one that has received surprisingly little attention over the years. Specifically, the computation of positive predictive value is usually demonstrated as if the values for sensitivity, specificity, and prevalence were known exactly. In reality, these quantities are associated with considerable uncertainty (Baron, 1994). A correct interpretation of the test outcome requires that this uncertainty is first quantified and then propagated to a derived measure of interest such as positive predictive value.

Here we demonstrate how the correct interpretation of medical tests results can be facilitated through the use of the `Binary Classification` module in JASP, an open-source statistical software program developed at the University of Amsterdam (`jasp-stats.org`).[1] The `Binary Classification` module was partly inspired by existing software applications (e.g., Crawford, Garthwaite, & Betkowska, 2009; Lenhard & Lenhard, 2014; Watson, Richter, & Deeks, 2020; Watson, Whiting, & Brush, 2020). Below we showcase the functionality of the `Binary Classification` module, first for the common scenario in which the parameter values are assumed to be known exactly, and then for the more realistic scenario in which they are associated with considerable uncertainty.

[1] Annotated `.jasp` files with examples are provided at `osf.io/kue5h`.

9.2 Example 1: Parameters Known Exactly

Consider again the example of Binarius who tested positive for malaria and wishes to assess the probability that they have actually contracted the disease. In most educational programs, such an example would be used to expose students to Bayes' theorem, which can be employed to calculate the desired positive predictive value:

$$\begin{aligned} &p(\text{Condition} = \text{positive} \mid \text{Test} = \text{positive}) \\ &\quad = \frac{p(\text{Condition} = \text{positive}) \times p(\text{Test} = \text{positive} \mid \text{Condition} = \text{positive})}{p(\text{Test} = \text{positive})} \\ &\quad = \frac{\text{Prevalence} \times \text{Sensitivity}}{\text{Prevalence} \times \text{Sensitivity} + (1 - \text{Prevalence}) \times (1 - \text{Specificity})}. \end{aligned} \tag{9.1}$$

Instead of having students apply the theorem manually, JASP calculates all necessary information and provides detailed output with illustrative figures in order to build a better intuition for the correct conclusion. Having entered values for the three key parameters *sensitivity*, *specificity*, and *prevalence*[2], the sole focus can then be on the proper interpretation of the JASP output.

For the interpretation of the malaria test result of Binarius, we take parameter values from Berzosa et al. (2018) who reported sensitivity and specificity of the relevant RDT to be 83.74% and 89.11%, respectively. Deciding on the value of prevalence is more challenging, as we need to estimate the proportion of people who have malaria, out of the set of people who have recently visited Central Africa and show symptoms of fatigue and headache. However, these symptoms can have many other causes (such as influenza or jet lag); we therefore make an educated guess and set prevalence to $\sim 10\%$. We enter these estimates in the JASP `Binary Classification` module as shown in Figure 9.1. We also choose to display plots that are meant to aid intuition.

Upon receiving the parameter values, JASP immediately updates the calculations and shows the results in the output panel. The primary output is shown

[2] In some cases it is appropriate to replace the term *prevalence* by the term *pre-test probability* or *base rate probability* to indicate that the *a priori* probability of someone having a disease may not only depend on the proportion of the population with a condition at a specific point in time, but on other factors as well – exposure to a pathogen, medical history, etc. Here we will use these terms interchangeably.

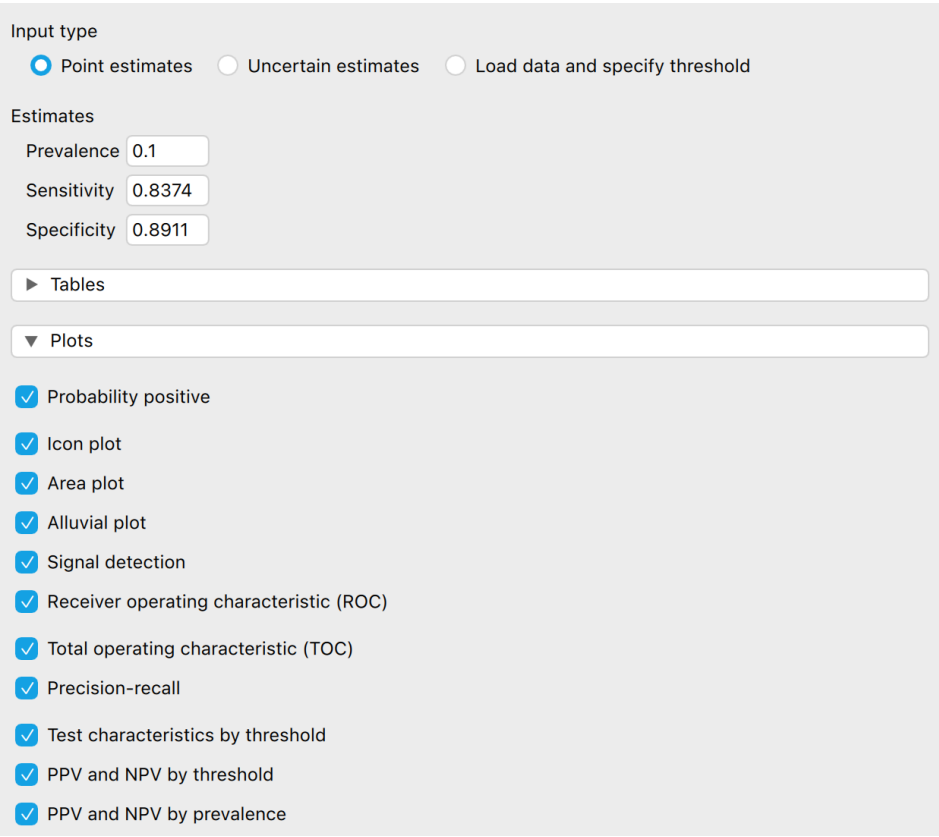


Figure 9.1: JASP screenshot of the input GUI for the `Binary Classification` module with parameters known exactly. All plots have been selected for display.

in Table 9.1 and contains the common quantities of interest. We may confirm that the values of prevalence, sensitivity, and specificity correspond to the values that were entered in the input panel. This information is followed by the main entries from a confusion matrix (i.e., the estimated proportion of true positives, false positives, true negatives, and false negatives) that we would expect given the properties of the test and the prevalence. Next, the table presents the positive and negative predictive value, and their complements false discovery rate and false omission rate, respectively. Complements are also presented for sensitivity and specificity, that is, false negative rate and false positive rate, respectively. Finally, the table also presents general accuracy, that is, the proportion of cases for which the test outcome corresponds to the underlying condition, regardless of whether it is positive or negative.

Reading the results directly from the table, we would estimate Binarius's probability of having malaria to be about 46% – the positive test result has considerably increased the probability of Binarius having malaria (as the starting prevalence was 10%), but the most likely scenario remains that Binarius does *not* have malaria.

Obtaining the correct answer with JASP may be a straightforward task, but understanding how is it derived requires more insight. JASP provides a collection of figures that examine the problem from different perspectives in order to build an intuition for the underlying reasoning without resorting to mathematical formulae. Below we will discuss the figures one at a time. Table 9.2 provides an overview of the available figures, including how they can increase understanding of the concepts behind binary classification.

Figure 9.2 shows how the base rate probability of having the disease is updated upon observing either a negative or a positive test result. This output can be requested by selecting the `Probability positive` option in the JASP input panel. The red orange bars show that before obtaining the test result, the probability of Binarius having malaria was 10%. The cyan bars show the probability that Binarius has malaria after obtaining the test result; if the test would have been negative, the probability would have decreased; however, the test was positive, and the probability that Binarius has malaria equals about 46%.

To help create an intuition of how prevalence, sensitivity, and specificity together determine the interpretation of a test result, Figures 9.3, 9.4, 9.5, and

	Estimate
Prevalence	0.1000
Sensitivity	0.8374
Specificity	0.8911
True positive	0.0837
False positive	0.0980
True negative	0.8020
False negative	0.0163
Positive predictive value	0.4607
Negative predictive value	0.9801
False discovery rate	0.5393
False omission rate	0.0199
False positive rate	0.1089
False negative rate	0.1626
Accuracy	0.8857

Table 9.1: Excerpt from the primary JASP output table for the malaria example of Binarius, with parameters known exactly. The top three entries (i.e., prevalence, sensitivity, and specificity) determine the values for the entries below. The table in JASP features additional columns that contain mathematical notation and a verbal description.

	Name	Description & Use-case
Figure 9.2	Probability positive	Depicts the probability of having the disease (1) before any test was conducted, (2) after a negative test, or (3) after a positive test. Provides a visual representation of how the probability changes depending on the test's outcome.
Figure 9.3	Icon plot	Displays a theoretical population in terms of icons (i.e., in a frequency format, Galesic et al., 2009; Gigerenzer & Hoffrage, 1995). Provides a first insight into Bayes' theorem.
Figure 9.4	Area plot	Represents a theoretical population in terms of rectangular areas. Provides a more detailed insight into Bayes' theorem by separating the effects of prevalence, sensitivity, and specificity into three independent axes. Inspired by Sanderson (2019).
Figure 9.5	Alluvial plot	Shows the four cells in a confusion matrix as colored links between "Condition" and "Test" scaled by their proportion in a population. Helps visualize sizes of the cells in relation to the total positive cases or total positive tests.
Figure 9.6	Signal detection	Represents the population from a signal detection perspective. Provides visual justification for threshold, a concept that is important to understand the figures below.
Figure 9.7	ROC curve	Traces the true positive rate against the false positive rate for different values of threshold, creating the classical ROC curve useful to understand the diagnostic characteristics of the test.
Figure 9.8	TOC curve	An alternative to the ROC plot that takes prevalence into account. Compared to the ROC, this plot is useful when the sizes of the four individual cells in a confusion matrix are important.
Figure 9.9	PR curve	Traces the positive predictive value against recall for different values of threshold is useful in asymmetric scenarios (i.e., when prevalence is low).
Figure 9.10a	Test characteristics against threshold	Provides the same information as ROC, but with sensitivity and specificity plotted separately as a function of the threshold, making the threshold values explicit. Useful to understand how moving the threshold makes the test more or less conservative.
Figure 9.10b	Predictive values against threshold	PPV and NPV shown separately as a function of the threshold. Useful to understand how making the test more or less conservative affects the interpretation of the test result.
Figure 9.10c	Predictive values against prevalence	PPV and NPV shown separately as a function of prevalence. Useful to visualize how prevalence affects the outcome of the test result.

Table 9.2: Overview of figures in JASP meant to foster intuition for the correct interpretation of medical test outcomes.

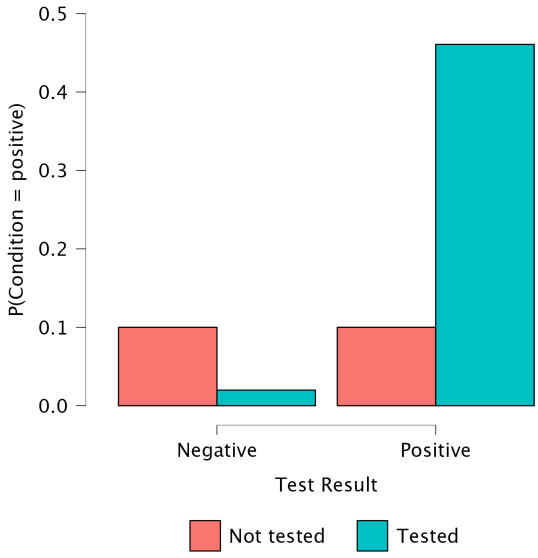


Figure 9.2: Probability of having the disease under different scenarios: Before testing, after obtaining a negative test, and after obtaining a positive test. Output from JASP.

9.6 show a representation of a hypothetical population, segmented by disease status and by whether or not they tested negative or positive.

People often have a better intuition for counts than for proportions (e.g., Galesic et al., 2009; Gigerenzer et al., 2007; Gigerenzer & Hoffrage, 1995). This can be exploited by presenting the relevant information as an *icon plot*. For instance, Figure 9.3 displays the situation as a population of one hundred icons. Given that prevalence is relatively low (90 out of 100 people do not have the disease) and specificity is relatively high, most of the population would be true negative; here, about 80 in 100 (the blue icons). However, even with a relatively high specificity, we nevertheless obtain a number of false positive cases; here, about 10 in 100 (the yellow icons). This number outweighs the number of true positives in the population; here, about 8 in 100 (the green icons), and so even after Binarius tested positive it is somewhat more likely that they are false positive than true positive.

Figure 9.4 (the area plot) builds on the intuition from the icon plot, but adds more information at the expense of representing the quantities of inter-

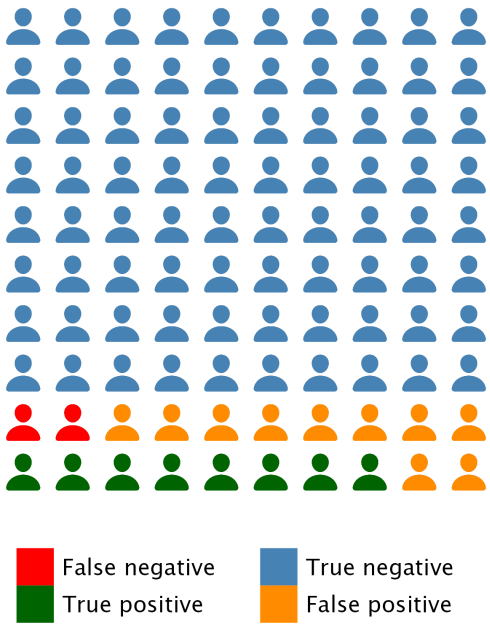


Figure 9.3: The icon plot displays a hypothetical population of a 100 people that either have or do not have the disease, and test either negative or positive. Output from JASP.

est with rectangular areas instead of counts. Specifically, the plot is divided into four rectangles corresponding to false negative, true negative, true positive, and false positive cases, which are arranged along three axes: prevalence on the bottom, sensitivity on the left, and specificity on the right. The areas of the individual rectangles are easily computed by multiplying appropriate quantities. For example, the green 'True positive' area corresponds to the proportion of the population that contains the disease and that would be correctly identified by a positive test. Therefore we can calculate the area of the green rectangle by multiplying the lengths of its adjacent sides: prevalence $\times$ sensitivity $= 0.1 \times 0.8374 = 0.08374$. Similarly, the orange 'False positive' area corresponds to the proportion of the population that does not contain the disease but is incorrectly tested positive. Thus, the area of the rectangle is $(1 - \text{prevalence}) \times (1 - \text{specificity}) = 0.9 \times 0.1089 = 0.09801$. The PPV is represented as the proportion of true positive cases relative to the proportion

of cases that tested positive, a sum of the 'True positive' and 'False positive' rectangles (highlighted in the plot by a dashed line). Thus, the PPV is calculated as:

$$\text{PPV} = \frac{\text{Green rectangle}}{\text{Green rectangle} + \text{Orange rectangle}} = \frac{0.08374}{0.08374 + 0.09801} \approx 0.46. \tag{9.2}$$

Thus, area plots such as Figure 9.4 can serve as a visual aid in applying Bayes' theorem. The advantage of this layout is that changing either prevalence, sensitivity, or specificity alters only a single aspect of the rectangles, which facilitates a better understanding of how the three parameters affect the interpretation of the test outcome. For example, increasing prevalence would increase the width of the green rectangle and decrease the width of the orange rectangle while keeping their heights constant, meaning that the proportion 'True positives' increases and the proportion 'False positives' decreases. Thus, increasing (decreasing) prevalence while keeping sensitivity and specificity constant can be easily seen to increase (decrease) the PPV. The area plot was inspired by Sanderson (2019); to the best of our knowledge it has not yet been applied in educational practice, other than by ourselves.

Figure 9.5 shows the same information as the previous figures but rearranged in a so-called 'alluvial plot' (Brunson, 2020); the left bar 'Condition' shows the proportion of the population that either has the disease ('Positive') or not ('Negative'), whereas the right bar 'Test' shows the proportion of the population that tested 'Positive' or 'Negative'. The colored links that connect the two bars correspond to true positives, false positives, false negatives, and true negatives. For example, the green link corresponds to people who have the disease and also test positive; the orange link corresponds to people who do not have the disease but test positive.

Figure 9.6 shows the testing scenario through the lens of signal detection theory (Lee, 2008; Paulewicz & Blaut, 2020). For people who do not have the disease, the values of a hypothetical 'Marker' variable are assumed to follow a standard normal 'noise' distribution. For people who do have the disease, the 'Marker' values follow a 'signal' distribution that is shifted to the right of the 'noise' distribution. Here, the two distributions are weighted by prevalence;

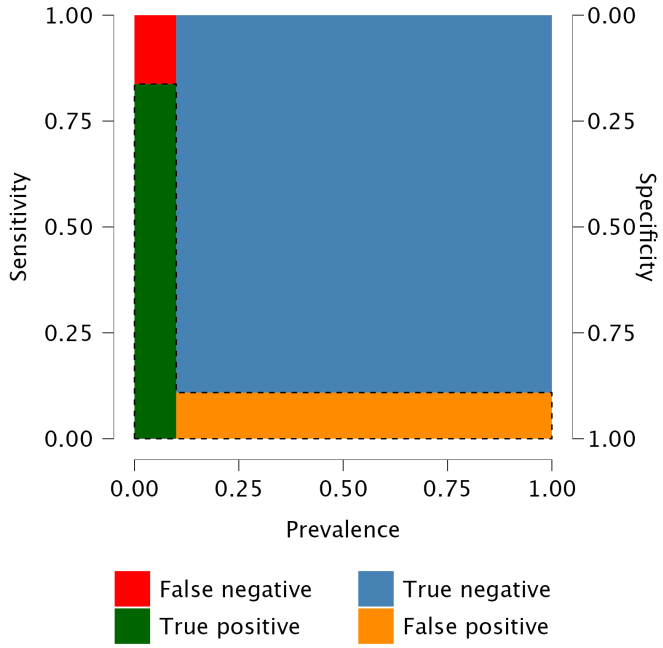


Figure 9.4: The area plot displays the proportions of true positives, false positives, false negatives, and true negatives as rectangles on a unit square. Output from JASP.

for our running example, this means that the area under the 'noise' distribution integrates to 0.9 whereas the area under the 'signal' distribution integrates to 0.1. Next, we introduce a 'decision threshold' (i.e., the dashed vertical line). When the 'Marker' value falls below (vs. above) the threshold, the test result is labeled as 'negative' (vs. 'positive'). The threshold divides each of the two distributions into two parts, creating four areas that correspond to the false negative, true negative, true positive, and false positive cases. With the distributions fixed, heightening the decision threshold will increase specificity (i.e., more true negative cases and fewer false positive cases) but decrease sensitivity (i.e., more false negative cases and fewer true positive cases).

The trade-off between sensitivity and specificity introduced by varying the decision threshold is important to understand, as it is a basic idea behind understanding or potentially constructing new diagnostic tests. To evaluate the performance of a binary classification model, we can plot quantities such as

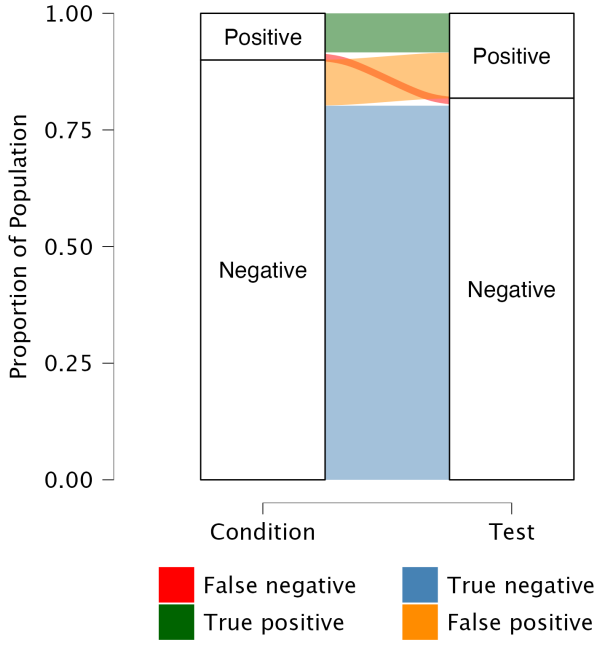


Figure 9.5: The alluvial plot. The 'Condition' bar on the left represents the *true* proportion of positive vs. negative cases in the population, whereas the 'Test' bar represents the proportion of those who test positive vs. negative. The links between the bars represent four possible combinations of the 'Condition' and 'Test'. Output from JASP.

sensitivity and specificity against each other for varying threshold values. This is illustrated by Figures 9.7, 9.8, and 9.9.

Figure 9.7 shows the popular ROC (Receiving Operating Characteristic) curve (Hoo, Candlish, & Teare, 2017). In the ROC plot, the x-axis represents the false positive rate, which measures the proportion of true negative cases that are incorrectly classified as positive (i.e., $1-$specificity), whereas the y-axis represents the true positive rate, indicating the proportion of true positive cases that are correctly classified as positive (i.e., sensitivity). As the decision threshold is 'pulled through' the distributions it traces out the relation between the false positive rate and the true positive rate: the ROC curve. The shape of the ROC curve helps researchers assess the performance of a classification model. A completely random classifier would result in a diagonal line (shown

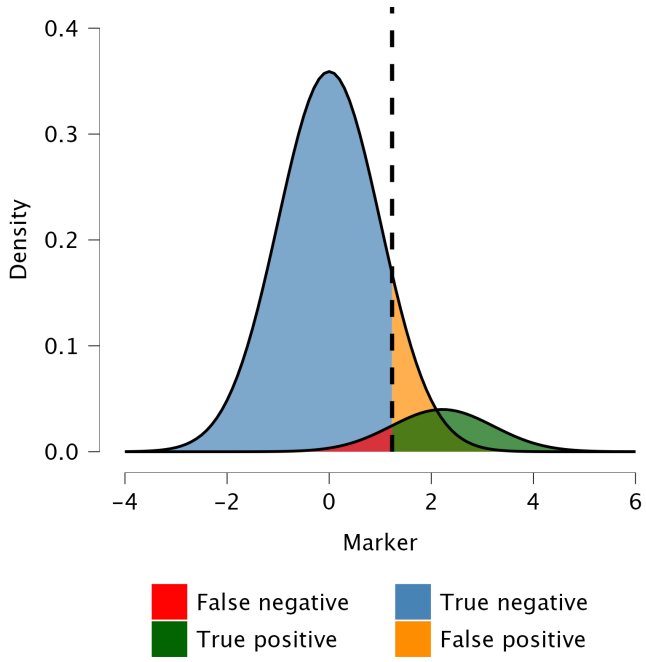


Figure 9.6: A signal detection theory perspective on the binary classification problem. Output from JASP.

as a dashed line in the plot) – the signal distribution and the noise distribution would overlap completely, and hence the false positive rate would equal the true positive rate. In contrast, a perfect classifier would create a curve that rises sharply from the origin to the upper left corner (100% sensitivity and 0% false positive rate). The area enclosed between the solid curve and the dashed diagonal can therefore be interpreted as an overall measure for the performance of the test. Note that the ROC does not feature prevalence.

Figure 9.8 shows the TOC (Total Operating Characteristic) curve (Pontius & Si, 2014). In the TOC plot, the x-axis represents the proportion of true positives plus the proportion of false positives, whereas the y-axis represents the proportion of true positives. By moving the decision threshold through the distributions the TOC curve is traced out. The curve is surrounded by a parallelogram that defines the theoretically possible bounds of the curve, given the restrictions imposed by the specified prevalence. In some situations, the TOC plot can be more informative than the ROC plot, as it is informed not only by

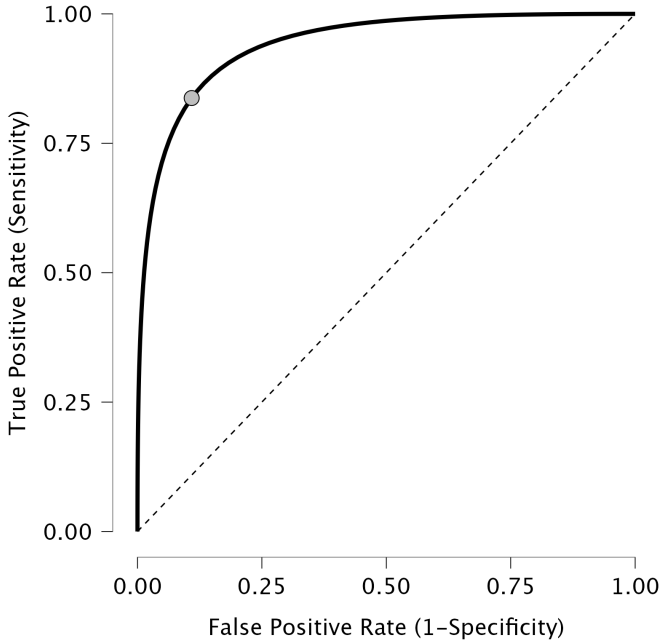


Figure 9.7: A receiving operating characteristic plot. The curve depicts values of true positive rate and false positive rate for different values of the decision threshold. The closer the curved line is to the diagonal dashed line, the lower the diagnostic value of the test. The grey dot represents the value at the decision threshold shown in Figure 9.6. Output from JASP.

sensitivity and specificity, but also by prevalence. Specifically, the plot is constructed such that it features all four elements of the confusion matrix (i.e., false negative, true negative, true positive, and false positive cases), by considering the distance of the curve to each of the sides of the parallelogram (Pontius & Si, 2014). Because the y-axis shows the proportion of true positives, the distance to the bottom side of the parallelogram corresponds to the proportion of true positives. The top side of the parallelogram is drawn at the value of prevalence, that is, the proportion of positive cases in the population. Therefore, the distance of the curve to the top side of the parallelogram corresponds to the proportion of false negatives. Along the horizontal axis, the distance between the left and right side of the parallelogram is the complement of prevalence, that is, the proportion of negative cases in the population. The TOC curve splits that distance

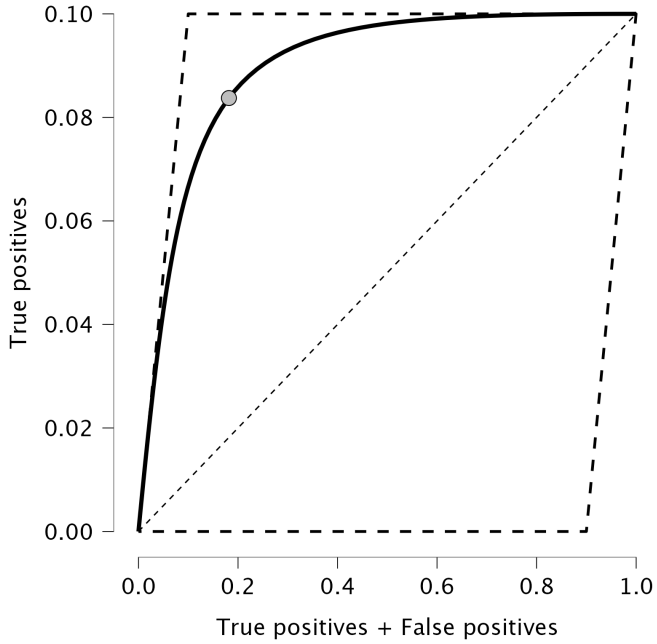


Figure 9.8: A total operating characteristic plot. The curve depicts proportions of true positives and true positives + false positives for different values of the threshold. The closer the curved line is to the diagonal dashed line, the lower the diagnostic value of the test. The grey dot represents the value at the decision threshold shown in Figure 9.6. Output from JASP.

in two parts, corresponding to the proportion of true negatives and false positives; the distance of the curve to the right side of the parallelogram corresponds to the proportion of true negatives, and the distance to the left side corresponds to the proportion of false positives. A completely random test would result in a diagonal line (shown as a dashed line in the plot). When the curve is closer to the upper left corner of the parallelogram, the test results in fewer false positives and false negatives, and more true positives and true negatives. A perfect classifier would create a curve that rises sharply from the origin to the upper-left corner of the parallelogram, indicating that the classifier judges all cases that have the disease to be positive, while having zero false-positive cases.

An alternative view of the TOC curve comes from considering the predictive value of a positive test that Binarius obtained. Notice that the y-axis shows

the numerator and the x-axis shows the denominator of Equation 9.1. The current threshold (depicted as a grey point) results in $\approx$0.08 on the y-axis, compared to $\approx 0.08 + 0.1 = 0.18$ on the x-axis, corresponding to the calculation of the PPV derived in Equation 9.2. In general, to obtain high PPV, the test threshold should be set to a value that results in a high value on the y-axis relative to the value on the x-axis.

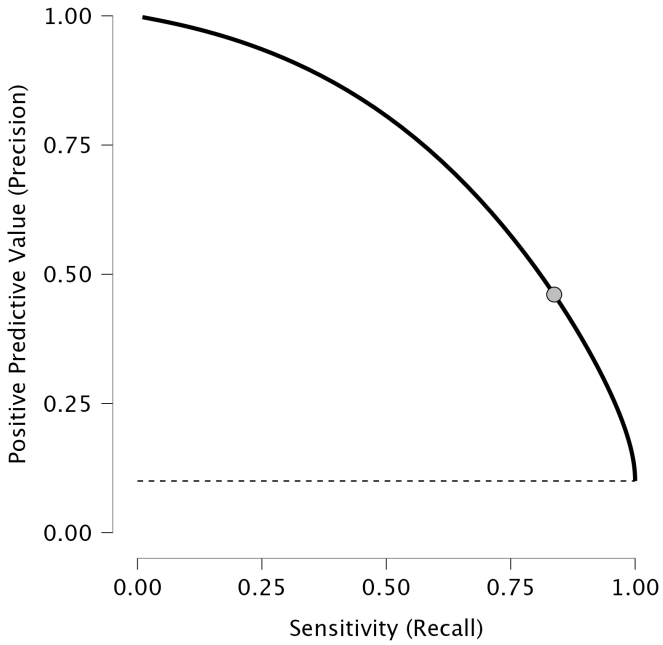


Figure 9.9: A precision-recall plot. The curve depicts precision and recall for different values of the threshold. The closer the curved line is to the prevalence (dashed line), the lower the diagnostic value of the test. Output from JASP.

Figure 9.9 shows the so-called precision-recall or PR curve (Cook & Ramadas, 2020). The PR curve focuses on two key metrics: 'precision' (i.e., positive predictive value) on the y-axis and 'recall' (i.e., sensitivity) on the x-axis. The focus on positive cases in this plot is particularly useful when negative cases dominate the population, making metrics such as overall accuracy less informative. Specifically, increasing sensitivity by decreasing the test's threshold results in a seemingly counterintuitive effect of decreasing positive predictive value. As the test becomes less conservative, it becomes better at detecting positive cases,

but also indicate more negative cases as positives. The set of negative cases that tested positive eventually start to dominate the set of positive cases that tested positive. A good test will result in a curve that turns sharply around the upper-right corner, indicating high precision and high recall. In contrast, a poor test would result in a curve that turns sharply around the lower-right corner and following the value of 'prevalence' (the dashed line) for PPV for any value of sensitivity, indicating that no matter what threshold we choose, obtaining a positive test does not change the base-rate probability of having the disease.

The ROC curve, the TOC curve, and the PR curve each highlight how two test characteristics trade off when the threshold is varied. These plots are useful to assess the general diagnostic value of the test: the larger the 'area under the curve' in each of the plots, the more diagnostic the test outcome, irrespective of the specific value for the threshold (Faraggi & Reiser, 2002; Pontius & Si, 2014; Sofaer, Hoeting, & Jarnevich, 2019). One downside of these plots is that the threshold values remain implicit; for instance, it is not possible to consider an ROC curve and learn about the false and true positive rates that accompany a threshold of, say, 1.2.

As a remedy, these curve plots can be supplemented with graphs where the two relevant characteristics are plotted separately as a function of a third quantity of explicit interest. For example, Figure 9.10 shows three panels. In panel (a), sensitivity and specificity are plotted separately as a function of threshold; in panel (b), PPV and NPV are plotted separately as a function of threshold; and in panel (c), PPV and NPV are plotted as a function of prevalence. We now discuss these panels in more detail.

First, Figure 9.10a shows the same trade-off as the ROC curve in 9.7 as it displays sensitivity and specificity when the threshold is set to different values. However, the threshold values are explicitly plotted on the x-axis, whereas sensitivity and specificity have their own separate curves on the y-axis. As a result, it now becomes clear that as threshold increases (i.e., the test becomes more conservative), sensitivity decreases and specificity increases.

Sensitivity and specificity are crucial to assess the performance of the test independent of prevalence. However, in a diagnostic scenario, we do care about prevalence as it directly affects the diagnosis. Instead of focusing on sensitivity and specificity, we may therefore wish to consider the positive and negative

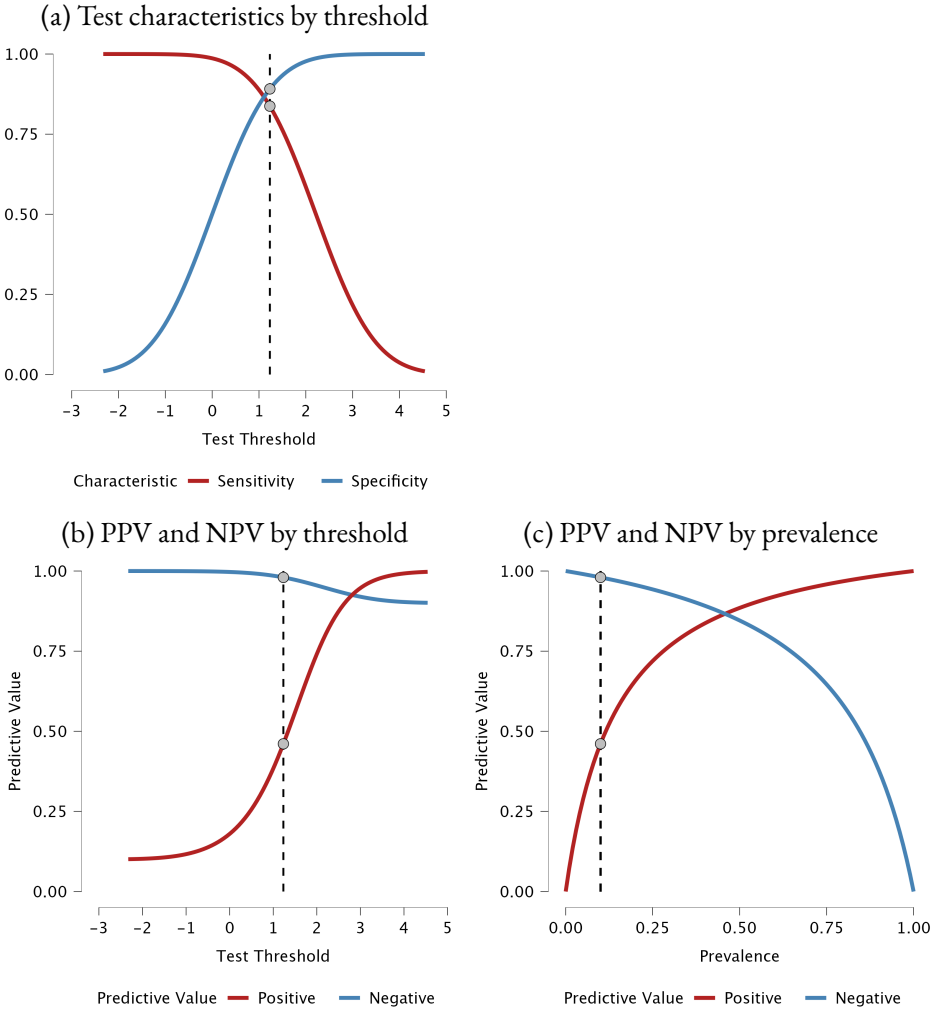


Figure 9.10: Trade-offs between two characteristics (y-axis) as a function of a third variable set to vary (x-axis). Plots on the left let threshold vary on the x-axis, whereas the plot on the right varies prevalence. The plot on the top shows sensitivity and specificity (i.e., test characteristics independent of prevalence), whereas the plots on the bottom show PPV and NPV (and thus take prevalence into account). Output from JASP.

predictive values. These values are shown in Figure 9.10b. When the threshold is set extremely low, essentially all cases test positive. This has the effect that a positive test conveys hardly any information and thus the PPV remains close to the base-rate probability. Conversely, as almost no cases are tested negative, obtaining a negative test conveys considerable information and thus the NPV tends towards 100%. As the threshold increases, the test becomes more and more conservative. As a result, obtaining a positive test provides more and more information and so PPV increases, whereas NPV decreases. At the extreme, the test is so conservative that to obtain a positive test one must have the disease, and therefore PPV tends towards 100%. In that case, however, essentially everyone tests negative, and so obtaining a negative test does not update the base-rate probability that one does *not* have the disease.

Plotting the NPV and PPV against the threshold is useful to understand how threshold affects the interpretation of a test outcome. However, in a typical medical scenario, the threshold is often already set to a fixed value and so inspecting how our results would have changed were the threshold set to a different value is perhaps not of immediate interest. What is of immediate interest, however, is how our conclusions would change if prevalence were different from what we assume. After all, the value of prevalence of 10% we set for the example of Binarius was just an educated guess, and it is likely that medical professionals would disagree about the best value to use. To understand how much our conclusions would change if prevalence were different, Figure 9.10c shows PPV and NPV as a function of prevalence. For example, while the PPV is around 0.46 under our original prevalence estimate of 10% prevalence, if that estimate had been set to 15% then the PPV would have been comfortably above 50%; and if the estimate had been 5% then the PPV would have been less than 30%. Interestingly, by changing only the estimate of prevalence, the values of both PPV and NPV can attain any arbitrary value. It is therefore clear that the correctness of our conclusions depends heavily on the correctness of our assumptions concerning prevalence.

Having examined the diagnostic test from many different angles, it would appear we are now in a good position to address the key question: "What is Binarius's probability that they actually have malaria after having received a positive test?" Analogous to the analyses presented above, the traditional analysis

presented in typical textbook scenarios would have us believe that Binarius has contracted malaria with a probability of $\approx 46\%$. However, we have already seen that the conclusions depend critically on the actual value of prevalence and the characteristics of the test. Therefore, we must ask ourselves how confident we really are that the prevalence is exactly 10%? Moreover, was the study reported by Berzosa et al. (2018) based on a large enough sample such that it would be reasonable to assume that sensitivity and specificity are exactly equal to 83.74% and 89.11%, respectively? Where there is room for doubt, there is a need for propagation of uncertainty. The next section demonstrates how uncertainty in prevalence, sensitivity, specificity can be taken into account when drawing conclusions from a fallible test.

9.3 Example 2: Parameters Subject to Uncertainty

The previous section outlined how to interpret the outcome of a medical test assuming known values of the three parameters: sensitivity, specificity, and prevalence. However, in almost every practical situation the values for these parameters are not known exactly. Sensitivity and specificity are based on estimates from previous studies with a limited number of participants, and so are themselves subject to measurement error and therefore uncertain. Moreover, sensitivity and specificity may change depending on when the disease was contracted, on viral load, and other relevant factors.

Prevalence is often even more uncertain; it may be estimated with relatively high certainty at the level of an entire country, but often this is not the most appropriate level of analysis. For instance, in the case of Binarius, we know that they live in New York, that they just returned from a trip to Africa, and that they experience some symptoms typical for malaria. For Binarius, the relevant prevalence is not the base rate of malaria across the entire US population. Rather, what is relevant is the prior probability that Binarius has the disease, taking into account all of the background information. However, such background information does not determine a single correct value of 'prevalence'. Instead, it identifies a range of values that are relatively plausible.

To the untrained eye, it is not immediately evident how this uncertainty ought to affect our inference, as it needs to be propagated through our calcu-

lations of the desired quantities of interest. Typically, medical tests should be accompanied by an information leaflet that reports estimates of sensitivity and specificity with some quantification of (un)certainty, such as confidence intervals. In the classical paradigm, we can apply different methods to propagate the uncertainty about sensitivity and specificity to calculate (for instance) the positive predictive value, while holding the value of prevalence fixed (Mercaldo, Lau, & Zhou, 2007; Zou, 2004).

An arguably more general and intuitive approach to propagate uncertainty is to use Bayesian inference. From the Bayesian perspective, the parameters (i.e., sensitivity, specificity, and prevalence) are considered unknown to begin with (Baron, 1994; Crawford et al., 2009; Mossman & Berger, 2001). Moreover, the Bayesian approach provides a natural way to update beliefs, that is, to combine prior uncertainty with available data in order to quantify posterior uncertainty. The starting point of a Bayesian analysis is to assign prior distributions to the parameters; these distributions reflect the relative plausibility for their values, before having seen the data. All three parameters range from 0 to 1, and this means that the family of *beta distributions* provides a natural and flexible set of candidate priors (cf. Albert, 2009; Bolstad, 2007; Wagenmakers & Matzke, 2023).

The beta distribution for an unknown proportion θ features two parameters, α and β, and the prior assignment is usually denoted $\theta \sim \text{beta}(\alpha, \beta)$. The mean of the beta distribution is $\alpha/(\alpha + \beta)$; therefore, when α is larger than β, the mean for θ is larger than $1/2$. Also, when α and β both grow large, the distribution becomes increasingly peaked around its mean, indicating more certainty about what values for θ are plausible. A $\text{beta}(1, 1)$ distribution is uniform from 0 to 1, and a beta $(100, 100)$ distribution is relatively peaked around $\theta = 1/2$. One may heuristically interpret the values for α and β as the number of hypothetical previously seen successes and failures, respectively.

For the analysis of medical tests, Mossman and Berger (2001) recommend to assign prevalence, sensitivity, and specificity each a so-called Jeffreys prior (Jeffreys, 1961; Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017), a beta distribution with both α and β equal to one half: $\mathcal{B}(1/2, 1/2)$. This U-shaped prior assigns most probability mass on values near 0 and 1, with relatively less mass on values in between. Alternative priors are also possible, see

Tuyl, Gerlach, and Mengersen (2008).

The main advantage of assigning each of the three parameters a beta prior distribution is that incoming data (e.g., the number of true positives and false negatives obtained from a study) result in an update of knowledge that is highly convenient. Specifically, when a beta prior distribution is updated with a certain number of 'successes' and 'failures' (i.e., binomial data), the posterior distribution is also a beta distribution, but with updated shape parameters. This general property is known as conjugacy (Diaconis & Ylvisaker, 1979; Raiffa & Schlaifer, 1961). For instance, when a diagnostic test from a study yields a particular number of true positives (a), false negatives (b), true negatives (c), and false positives (d), the Jeffreys beta priors for sensitivity and specificity are updated to the following beta posterior distributions (Mossman & Berger, 2001):

$$\begin{aligned} \text{sensitivity} &\sim \mathcal{B}(1/2 + a, 1/2 + b) \\ \text{specificity} &\sim \mathcal{B}(1/2 + c, 1/2 + d). \end{aligned} \tag{9.3}$$

Optionally, if we consider the sample reported in the study representative of the population of interest, we may also update prevalence with the available data:

$$\text{prevalence} \sim \mathcal{B}(1/2 + a + b, 1/2 + c + d). \tag{9.4}$$

Alternatively, in the presence of strong background knowledge we may want to use informed priors for prevalence instead.

With the posterior probabilities in place it is then relatively simple to propagate the uncertainty, as one can draw random values from the posterior distributions and compute the desired quantity of interest (Mossman & Berger, 2001). Repeating this sampling process many times leads to approximately correct uncertainty distributions for any derived quantity of interest, and the error of approximation can be made arbitrarily small by increasing the number of samples that are drawn from the posterior distributions; typically a few thousand draws suffice for highly precise estimates.

Although this principled Bayesian approach works well in general (Mossman & Berger, 2001), it does overlook two subtle ideas that can be leveraged to our advantage. The first idea builds on the fact that it is often reasonable to assume that the performance of a diagnostic test is better than chance; in the

hypothetical case of a diagnostic test that performs reliably worse than chance, we can simply switch the labels that it predicts, thereby constructing a new test that performs better than chance.[3] The model can accommodate the assumption that a test performs better than chance by imposing the following constraint on the parameters:

$$\text{sensitivity} \geq 1 - \text{specificity}. \tag{9.5}$$

From a signal detection perspective, this constraint corresponds to restricting the distance between the 'noise' and 'signal' distribution to be positive: $d' \geq 0$ (Lee, 2008; Paulewicz & Blaut, 2020).

The second idea follows directly from adhering to the Bayesian adage that all available information ought to be taken into account. When we observe a positive (or negative) test outcome, that result itself provides information about the unknown parameters, despite the fact that we do not know whether or not the test result was actually correct (Winkler & Smith, 2004). For example, when Binarius tested positive, we know that they are either true positive or false positive. In the true positive case, prevalence should be updated by one positive case and sensitivity should be updated by one correctly classified case. In the false positive case, prevalence should be updated by one negative case and specificity should be updated by one incorrectly classified case. This results in posterior distributions that are essentially mixtures of the two possible options (Winkler & Smith, 2004).

Thus, when we wish to assess Binarius's probability that they have malaria given that the test was positive, we need to take into account that the test came out positive in the first place. We do that by first updating the parameters, and only then calculating the *positive predictive value* using Equation 9.1 (Winkler & Smith, 2004). Depending on the amount of certainty about the parameters before obtaining the positive test, this correction may or may not make a meaningful difference. When there is a a high degree of uncertainty to begin with, it may lead to a large enough difference to change the outcome of the medical diagnosis (Winkler & Smith, 2004).

Now that the prior distributions have been defined, the updating rules are

[3] In practice, if a test performs worse than chance, it would of course be wise to investigate the reason for the anomalous behavior.

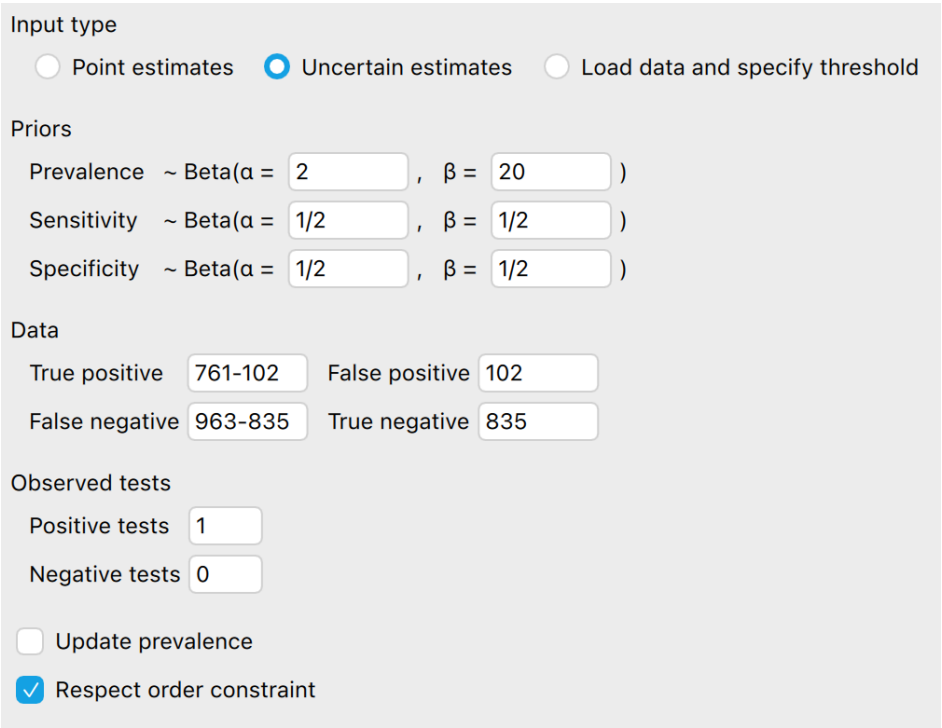


Figure 9.11: JASP screenshot of the input GUI for the `Binary Classification` module with parameters subject to uncertainty.

known, and the parameter restrictions have been implemented, we may revisit the case of Binarius and conduct a Bayesian analysis that acknowledges the uncertainty in the three key parameters.

We start with the Jeffreys $\mathcal{B}(1/2, 1/2)$ priors on sensitivity and specificity (Mossman & Berger, 2001). In the sample from Berzosa et al. (2018) there were 963 negative cases (as indicated by the gold standard PCR method) of which the RDT correctly identified 835 as negative. There were also 761 cases that tested positive using the RDT method, of which 102 were false positives. Thus, the data is comprised of $761 - 102 = 659$ true positives, 102 false positives, $963 - 835 = 128$ false negatives, and 835 true negatives. Since Binarius tested positive, we also add one observed positive test to the data.

As for prevalence, the study by Berzosa et al. (2018) was conducted in Equatorial Guinea, and the prevalence of the disease in that country has arguably limited relevance for students from NYU such as Binarius. Thus, we should

not update prevalence with the data from the study by Berzosa et al. (2018). However, Binarius did recently visit Central Africa, so the prevalence of malaria in the US is also of limited relevance. In addition, Binarius suffers from fatigue and headache. The prevalence estimate that would be relevant is for students at NYU who have recently visited Central Africa and afterwards experience fatigue and headache. Unfortunately, no prevalence information is available for this specific group, and a large range of values may appear plausible. We therefore assign prevalence a weakly informative prior: prevalence $\sim \mathcal{B}(2, 20)$. This prior is centered around a prevalence value of 0.1 that was used in the case of fixed parameters, but is associated with considerable uncertainty; 95% of the probability mass falls between ≈ 0.02 and ≈ 0.24. Lastly, we will impose the restriction that the test works better than chance.

The input panel of the JASP analysis is displayed in Figure 9.11. JASP uses the general-purpose Markov chain Monte Carlo program 'JAGS' (Plummer, 2003) to draw samples from the posterior distributions. Once JAGS has finished sampling, summary results are displayed in the output panel.

The main JASP output table (not shown) presents numerical information that summarizes the posterior uncertainty. Here we display the full posterior distributions of the quantities of interest; in JASP, these can be obtained by ticking the option `Estimates` in the `Plots` section, and then selecting `Prevalence`, `Sensitivity`, `Specificity`, and `Positive predictive value`. The resulting output is displayed in Figure 9.12.

For prevalence, the posterior mean equals 0.11 with a 95% central credible interval ranging from 0.02–0.27; for sensitivity, the posterior mean equals 0.84, 95%CI $= [0.81, 0.86]$, whereas the posterior mean for specificity equals 0.89, 95%CI $= [0.87, 0.91]$.

As for Binarius's probability that they have malaria, the posterior mean equals ≈ 0.43. However, the uncertainty associated with that point estimate is considerable: 95% of the posterior mass falls between 0.10 and 0.73. Thus, while the positive test outcome has noticeably increased the probability that Binarius contracted malaria (as can be appreciated by contrasting the posterior distribution for positive predictive value with the $\mathcal{B}(2, 20)$ prior distribution for prevalence[4]), we are still quite unsure about the probability that Binarius has

[4]This can be done in the `Binary Classification` module by selecting the option

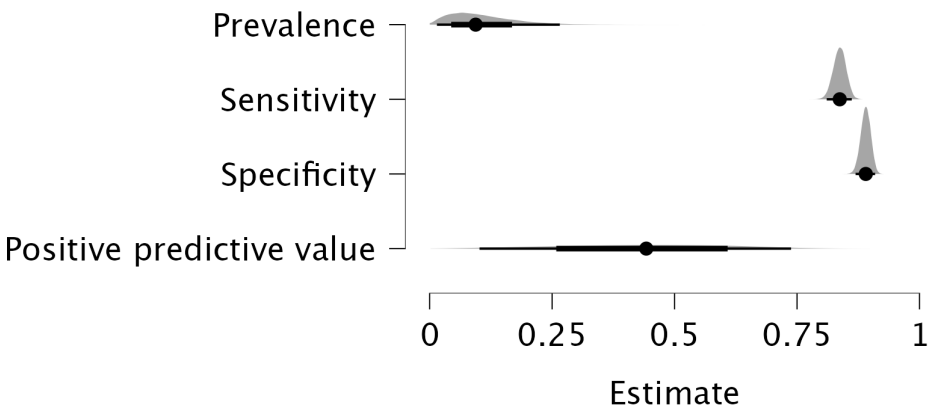


Figure 9.12: Posterior distributions of prevalence, sensitivity, specificity, and positive predictive value. The entire distributions are plotted as density estimates, with the black dot corresponding to the mean, the thick black line to the 67% and the thin black line to the 95% central credible interval, respectively. Output from JASP.

malaria.

In general terms, the data reported by Berzosa et al. (2018) has allowed sensitivity and specificity to be estimated with relative high certainty: the posterior distributions are relatively narrow. However, the $\mathcal{B}(2, 20)$ prior distribution for prevalence reflects considerable uncertainty, and this is not much reduced by the observation of Binarius's positive test result. This high uncertainty in prevalence combines with the uncertainty in sensitivity and specificity to yield an even greater uncertainty in positive predictive value.

To illustrate how uncertainty in the three parameters affects the conclusions, Figure 9.13 displays the same plots as Figure 9.10, but now with 95% credible intervals around the posterior means. Figure 9.13a shows estimates of sensitivity and specificity for different threshold values. The credible intervals are narrow, suggesting that the posterior uncertainty is small, thanks to the relatively large sample size of the study conducted by Berzosa et al. (2018) and the fact that sensitivity and specificity are directly estimated from the data: uncertainty in prevalence does not play a role.

`Probability positive` in the `Plots` section.

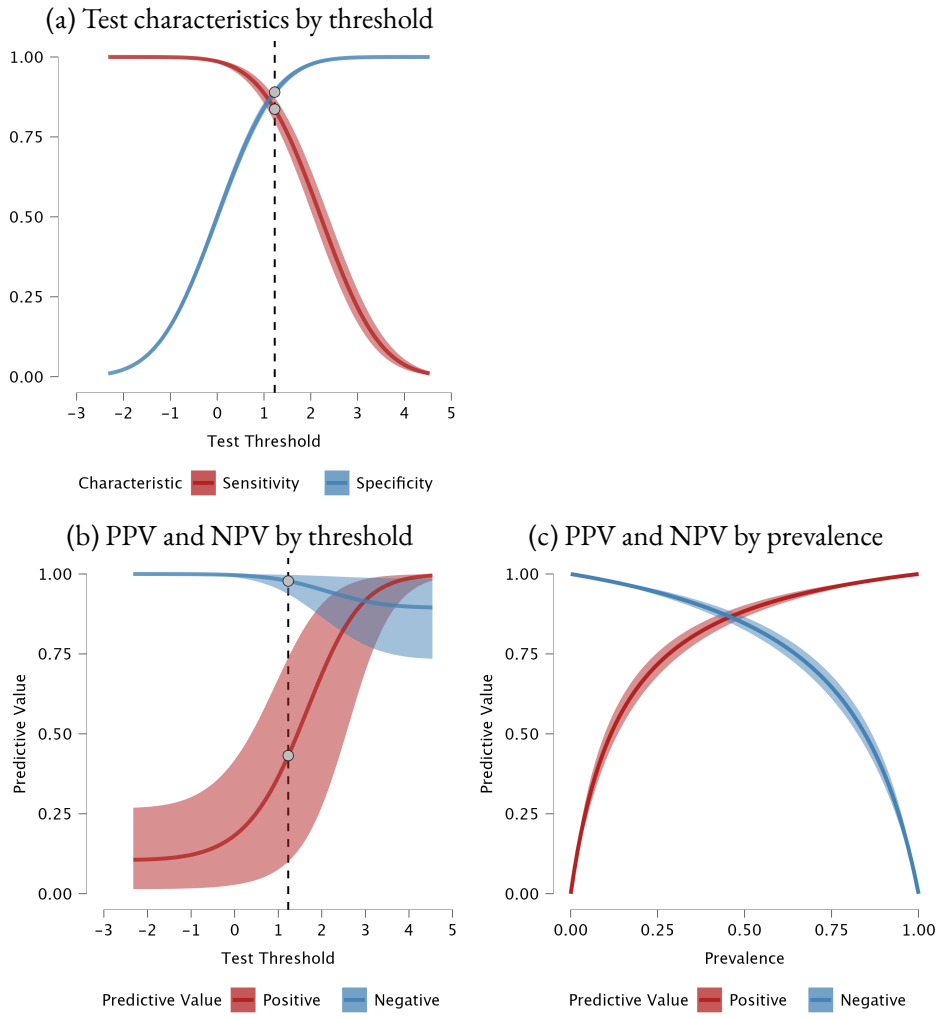


Figure 9.13: Trade-offs between two characteristics (y-axis) as a function of a third variable set to vary (x-axis). Plots on the left let threshold vary on the x-axis, whereas the plot on the right varies prevalence. The plot on the top shows sensitivity and specificity (i.e., test characteristics independent of prevalence), whereas the plots on the bottom show PPV and NPV (and thus take prevalence into account). Ribbons around the lines depict the 95% central credible intervals. Output from JASP.

Figure 9.13b shows PPV and NPV as a function of the threshold; the posterior uncertainty for these parameters is much larger, as indicated by the wide credible intervals. Sensitivity and specificity are relatively precisely estimated, and this means that the uncertainty in PPV and NPV is driven largely by the uncertainty in prevalence.

For a deeper appreciation of the effect of uncertainty in prevalence, Figure 9.13c shows posterior means of PPV and NPV and their 95% credible intervals for various fixed values of prevalence. The plot demonstrates that for any single value of prevalence, both PPV and NPV can be estimated precisely: their posterior distributions are relatively narrow. In other words, sensitivity and specificity are estimated relatively precisely, and consequently PPV and NPV can also be determined relatively precisely, just as long as the value of prevalence is known. However, as our knowledge of the base-rate is relatively imprecise, so is our final estimate of PPV.

The impact of prior uncertainty in prevalence can be further demonstrated by altering it. Imagine that NYU would have monitored the health of students who return from a study trip to Central Africa and show symptoms of malaria, and that NYU's records show that 50 out of 550 of these students had in fact contracted malaria. We could use that information as a prior for prevalence: $\mathcal{B}(50.5, 500.5)$. This greatly reduces our uncertainty in the base rate probability that Binarius has the disease, and as a result our uncertainty that Binarius has the disease after testing positive is also much reduced, as shown in Figure 9.14.

In sum, it may matter greatly whether or not the analysis assumes that prevalence, sensitivity, and specificity are known exactly. After quantifying and incorporating the uncertainty in the three parameters, the probability that Binarius has the disease is associated with a surprisingly large amount of uncertainty. This example highlights that one should consider all sources of uncertainty when drawing conclusions from a fallible medical test.

9.4 Concluding Comments

This manuscript presented a tutorial-style exposition on drawing correct conclusions from a fallible medical test. First, we discussed the standard problem in which parameters are assumed to be known exactly, and the conclusion is based

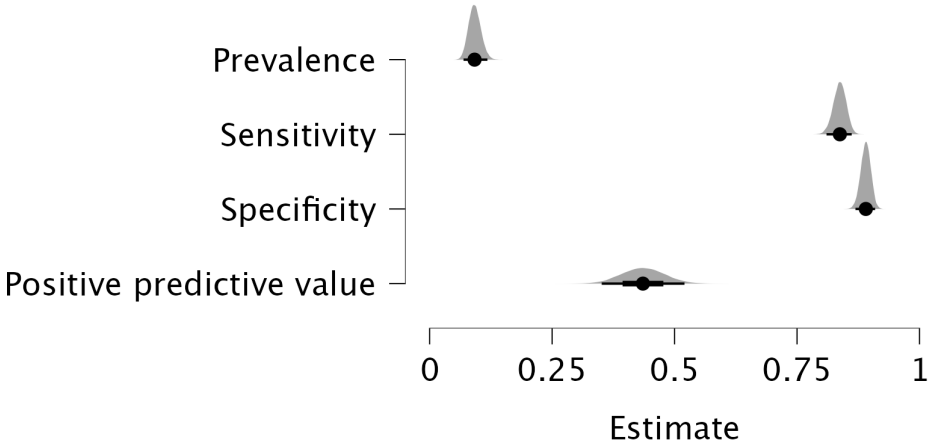


Figure 9.14: Posterior distributions of prevalence, sensitivity, specificity, and positive predictive value, after using an informed prior on prevalence. The entire distributions are plotted as density estimates, with the black dot corresponding to the mean, the thick black line to the 67% and the thin black line to the 95% central credible interval, respectively. Output from JASP.

on a a straightforward application of the rules of probability theory, namely Bayes' theorem. The relevant section presented a collection of partly overlapping approaches for explaining the solution from different perspectives, with the goal to strengthen the intuition without relying on mathematical expressions.

Although the standard problem is commonplace in introductory texts explaining Bayes' theorem, the solution is arguably not truly Bayesian; a fully Bayesian solution is to acknowledge and quantify all uncertainty, and particularly the uncertainty with respect to the values for the three key parameters sensitivity, specificity, and prevalence. The second part of this tutorial therefore focused on how to express and propagate this uncertainty, leaning on visual intuition rather than on mathematical derivations. By presenting the problem from different perspectives, we demonstrated how different sources of uncertainty interact, and how the largest sources of uncertainty may be identified.

The final example highlighted how the current uncertainty may be reduced by incorporating additional sources of background information. This hints at another feature of Bayesian reasoning, reflected in Dennis Lindley's adage

"Today's posterior is tomorrow's prior" (Lindley, 1972, p. 2): by adding new sources of information more and more uncertainty can be eliminated. This information can take different forms. Often, new information consist of recently collected observations or measurements, but it may just as well consist of relevant background knowledge that has become available at some point in time; in the latter case, the exact same data (e.g., a positive test result) may warrant a very different conclusion.

This way of reasoning should feel natural, as we rarely have at our disposal only a single piece of information when making a medical diagnosis; instead, we usually have access to the medical history of the patient, exposure to risk factors, other symptoms, and part of this knowledge becomes available sequentially, over the course of interacting with the patient. As more information becomes available, the medical professional will continually adjust and sharpen their opinion. From this perspective, the distinction between 'prevalence' ('base-rate', 'prior probability') and 'positive predictive value' ('posterior probability') becomes semantic, as both concepts express the probability of having the disease for the individual patient, given all information available at a certain point in time.

One of the main messages of this manuscript is that the `Binary Classification` module in JASP can assist in the proper interpretation of test outcomes, an important goal that has proven elusive using formal methods of instruction that rest on a pen-and-paper application of Bayes' rule. The module can act as a cognitive prosthesis, or as an inoculation against the pervasive bias to undervalue the impact of prevalence. Instead of carrying out the calculations by hand, one only needs to be able to input the relevant values and interpret the output. This aspect becomes even more important when uncertainty about the three parameters needs to be quantified and propagated to derived measures of interest, such as positive and negative predictive value.

Armed with JASP, medical professionals can prevent themselves from falling for the base-rate fallacy, and they can easily quantify the uncertainty in their inferences. Medical students can learn to reason about medical testing without having to learn the application of Bayes' theorem by heart: instead, a few mouse clicks yield all of the necessary information, either through a table or a figure.

The `Binary Classification` module is not only useful for students

and doctors, but also for patients: at-home medical testing is becoming more commonplace, including tests for infectious diseases, pregnancy, and more (Jean et al., 2022), and may become especially beneficial for vulnerable populations by increasing the accessibility of care (Hubach et al., 2021; Kattari et al., in press; Rasti, 2022). It is important that patients who take tests at home without the support of a medical professional are provided with information that allows them to draw correct conclusions. Ideally, tests that are sold in pharmacies for at-home testing should come with the required information to carry out the inference including assessing the uncertainty (Tidy, Shine, Oke, & Hayward, 2018). Thus, even the general population can benefit from a tool that allows a principled assessment of an outcome of a medical test, without falling for common fallacies.

For those interested in the use of the `Binary Classification` module in JASP, we prepared annotated .jasp files that are available at `osf.io/kue5h`. One of the files includes the running example presented in this manuscript. For those who are more interested in the methodological aspects, we also provide a .jasp file that contains the JAGS model as implemented in the `JAGS` module of JASP. These files can be easily downloaded, and subsequently opened and modified in the JASP application (`jasp-stats.org`).

Open Practices Statement

The JASP module used in this article is openly available as part of the JASP code base (`github.com/jasp-stats/jaspLearnBayes`). The annotated .jasp files are available at `osf.io/kue5h`.

Declarations

The authors declare their involvement in the open-source software package JASP (`jasp-stats.org`), a non-commercial, publicly funded effort to make Bayesian statistics accessible to a broader group of researchers and students. The authors have no financial or proprietary interests in any material discussed in this article. This article does not contain empirical data.

Quality without results is pointless. Results without quality is boring.

–Johan Cruyff

Chapter 10

Accessible and Sustainable Statistics with JASP

This chapter is preprinted as Wagenmakers, E.-J., Kucharský, Š., van den Bergh, D., and van Doorn, J. (2023). Accessible and Sustainable Statistics with JASP. *PsyArXiv*. doi: 10.31234/osf.io/ud2vj

Abstract

JASP is open-source software that aims to make statistical methodology available to a wide audience. JASP's user-friendly interface enables students, teachers, and researchers to conduct complex analyses in an instant, allowing the focus to lie on the interpretation of the results rather than on the underlying computer code. JASP includes both well-established and state-of-the-art methods and presently offers a full-fledged, modern replacement for commercial software such as SPSS or Minitab. JASP has been developed primarily at the University of Amsterdam and is now supported by a Community of several universities and colleges.

10.1 Introduction

In 2012 we initiated the development of the open-source statistics program JASP (jasp-stats.org) at the University of Amsterdam. The original goal of JASP was to make Bayesian statistics and frequentist statistics equally accessible: the aim therefore was to provide a Bayesian *addition* to SPSS, but open-source and friendlier to use. As more and more frequentist functionality was added to JASP, this objective shifted and it became increasingly realistic to aim for a wholesale *replacement* of SPSS. We have now arrived at this stage.

Currently, hundreds of universities worldwide use JASP to support their statistics education (see Figure 10.1). This is possible in part because JASP is available in several languages: English, Dutch, German, Spanish, Portuguese, French, Galician, Polish, Russian, Chinese, Japanese, and Indonesian.

It is impossible to discuss or even summarize all of JASP's functionality concisely – currently, JASP uses 475 different R packages (jasp-stats.org/r-package-list) spread out across seven standard analysis techniques (descriptive statistics, t-tests, ANOVA, mixed models, regression, frequencies, and factor analysis) and 25 modules that the user can activate and deactivate at will (e.g., for power analysis, meta-analysis, structural equation modeling, etc.). Below we present seven sample analyses, but first we would like to draw attention to some important features that distinguish JASP from most other statistics programs.

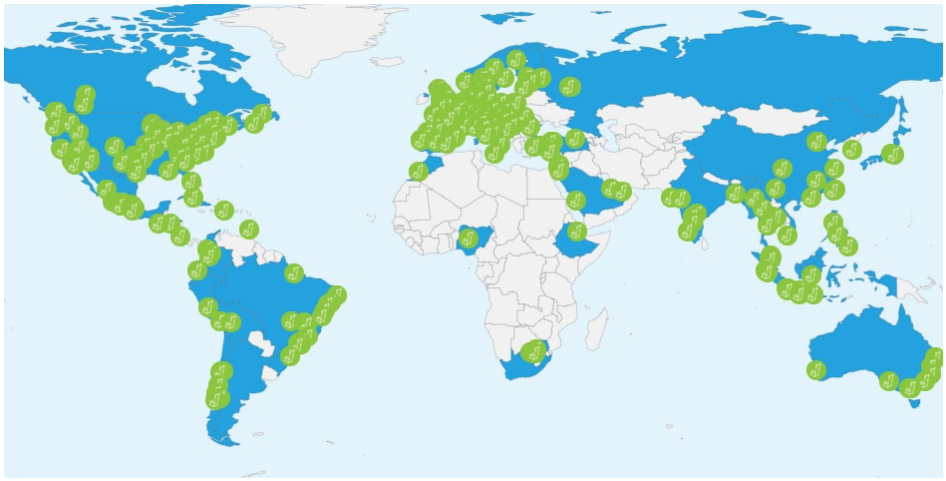


Figure 10.1: The 292 universities (in 67 countries) we know of that use JASP in their statistics teaching. JASP is currently downloaded about 75,000 times per month.

10.2 Characteristic Features of JASP

The development of JASP has been driven by desiderata of ease of use, accessibility, clarity, reproducibility, sustainability, and statistical inclusiveness. These desiderata are reflected in the following characteristic features:

- JASP has a **graphical user interface** (GUI). This means that the user selects options from a menu, just like in SPSS. Note that the JASP GUI is 100% reproducible – a saved **.jasp** file contains not only the data, the results, and any annotations, but also the ticked analysis options.

- JASP allows the user to store the **corresponding R code** for standard analyses. In JASP, the relationship between R code and GUI is a two-way street: modifications to the GUI change the R code, but modifications to the R code change the options checked in the GUI. In this way, JASP can be controlled both by R code and by the menu. We hope to achieve full integration with R this year.

- The JASP GUI provides **immediate feedback**. For instance, if the 'mean' option is checked in the input panel, then in the output panel

the mean is added directly to the corresponding table; if the 'mean' option is unchecked and the 'median' option is checked instead, then the effect of these actions is also immediately visible: the mean is replaced by the median. In this way, the student is invited to discover *interactively* what certain statistical procedures do.

- JASP is founded on the principle of **'progressive disclosure'**. Initially, the user is presented with relatively simple output – often a single table with the primary results. Additional information is within reach, but requires the user to select it first. This makes it clearer what the connection is between input and output; it also prevents the user from being overwhelmed with results and not seeing the forest for the trees.

- JASP provides extensive **help files** that are available by clicking the 'i' icon on top of each analysis. These help files document the input options, list the R packages used, and provide relevant references. Analyses aimed at statistics students are accompanied by an `introductory text` button that incorporates explanations of the analysis directly into the analysis output. This makes it easier for statistical novices to understand the main concepts behind the analysis. In the future, we plan to expand on this functionality.

- Older statistics programs are often cluttered – over the years they have gradually added more and more functionality without paying sufficient attention to organisation. In JASP, we strive for **clarity** by offering more complex functionality in separate tabs and in modules that need to be activated separately.

- In JASP, all output elements – figures, tables – can be provided with **custom user annotations**. This way, teachers can provide data-specific explanations, or students can ask questions. Annotations recently also allow LaTeX code for typesetting mathematical equations (see Figure 10.11 for an example) and embedding videos.

- JASP is **open-source**. This means that anyone can install and use JASP without any obligation, financial or otherwise. We take this opportunity

to denounce SPSS's revenue model.[1] The annual SPSS campus licences are priced *relatively* low so that universities continue to use the program for education, thereby making their students dependent on SPSS. Alumni will then advocate the use of SPSS at the companies and healthcare institutions at which they work; however, in these non-academic contexts a single SPSS licence quickly costs a small fortune – annually. An open-source fanatic might argue that universities that deploy SPSS are unwittingly complicit in wasting public money and creating an expensive lifelong addiction for their alumni. To be clear, the amount of money involved is astronomical. The SPSS website states that 80 per cent of US universities and colleges use SPSS. That amounts to about 4,000 institutions; if each institution pays $25,000 a year, then US academia alone transfers $100 million to IBM – every year. This situation becomes all the more glaring in the realization that SPSS is mainly used for trivial analyses (e.g. standard ANOVA and linear or logistic regression). For more complex analyses, data professionals use other programs, such as `R`, `Python`, or `Julia`.

- Transitioning to JASP from other GUI statistical software is easy. In our experience, anybody who is used to SPSS, Stata, SAS, Minitab, and alike finds JASP intuitive and does not require additional training. Given JASP's integrated `R` code support it is also easy to complement JASP with `R` if desired. JASP is **fully encapsulated**, meaning that the code and dependencies available in JASP remain fixed (within a specific version). Maintaining JASP on workstations is easier for ICT departments than maintaining different `R`, `Python`, or `Julia` versions and their packages. Thus, JASP is an ideal stepping stone for organizations that wish to transition away from GUI software towards scripting statistical pipelines with code, but are saddled with technical debt or shortage of expertise or resources.

- JASP does not use any tracking. We do not collect any user data whatsoever in the software itself. The data and analyses loaded in

[1] The issues raised here are common to all closed-source software that follows a pay-to-use business model – SPSS is not unique in this regard. Eventually, all commercial statistical software programs lead to a more or less severe vendor lock-in.

JASP are **completely confidential**. JASP does not contain any malware; see jasp-stats.org/2018/01/23/softpedia-review-award-jasp for an independent review. We may collect some of your data if you decide to share it with us, as outlined in our privacy policy at jasp-stats.org/privacy.

- JASP is **designed by universities, for universities**, and is therefore closely aligned with current teaching and research practices. A modest example is that the tables in JASP are formatted in 'APA' style, i.e. without vertical lines. Another example is that JASP has specific modules tailored to education, such as the 'Learn Bayes' module and the recently added 'Learn Stats' module. Furthermore, JASP has a built-in 'Data Library' with more than 50 example data sets based on popular course books and scientific articles. The library is also available online at johnnydoorn.github.io/DataLibraryBookdown. The JASP website contains a comprehensive listing of books, articles, YouTube videos, and other instructional materials on JASP (jasp-stats.org/resources).

- JASP provides traditional statistical procedures such as ANOVA and regression, but also includes **state-of-the-art analyses** that are useful both in academic research and in applied statistics. These analyses are mostly available in specific JASP modules (e.g. machine learning, quality control, network analysis, SEM, auditing, time series, etc.).

- JASP provides relatively many Bayesian methods, and so there may be a misconception that JASP is specifically Bayesian software. This impression is incorrect – JASP has a wide variety of **both Bayesian and frequentist methods**.

- Modules in JASP are essentially R packages with a couple of additional files that define the GUI. For R developers it is relatively easy to create a new module for JASP. In the future, we aim to establish an **Online Module Library** that will provide users the opportunity to upload and download user-contributed modules.

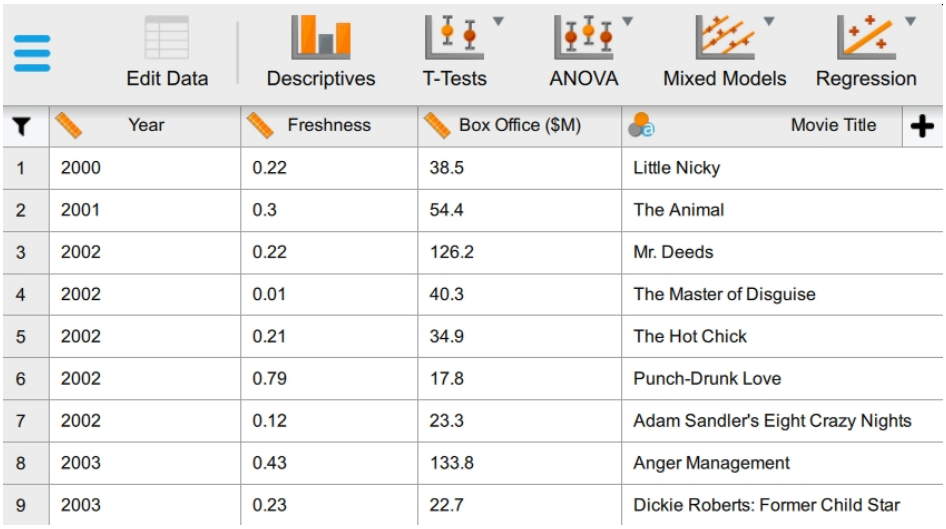
- In JASP, much attention has been paid to the **quality of figures**, in accordance with 'A Compendium of Clean Graphs in R', accessible at shinyapps.org/apps/RGraphCompendium. The aim is to provide figures that are elegant, insightful, and 'publication-ready'. The figures can be edited internally, but they can also be saved (e.g., as a pdf or in Powerpoint format) for further external editing.

- JASP is developed by a **dedicated team** of software developers, postdocs, and PhD students. This team is in direct contact with users through the JASP GitHub repository (github.com/jasp-stats/jasp-issues), which enables efficient processing of bug reports and feature requests.

- An important part of our effort is to adhere to **best practices in software development**. This means that we test our code (both manually and automatically), conduct code review, implement version control, and work with continuous integration pipelines. We also **validate the software** as it is being developed and maintained; that is, we compare our results to those produced by other statistical software in order to ensure that JASP yields the same results (or diverges for good reason); for more information, see jasp-stats.github.io/jasp-verification-project/.

- Recently we started the **JASP Community**, a consortium of institutions of higher learning that are joining forces to ensure that JASP can continue to be actively developed in the future. The current list of member institutions can be found at jasp-stats.org/cooperative-institutional-members.

This list of features is not exhaustive, and like all software, JASP can only be properly appreciated when it is applied in practice. To give a more complete impression of JASP, we therefore present seven example analyses on the following topics: (1) Descriptive Statistics, (2) the Bayesian t-test, (3) the Learn Stats module, (4) the Learn Bayes module, (5) the Distributions module, (6) Network analysis, and (7) Time series & Forecasting.



	Year	Freshness	Box Office ($M)	Movie Title
1	2000	0.22	38.5	Little Nicky
2	2001	0.3	54.4	The Animal
3	2002	0.22	126.2	Mr. Deeds
4	2002	0.01	40.3	The Master of Disguise
5	2002	0.21	34.9	The Hot Chick
6	2002	0.79	17.8	Punch-Drunk Love
7	2002	0.12	23.3	Adam Sandler's Eight Crazy Nights
8	2003	0.43	133.8	Anger Management
9	2003	0.23	22.7	Dickie Roberts: Former Child Star

Figure 10.2: Spreadsheet overview of the Adam Sandler data in JASP.

10.3 Example 1: Descriptive Statistics

At start-up, the user can choose to enter data (using an internal data editor) or open a data file. In this example, we open a data file from the JASP Data Library. From the main menu, we navigate via `Open` → `Data Library` → `4.Regression` to the 'Adam Sandler' data set. This data set contains a list of 31 movies from 2000 to 2015 starring Adam Sandler. After opening the data file, the data are displayed in a spreadsheet format, with one row per movie (see Figure 10.2). The relevant dependent variables are 'Freshness' (i.e., an estimate of the movie's quality on a scale of 0 to 1, provided by the movie website 'Rotten Tomatoes') and 'Box Office ($M)' (i.e., how many millions of US dollars the film earned at the box office).

For a first impression of the data, we choose `Descriptives` on the ribbon at the top. Part of the corresponding interface is shown in Figure 10.3. The user can use 'drag-and-drop' to make a selection among the available variables. Of the selected variables, a table is then immediately generated that includes the mean and standard deviation. The table can be modified at will.

We open the tab `Basic plots` and check the option `Correlation plots`. The result is shown in Figure 10.4. In this figure, the histograms show that the variables are unlikely to be normally distributed; the scatter plot sug-

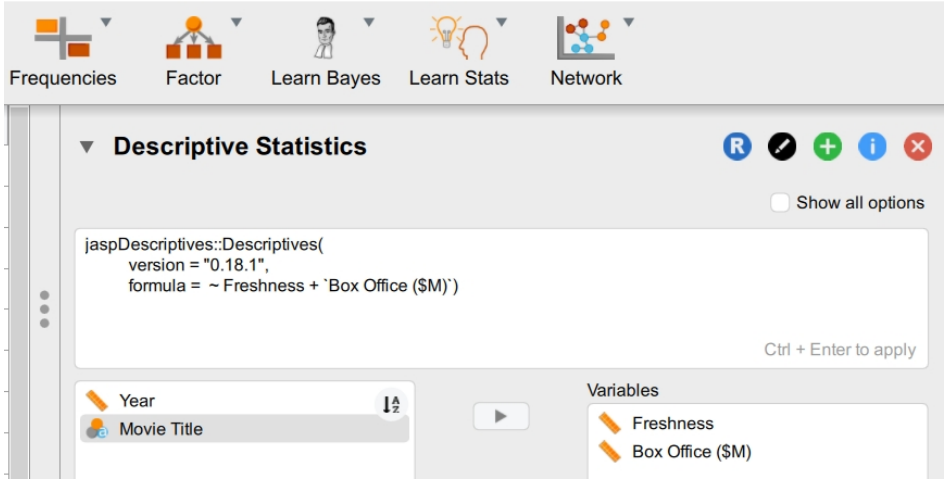


Figure 10.3: Drag-and-drop selection of the Adam Sandler variables 'Freshness' and 'Box Office ($M)' for the purpose of descriptive statistics.

gests that relatively good Adam Sandler movies *do not* attract more audiences than relatively bad Adam Sandler movies.

10.4 Example 2: A Bayesian T-test of the Facial Feedback Hypothesis

In 2016, we were involved in a 'Registered Replication Report' of the facial feedback hypothesis (Wagenmakers, Beek, et al., 2016). Briefly, the study examined whether participants found cartoons to be funnier when they clamped a pen between their teeth (i.e., with the facial muscles in the smile position) rather than between their lips (i.e., with the facial muscles in the pout position). The project involved 17 replication experiments with a combined total of 1894 participants; here we analyze only the data from the experiment conducted at the University of Amsterdam. After applying preregistered exclusion criteria, this experiment involved 130 participants who each scored four cartoons for funniness; 65 participants were assigned to the 'teeth' condition, and the other 65 were assigned to the 'lips' condition. Descriptive statistics show that the mean funniness rating was 4.94 in the teeth condition (SD = 1.14), and 4.79 in the lips condition (SD = 1.30). A frequentist independent t test gives a p-value of 0.48

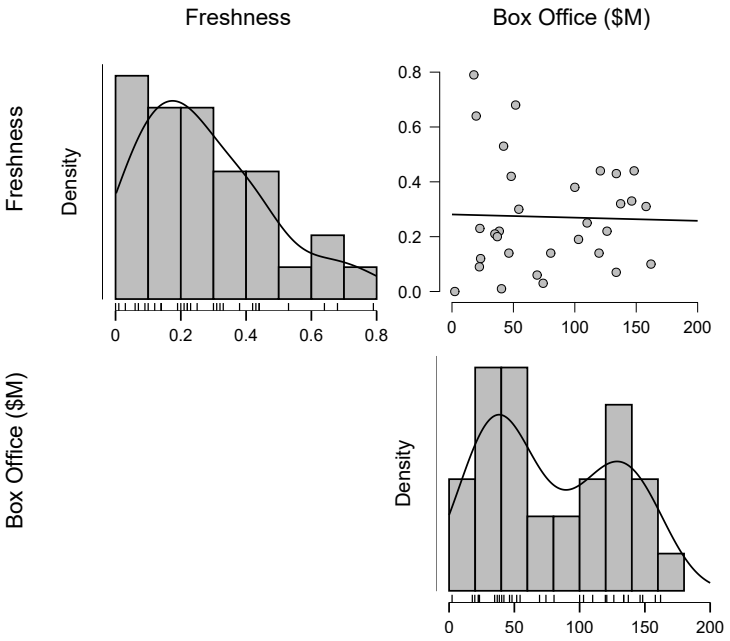


Figure 10.4: Correlation plot of the Adam Sandler data in JASP. The rightmost data point in the scatter plot is for the 2010 movie 'Grown Ups,' which grossed an impressive $162 million but was rated only 10 percent 'fresh'.

and a t-value lower than 1. This, of course, gives no reason to reject the null hypothesis. But to what extent do these data now provide support for the null hypothesis? In other words, is there absence of evidence, or is there evidence for absence? To determine this, we can perform a Bayesian t test in JASP.

Via `T-Tests` (on the ribbon at the top) → `Bayesian` → `Independent Samples T-Tests` we activate the Bayesian t-test interface. Using drag-and-drop, we select the relevant variables (i.e., 'condition' and 'meanCartoonRating'). Ticking the 'Raincloud plots' option results in Figure 10.5.

Next, we perform the Bayesian test. This requires a specification of the population effect sizes δ deemed plausible under the alternative hypothesis $\mathcal{H}_1$. We can simply use the standard options (i.e., the *default* prior distributions for hypothesis testing), but it is more interesting to think a little deeper about the available background knowledge. After all, this is a replication experiment, and it is therefore defensible to let the prior distribution for effect size be de-

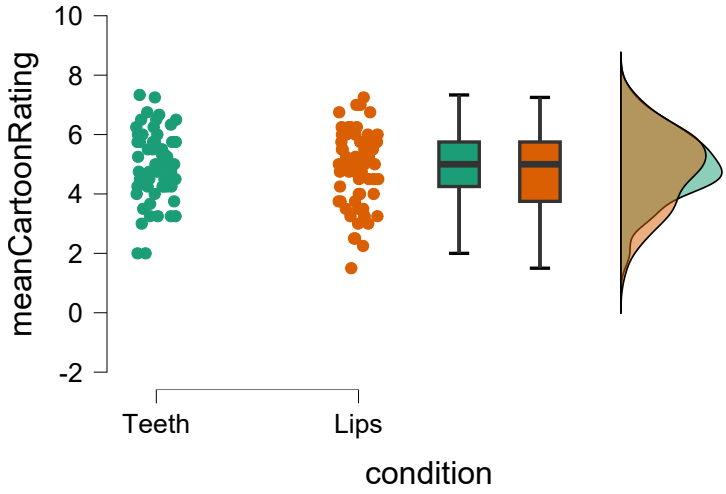


Figure 10.5: Average "funniness ratings" of four cartoons, separately for 65 participants holding a pen between their teeth ("Teeth") and 65 participants holding a pen between their lips ("Lips"). From left to right: a rain cloud plot (with jitter), box plots, and non-parametric density estimators.

termined by the posterior distribution of the original experiment (Verhagen & Wagenmakers, 2014). That posterior distribution is approximately normally distributed with mean of 0.4 and a standard deviation of 0.25.

In the GUI, we open the tab `Prior` and define the desired normal distribution. Finally, we indicate that the alternative hypothesis has a direction (i.e. cartoons would be funnier with a pen between the teeth, not less funny). Ticking the option `Plots` → `Prior and posterior` results in Figure 10.6. This figure provides a relatively complete overview of the Bayesian inference. In particular, we see that the *Bayes factor*, BF_{0+}, is equal to 2.035. This means that the null hypothesis (i.e., $\mathcal{H}_0 : \delta = 0$) predicted the observed data about two times better than the alternative hypothesis (i.e., $\mathcal{H}_+ : \delta \sim N(0.4, 0.25^2)I(0, \infty)$). Such weight of evidence would increase the prior probability for $\mathcal{H}_0$ from 0.50 to $2.035/3.035 \approx 0.67$ – according to Jeffreys (1961, Appendix B), this is not worth more than a bare mention. Thus, for comparing these two specific hypotheses, there is 'absence of evidence' rather than 'evidence for absence'.

Producing such analyses in JASP takes a few seconds – often ticking one or two options is enough.

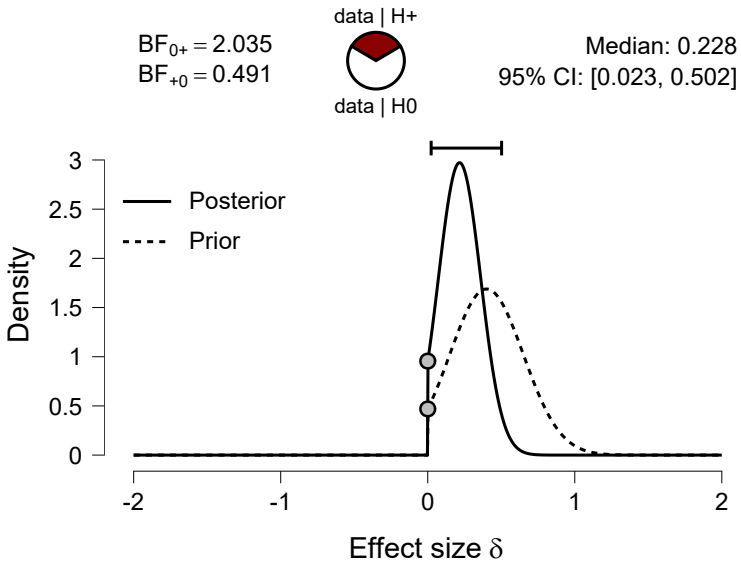


Figure 10.6: Prior and posterior distributions of effect size for one of the replication experiments reported in Wagenmakers, Beek, et al. (2016)

10.5 Example 3: The Learn Stats Module

Recently we added a 'Learn Stats' module to JASP. This module contains a coherent set of demonstrations that can help teachers explain key statistical concepts, namely: the normal distribution, the binomial distribution, the central limit theorem, the standard error, descriptive statistics, sampling variability, p-values, confidence intervals, and effect sizes. The Learn Stats module also has a statistical decision tree.

The underlying idea is that many key statistical concepts are associated with a visual representation. The demonstrations in the Learn Stats module try to bring those visual representations to the fore. A first example: after activating the Learn Stats module, we select `Confidence Intervals`. We set the `Confidence level` to 80 and the number of `Repetitions` to 50. The result is shown in Figure 10.7. The figure helps solidify the correct definition of an 80% confidence interval, namely that it is an interval generated by a procedure that encloses the true value in 80% of the experiments.

A second example: after activating the Learn Stats module, we choose `Effect Sizes`, and click on `Pearson correlation coefficient` ρ.

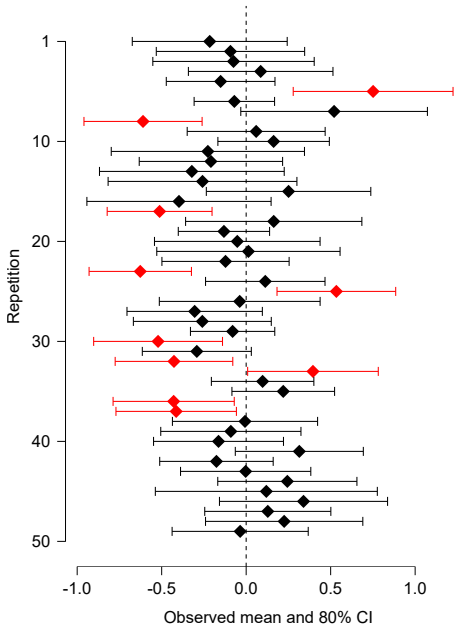


Figure 10.7: A visualisation from the Learn Stats module to clarify the concept of a confidence interval.

The GUI can then be used to specify a bivariate normal distribution; the corresponding figure shows this normal distribution with ellipses, possibly together with a sample of random size and the corresponding regression line (see Figure 10.8).

In the near future, we hope to add more and more education-specific material to JASP. Here, we believe, lies a great opportunity to make statistics education even more insightful and accessible.

10.6 Example 4: The Learn Bayes Module

JASP contains an extensive suite of Bayesian procedures. In order to support a better understanding of key Bayesian concepts we have developed a 'Learn Bayes' module that acts as a Bayesian extension of the 'Learn Stats' module.

The Learn Bayes module currently contains various straightforward demonstrations of Bayes' theorem, designed to assist teaching and foster a solid intuition of Bayesian inference (Ly, van den Bergh, Bartoš, & Wagenmakers,

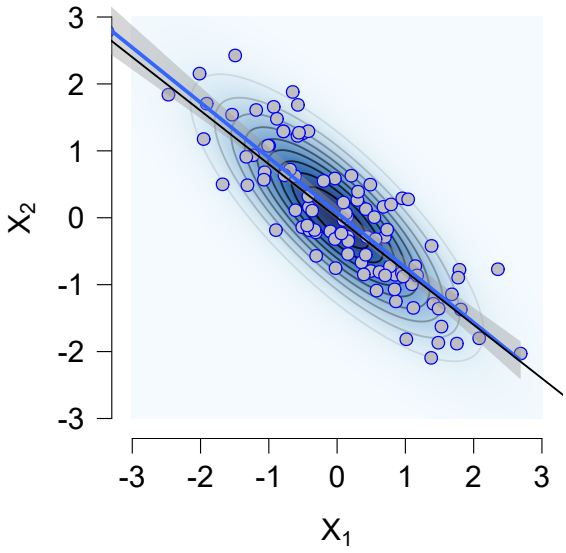


Figure 10.8: A visualisation from the Learn Stats module to clarify the concept of the Pearson correlation coefficient. The population coefficient here is -0.80; the ellipses describe the shape of the corresponding bivariate normal distribution; each of the 100 circles is a draw from the population distribution.

2021). Currently, these demonstrations feature `Binary classification` (i.e., applying Bayes' theorem in the context of diagnostic tests), `Binomial estimation` and `Binomial testing` (i.e., a detailed account of Bayesian inference for binomial data, used extensively in the free course book 'Bayesian inference from the ground up: The theory of common sense', Wagenmakers & Matzke, 2023), the `Problem of points` (i.e., the classic gambling problem that gave birth to the field of probability and statistics) and `Buffon's needle` (i.e., a geometric demonstration of estimating π by throwing needles on a planked floor).

For a glimpse what Learn Bayes has to offer we focus on the `Binary classification` submodule. The analysis of interest centers on the question 'given that a person tests positive (or negative) for a particular disease, what is the probability that they have that disease?' Crucially, the answer depends not only on the characteristics of the test (i.e., *sensitivity*[2] and *specificity*[3]; both set

[2] The probability that the test correctly detects the presence of the disease.
[3] The probability that the test correctly claims the absence of the disease.

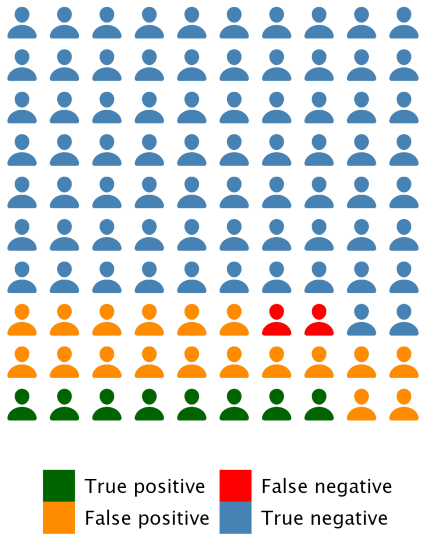


Figure 10.9: In the `Binary classification` tool from the Learn Bayes module, the icon plot shows why prevalence matters when interpreting the results from a medical test.

by default to 80%), but also on the *prevalence* (i.e., the base-rate probability of having the disease, before obtaining the positive test; set by default to 10%). For instance, with the default settings for sensitivity, specificity, and prevalence, a person who tests positive has a 31% probability of actually having the disease, considerably lower than may be expected based on the performance of the test alone.

The main philosophy of the `Binary classification` submodule is to build an intuition for the correct answer by providing informative visualizations rather than mathematical derivations. For instance, the surprising results are often explained using an icon plot. Ticking the `Icon plot` under the `Plots` section yields Figure 10.9. By default, 10 out of 100 people have the disease (i.e., the prevalence equals 10%). With a sensitivity of 80%, 8 of those 10 people would correctly test positive (i.e., the 8 'True positive' cases in green) whereas 2 would falsely test negative (i.e., the 2 'False negative' cases in red). With a specificity of 80%, $0.8 \times 90 = 18$ people who do not have the disease would falsely test positive (i.e., the 'False positive' cases in amber). Thus, among

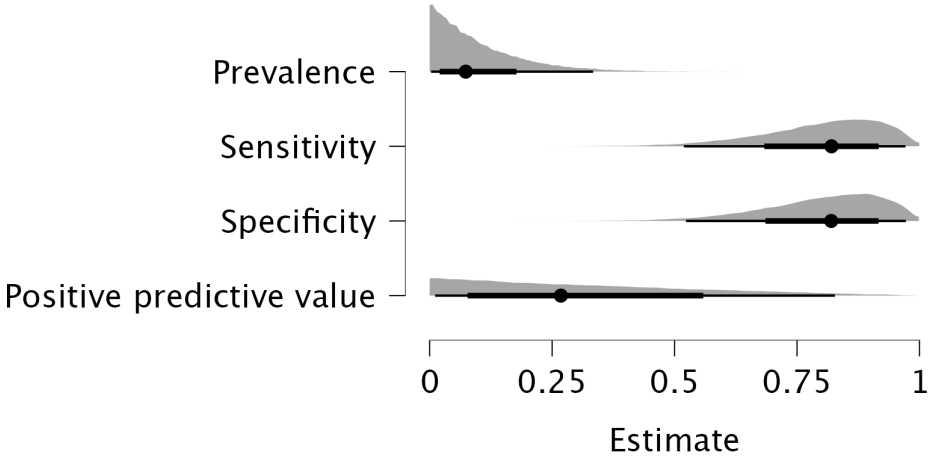


Figure 10.10: The `Binary classification` tool can propagate uncertainty about sensitivity, specificity, and prevalence to uncertainty about positive predictive value, providing a fully Bayesian account of the classic medical example that is commonly used to demonstrate Bayes' theorem.

the $8 + 18$ people who tested positive (i.e., the green and amber icons), only 8 actually have the disease, for a probability of $\approx 31\%$ (i.e., the *positive predictive value*).

In addition to the icon plot, the `Binary classification` tool also offers a large collection of other visualisations, including an ROC plot, an alluvial plot, and signal detection plots. Moreover, it is also possible to quantify uncertainty in sensitivity, specificity, and prevalence by assigning prior distributions to these quantities; the associated uncertainty then propagates to the key outcome of interest, that is, the probability of having the disease. Specifically, the `Uncertain estimates` option provides the opportunity to assign beta prior distributions to sensitivity, specificity, and prevalence. By default, these beta priors are chosen such that the distribution means correspond to the default point values illustrated in the previous example. The prior distributions can be updated by the data; in addition, the uncertainty reflected in the beta distributions propagates to all derived measures. For example, ticking the `Estimates` checkbox under the `Plots` section, and selecting `prevalence`, `sensitivity`, `specificity`, and `positive predictive value` yields Figure 10.10. Now instead of point values, each quantity is represented by an

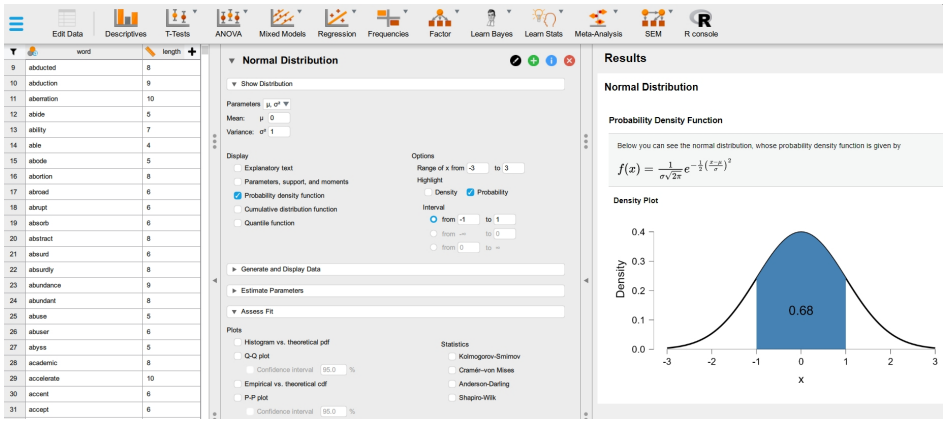


Figure 10.11: A screenshot of the Distributions module with a normal distribution selected. The left panel shows part of the data set; the center panel shows the input options for the normal distribution; and the right panel shows the corresponding output.

entire distribution; in this case, it is evident that the point value for positive predictive value is associated with a relatively high level of uncertainty.

10.7 Example 5: Distributions

Another natural extension of the Learn Stats module is the 'Distributions' module. As the name suggests, this module contains various probability distributions, ranging from the common (e.g., the normal and the binomial) to the exotic (e.g., the Amoroso and the zero-inflated negative binomial). The Distributions module now contains 31 continuous distributions and 9 discrete distributions.

The main purpose of this module is to familiarize students with different probability distributions and their properties. The input panel allows students to set the parameters of the distributions, and the output panel displays the corresponding probability density function, cumulative distribution function, or quantile function. In addition, the input panel allows students to query the distributions in order to highlight particular values for probability density or probability mass. An example for the normal distribution is displayed in Figure 10.11. This functionality is also helpful for deciding what distribution to

use as a prior for a parameter in a Bayesian analysis.

A secondary purpose of the Distributions module is to offer students the option to work with actual *samples* from the selected distribution. Synthetic samples may be generated from a particular distribution at will; these samples may then be fit using any distribution, and the quality of the fit can be assessed using a range of plots and statistics. For instance, we may generate data from a log-normal distribution, and fit these data using a normal distribution; for large enough sample sizes, the misfit should be apparent both from an inspection of the figures and from an interpretation of the statistics.

10.8 Example 6: A Network Analysis

Network analysis is used to discover and quantify the unique relationships between variables. The importance of this endeavor can be highlighted with a classic example: the relationship between ice cream sales and the number of drownings. These two variables are highly correlated; when ice cream is in high demand, more people drown. Of course, the statistical association between these two variables is entirely due to a third variable: temperature. This kind of 'spurious' relationship can be revealed by calculating a network with partial correlations. Such a network merely shows relationships between variables that cannot be explained by the remaining variables in the data set.

The example below concerns data from 2287 eighth-graders (publicly available in the `R` package `MASS`; Bosker & Snijders, 2011). These children took a language test (Language) and an IQ test (IQ); furthermore, the family's socioeconomic status was quantified (SES), class size was tracked (Class Size), and researchers recorded whether or not they were combination classes (i.e. a class with both children in grade 7 and grade 8; Combination 7-8). A statistical problem in estimating the partial correlations between these variables is that it is *a priori* uncertain which network structure matters. How likely is a network in which only IQ and SES are connected? How much better (or worse) does a network perform when the relationship between SES and Language is added?

Bayesian network analysis in JASP (Huth et al., in press) explores the space of *all possible network models* in a stochastic way. Network models that predicted the data relatively well enjoy a boost in probability, whereas network

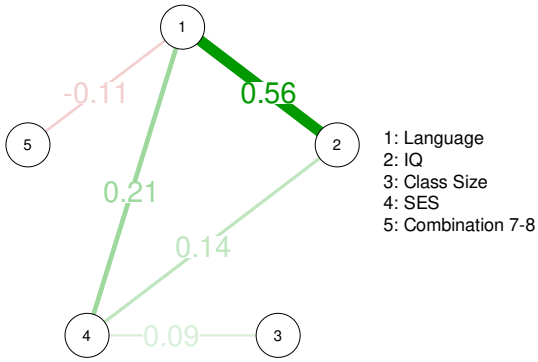


Figure 10.12: The strength of relationships between variables in a network structure. The weights were determined by considering all possible networks simultaneously.

models that predicted the data relatively poorly suffer a decline. The final conclusions are based on a weighted average of all network models, where the weights are determined by the models' posterior probabilities. The weighted conclusions for the sample data are summarized in Figure 10.12.

The analysis shows that IQ and Language are positively related. In addition, there are positive relationships between SES and both Language, IQ, and Class Size. Furthermore, there is a small negative relationship between Combination 7-8 and Language. This finding could motivate researchers to investigate whether being in a combination class negatively affects language development. Note that this analysis does not provide evidence for a causal relationship and therefore cannot be interpreted that way.

10.9 Example 7: Time Series & Forecasting

An important and exciting part of statistics concerns the analysis of *time series*: a temporal sequence of observations that may contain underlying structure and hence allows explanation and forecasting of future values. We recently extended JASP to include several time series analysis tools.

As an example, we analyze a time series of the number of visits to the JASP website, counted for each day starting from January 1st 2020 until October 8th

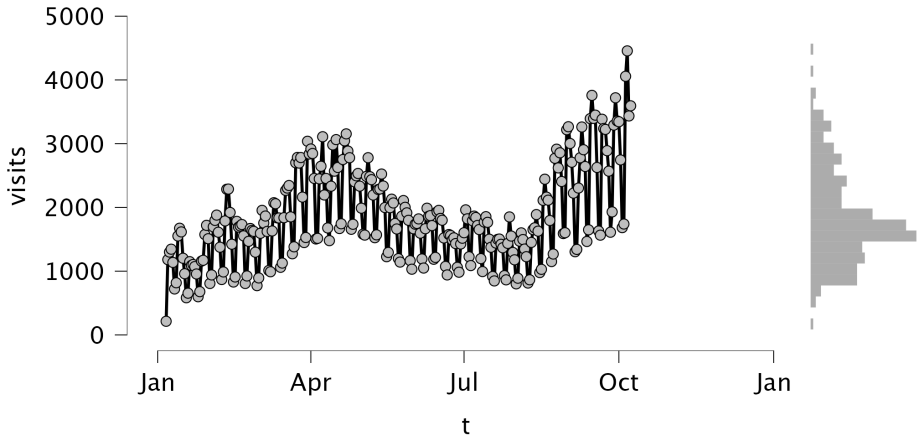


Figure 10.13: The number of visits to the JASP website counted for each day starting from January 1st 2020 until October 8th 2020. On the left, a line plot shows the number of visits on the y-axis plotted against the corresponding date. On the right, the marginal distribution of the number of visits is plotted as a histogram.

2020. The data set can be accessed via `Open` → `Data Library` → `15.Prophet` → `JASP Webpage Visits`.

To analyze the data, we will use the 'Time Series' module. This module provides several popular tools such as descriptives, stationarity checks, and spectral analysis. For the sake of brevity, we will turn straight to modeling the data using the autoregressive integrated moving average (ARIMA) framework. Open the analysis via `Time Series` → `ARIMA`. To initiate the analysis, we add the variable 'visits' as a `Dependent variable` and 'date' as `Time`.

To visualize the data, tick the `Time series plot` checkbox. To add a marginal distribution plot, select `Distribution` → `Density`. The resulting visualization is shown in Figure 10.13.

To analyze the time series data, we need to select the model-building strategy. Under the `Model` section, we tick the `Intercept` checkbox to estimate a general intercept. As it is likely that the number of visits depends on the day of the week (e.g., fewer people may visit the JASP website over the weekend), we will specify a weekly seasonality trend: tick the `Seasonality component` option and select `Custom period` with a periodicity of 7 days. For the selec-

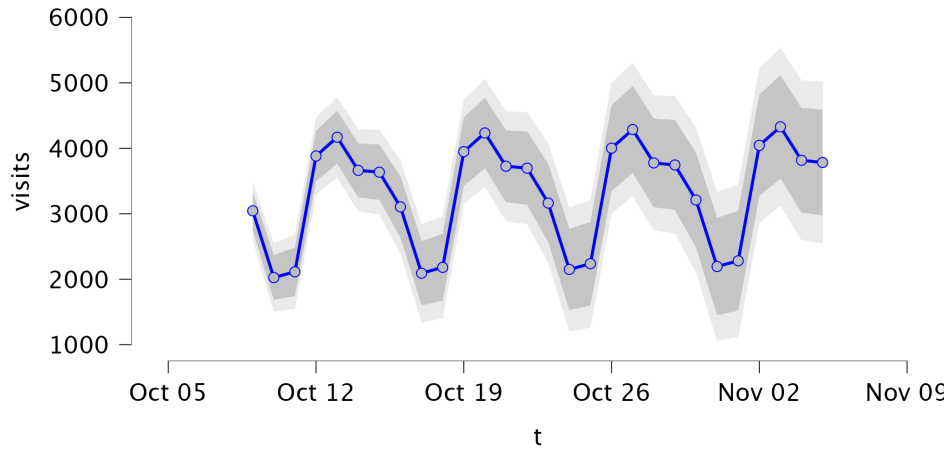


Figure 10.14: ARIMA predictions for the number of visits for the 28 days following October 8th 2020. The blue line represents the point predictions, dark grey ribbon represents the 80% prediction interval, and light grey ribbon represents the 95% prediction interval.

tion of the ARIMA terms, JASP automatically selects the best model according to the BIC criterion.

With the models fitted, JASP can generate predictions for the number of website visits beyond October 8th 2020. Under the `Forecasting` section, we specify `Number of forecasts` to 28 to compute the predictions for the next 28 days (four weeks). Select `Time series plot` to obtain the plot shown in Figure 10.14. It is also possible to obtain the forecasts in form of a table, or export them as a `.csv` file onto your computer.

From the forecast plot in Figure 10.14, it is evident that the weekly seasonality component is informative, as there is a clear dip in the number of website visits during weekends. However, the original website visits shown in Figure 10.13 appear to grow over time, which can be caused by a general upward trend, or perhaps monthly or yearly seasonality patterns. The traditional ARIMA framework does not take general trends into account explicitly, and so it is possible that the forecast will lead to increasingly inaccurate predictions the further ahead in time we predict.

A classical remedy would be to detrend the data before fitting the ARIMA model. Alternatively, JASP also provides state-of-the-art time series models de-

signed to handle data with complex trend and seasonality structures, estimating change points, and more: `BSTS` (Bayesian structural time series; Scott & Varian, 2014) and `Prophet` (Letham & Taylor, 2017). Interested readers can find the analysis of the JASP website visits using `Prophet` in the 'Data Library'.

10.10 Integration with R

Analyses in JASP run predominantly in the R programming language (R Core Team, 2020). For years, a closer integration between JASP and R was one of the most popular feature requests. The design and implementation of such integration is not an easy task, if only for the reason that the term 'R integration' may mean different things to different people. However, we are happy to report that through a lot of hard work on the part of the JASP software engineers, this major project is now nearing completion. Several crucial components are in place, with many more to come soon.

Below we first describe the R features currently available, and then foreshadow the R features that are still to come.

R console

JASP allows the execution of custom R code through its R console, which can be activated and deactivated as a module. The data set loaded in the JASP instance is automatically included in the environment's `data` object; if any filters were used in the JASP data viewer, the filtered data are available through the `filteredData` object.

To illustrate the use of the R console, we will replicate the time series analysis above using R code in JASP. Re-open the `JASP webpage visits` data set in JASP, and activate the R console window by navigating to the modules menu and selecting the console.

First, we adjust the environment of the R console by selecting 'Time series' in the dropdown button on the bottom right. This ensures that all packages associated with the 'Time series' module are available with the correct version. Next, we can execute the following commands:

```
library(forecast)
y <- ts(data$visits, frequency = 7)
fit <- auto.arima(
  y          = y,
  allowmean  = TRUE,
  seasonal   = TRUE,
  ic         = "bic"
)
predictions <- forecast(fit, 28)
```

The first line loads the `forecast` package (Hyndman et al., 2023; Hyndman & Khandakar, 2008) that implements common time series procedures. Next, we specify that the variable 'visits' is a time series with a periodicity of 7 days, and store that time series in an object called `y`. The next command runs the `auto.arima` function, which selects the ARIMA model that best fits the time series, and store the model in an object called `fit`. We specify `allowmean = TRUE` to allow an intercept, `seasonal = TRUE` to allow seasonal models, and `ic = "bic"` to specify that we wish to select the best model based on the BIC criterion (Schwarz, 1978). Lastly, we make predictions for the next 28 days by calling the `forecast` function and store the predictions in an object called `predictions`.

Figure 10.15 displays the R console next to the results from the 'Time series' module. The best model, estimated coefficients, and fit statistics all match the output in JASP – the only difference being that the JASP output has been organized in easy-to-read tables. Similarly, printing the `predictions` object returns the forecasts for the next 28 days that are plotted in Figure 10.14.

The R console is of course much more flexible, as users can run custom code independently from the JASP analysis. Therefore, the console is not only useful for reproducing results given by JASP, but for supplementing JASP results with additional analyses and calculations as well.

JASP syntax for R

The R console provides a convenient interface to R in JASP. One may wonder about the opposite: a convenient interface to JASP in R . The JASP team is

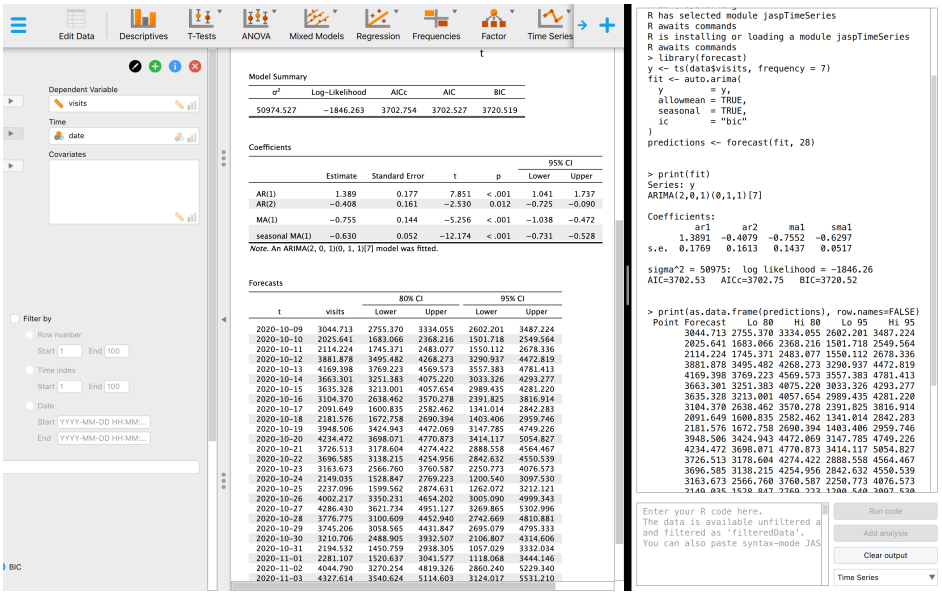


Figure 10.15: The R console (right panel) can run custom code in JASP independently from JASP analyses (left panel).

currently in the phase of implementing the 'R syntax' functionality that aims to do just that.

In principle, JASP modules are R packages. The R syntax project aims to provide convenient wrapper functions for JASP analyses that can be executed as functions in R which will return nicely formatted output in the R session. The analysis wrappers will be accompanied by standard help files that R programmers are accustomed to. In essence, R programmers will be able to run JASP analyses completely independently of whether or not they have the JASP application installed on their computer, or even without ever working with the JASP GUI in the first place.

Another exciting feature of the 'R syntax' in JASP is that JASP and R become interoperable. When specifying the analysis in the GUI, JASP can generate the corresponding R function call at the top of the analysis, as shown in Figure 10.3. Moreover, the R syntax within JASP GUI is a two-way street: changing an option in the GUI updates the syntax shown in the R window, but applying changes to the R syntax also updates the GUI. This provides an opportunity for statisticians to experiment with the analyses and learn how to produce valid R

syntax.

The use of the code generated by the JASP GUI is also twofold. We can execute it in the R console which will run the corresponding analysis in JASP, without having to specify anything in the GUI. For example, we can run the descriptive analysis of the Adam Sandler data as demonstrated in the first example of this article. First, reopen the Adam Sandler data set in JASP and activate the R console. In the R console, activate the 'Descriptives' environment in the dropdown menu at the bottom right of the console window. Paste the following code in the console:

```
jaspDescriptives::Descriptives(
  formula = ~Freshness+`Box Office ($M)`,
  correlationPlots = TRUE)
```

The `Add analysis` button activates when the R console recognizes the command as a valid call of a JASP analysis. A click on the button creates a new analysis in the JASP window, including the GUI and results. Thus, one can share JASP analyses without sharing the full `.jasp` files - the data and syntax is all one needs to reproduce an analysis in JASP.

The final step of the R syntax project is to enable running these JASP commands as functions in R , outside and independently of JASP GUI itself, perhaps as part of a larger analysis script in R . This is one of the high-priority features that the JASP team currently focuses on.

Developer tools

Currently, modules in JASP are predominantly written and maintained by JASP team members or close associates and colleagues. However, there is tremendous potential in new modules being created independently of the JASP team. As the R integration intensifies, this process becomes easier to facilitate; writing a JASP module is similar to writing an R package, with a couple of extra steps. Therefore, R programmers that implement novel statistical techniques in a custom R package are just a couple of steps away from making their methods available in JASP as well.

In the future, these external contributions will be supported by an 'Online Module Library', which will serve as a curated list of modules to which users

can contribute, as well as download and use in JASP. New modules will no longer be bundled with the main application, but will still be easy to activate at will.

To further facilitate the development of user contributed modules, we created several developer tools. First, the `jaspBase` package mainly serves R developers as a tool for communicating with the JASP application – for example, it makes it possible to create a structured results panel, consisting of different tables, plots, and more.

Second, the `jaspGraphs` package provides functionality that facilitates the graph style outlined in the 'Compendium of Clean Graphs in R' mentioned earlier (`shinyapps.org/apps/RGraphCompendium`) by presenting a custom JASP style theme building on the popular graphing R package `ggplot2` (Wickham, 2016). The package also provides general functions for generating common plots that can be reused inside of the JASP modules so as to keep the output consistent across different analyses.

Last but not least, the `jaspTools` package is a JASP analogue to the popular R package `devtools` (Wickham, Hester, Chang, & Bryan, 2022), and provides functionality that is not necessarily useful during the runtime of a JASP analysis, but makes it easier to create, develop, and maintain JASP modules themselves.

An important part of code maintenance is regular testing. For unit testing we rely on the `testthat` framework (Wickham, 2011). `jaspTools` provides functionality that allows creating and running these tests in R for continuous testing and integration. However, we also view testing as a form of results verification, which means that in addition to testing *consistency* we also strive to test the *correctness* of the results by comparing the results to popular statistics books. Thus, results verification has a value outside of code development as well, by means of demonstrating that the results produced by the software are correct – which is not only useful for code contributors, but for regular users as well.

To make the *correctness* checks transparent, our verification project (available at `jasp-stats.github.io/jasp-verification-project/`) publishes the results of these tests to make them available to users who do not necessarily want to look inside of the code base. However, the maintenance of the

current setup is cumbersome, as the results of the verification project are disconnected from the code in our continuous testing pipeline. With the full R integration, `jaspTools` will provide verification functionality that facilitates writing custom 'testing vignettes'. These vignettes will serve three use-cases: (1) they will be used directly in our continuous testing pipelines, making sure that changes to the code do not break the verified results, (2) they will be human readable so that they eventually replace the painstakingly compiled documents for our users, and (3) they will serve as documentation on how to use JASP analyses using the R syntax. By being directly integrated inside of the code base of a module, the 'vignettes' will be easy to maintain. We believe this will further push the boundaries of transparent and verified code.

10.11 Why not just use R?

The methodologically sophisticated reader may wonder what the added value of JASP is over R (or other programming software such as `Python` and `Julia`). It is certainly true that for some courses (especially in mathematics and computer science), a programming language is preferred. However, it should be noted that for many of our own analyses, we prefer JASP to R . The main reason is efficiency: the analyses in JASP are completed with a few mouse clicks, including figures and tables, whereas in R we first have to search diligently for which package also offers the desired functionality, what the arguments are for the relevant function, and how best to represent the results. The task further increases in complexity when we do not only care about computing the correct results, but also wish to produce clean and consistently formatted output ready for publication. For us, the same analyses that cost *seconds* in JASP may result in *hours* of coding if done with R .

For other courses, a two-stage model is possible: all students start with JASP, and those with a methodological interest switch to R after one or two years. In the future, we will expand JASP to facilitate this transition. In fields such as medicine, coding in R is not an obvious option, but it is still necessary for students to be able to understand the literature and possibly perform tests themselves. JASP is ideal for this group of users.

JASP is currently a full-fledged replacement for SPSS, and a useful com-

plement to R . We are optimistic that JASP can grow into a widely supported inter-university project in the future, with significant benefits for both teaching and research.

Acknowledgments

The authors have been closely involved in the development and further dissemination of JASP, and it is therefore inevitable that this contribution evokes memories of the immortal Dutch advertising slogan loosely translated as "We, the people at Toilet Duck™, recommend Toilet Duck™". For their own judgment, sceptical readers can download JASP for free from jasp-stats.org. We are grateful to all members of the JASP team for their insight and commitment over the years.

This article is based on a translation of an associated article written in Dutch by Wagenmakers, van Doorn, and van den Bergh (in press).

It's difficult because this is my life, every year. I have trees out there and I saw every one small like that and they are now massive. I will greet every one of them before I leave and say 'thank you'.

–Arsène Wenger

Conclusion

Since fall of 2017, 'Gazing into a discrete world' steadily grew one project after another. These projects ranged from methodological articles about processing raw eye-tracking data, through uncovering systematic tendencies in human eye movements, to analysing high level cognitive strategies.

As there is a considerable heterogeneity between the projects, it may be difficult to see the forest for the trees. However, some chapters have more in common with each other than with others, and so in line with the spirit of 'classification', they were grouped in three discrete parts of this book. Before we step back to enjoy the panoramic view of the forest, we will venture into these three smaller woodlands and greet every individual tree. After all, each chapter should hold up to scrutiny on its own and therefore deserves (some) unique attention.

Discrete Patterns of Behavior

Mixture modeling and classification is the center of this part of the thesis. Across the four chapters, we have applied mixture modeling concepts in different context. The essential question is how to best model discrete patterns of behavior, and how to use eye-tracking as a stream of data additional to the traditional sources of information used in cognitive research. Each individual chapter showcases how discrete patterns in behavior can be modeled and studied and how understanding discreteness of data can help us understand the underlying mechanisms behind phenomena of interest.

Chapter 1 introduces a novel approach to eye movement analysis, focusing

on descriptive and generative aspects. It poses the question: Is it possible to develop a model that not only discerns various eye movement events, such as 'fixations' and 'saccades', from raw eye-tracking data but also characterizes their typical features or even generates synthetic data closely mirroring real observations? The central challenge here is the latent nature of eye movement events, which are not directly observable and must be inferred from the data. Given the discrete nature of these events, a Hidden Markov Model (HMM) emerges as an apt choice. This model allows for the inference of transition patterns between different eye movement events from the raw data, treating these movements as hidden causes of observable patterns. The advantage of this methodology for eye movement classification is its ability to learn directly from the data it analyzes, thereby not depending on labeled datasets from other contexts, and eventually removing the necessity to rely on arbitrary thresholds. Although competitive with current classification algorithms, this model does display characteristic errors, such as confusing low-velocity 'smooth pursuits' with 'fixations', suggesting its status as more experimental than ready for immediate application. Yet, the model-based approach's ability to identify mismatches with data is invaluable, offering insights into eye movement characteristics and potential model enhancements. Furthermore, this work illuminates methods for validating existing algorithms, employing a three-step approach involving simulation, comparison with human coders, and assessment against existing algorithms. This study highlights the often-overlooked need for validation through simulation, a gap stemming from many algorithms' non-generative nature. The use of generative models, coupled with innovative experimental designs, could break free from the constraints of human-labeled data as the 'gold standard' (Hooge et al., 2018), paving the way for new benchmarks in evaluating eye movement classification algorithms.

Chapter 2 details the creation of a WALD-EM model tailored to analyze eye movements. This model interprets eye movements as sequences of distinct fixations, where each fixation is defined by a specific location and duration in the visual field. During these fixations, the eye remains mostly static, playing a crucial role in visual perception by enabling the brain to process visual information in segmented units. This approach aids in extracting key elements and forming an integrated mental image of the visual scene. The model's design is

flexible, allowing for adjustments and the integration of various elements that could predict or impact eye movement patterns. Such adaptability makes it an effective tool for investigating the intricate interactions between lower-level and higher-level cognitive processes that guide eye movement behavior.

Chapter 3 shifts focus from elementary explanations of eye movements to addressing the "inverse Yarbus problem" - the challenge of deducing an individual's cognitive processes or intentions from their eye movement patterns. Specifically, the chapter delves into the use of eye movements as a means to identify different strategies employed in cognitive reasoning tasks. It posits that varying strategies are likely to produce distinct eye movement patterns, and recognizing these unique patterns can reveal the underlying cognitive strategies, relying only on eye movement data. This perspective contrasts with prior research, which often presumed a link between diverse strategies and additional variables to start with, and using this presumed relationship to identify characteristic eye movement patterns that relate to these strategies (Hayes et al., 2011, 2015; Vigneau et al., 2006). The chapter argues that such assumption is not necessary as identifying distinct strategies can be effectively achieved through mixture models. Yet, it also demonstrates that by exploiting the logical structure of cognitive tasks, simpler methods like representing eye movements with transition matrices and classifying them into distinct groups using *k*-means clustering can also be effective. This method is not only straightforward to implement but also easy to interpret, offering a valuable supplemental tool for alternative analytical approaches.

Chapter 4 continues exploring discrete cognitive processes but shifts away from eye-tracking data for their identification. This chapter delves into the application of Hidden Markov Models (HMMs) in conjunction with evidence accumulation models (EAMs) for analysing data from speeded decision tasks. It challenges the conventional view of trade-off between speed and accuracy. Instead, it proposes that individuals might alternate between different cognitive states, yet allows for a continuous trade-off within those states. The chapter introduces an innovative model that combines HMMs with a simplified version of the Linear Ballistic Accumulation (LBA) model. This integration successfully addresses the usual computational challenges associated with EAMs while retaining the flexibility needed for continuous trade-off within states.

The complex computational nature of this model makes parameter estimation particularly demanding. To validate the model's computational efficacy, the chapter employs simulation-based calibration (SBC), a method that ensures the Markov Chain Monte Carlo (MCMC) technique used in the study correctly approximates posterior distributions. The effective application of this model confirms the utility of SBC as a robust tool for validating complex cognitive models.

Although these chapters extensively utilize mixture modeling methods, a key cautionary note is this: not every modeling situation necessitates the use of mixture models. These models are undoubtedly valuable tools, yet they should be used only when applicable - that is, when there is a reason to believe that the data are generated by underlying processes that are distinct from one another. The selection of topics in these chapters was intentional, chosen because they inherently suited mixture modeling ideas. This highlights the importance of selectively and thoughtfully applying mixture models where they are most appropriate, rather than defaulting to their use in all cases.

Addressing Imperfections

The second part of this book discusses work aimed towards improving research practices in developmental research, with a key emphasis on the habituation paradigm that is prevalent in this field.

Small samples may lack the statistical power and diversity needed for robust generalizations, while large samples may become unwieldy and resource-intensive. The challenge of striking the right balance can be difficult. Chapter 5 presented a tutorial on Bayesian sample size planning in the context of developmental research, where collecting large sample sizes is arguably more difficult than in other areas of psychological research. Bayesian sequential designs and Bayesian design analysis has the potential to overcome these limitations. In the case of sequential designs we can continuously evaluate the evidence provided by the data as it comes in, and stop data collection once we can draw strong enough conclusions. Bayesian design analysis can help researchers planning an appropriate study design to answer questions of interest, or evaluate whether it is feasible to collect required sample sizes in the first place. Both approaches

thus lead to saving resources and increase transparency of empirical research.

Individual studies, while informative, may only offer fragmented answers. A holistic understanding requires weaving together diverse strands of research to construct a more comprehensive theoretical accounts of empirical phenomena. In chapter 6, we focused on the study of infant habituation, and designed a collaborative systematic review and meta-analysis project from over 700 articles that use habituation paradigms. The aim of the project is to better understand current practices in habituation research as well as to try to disentangle which practices lead to more robust results. Unfortunately, the chapter only presents the first stage registered report, and as such presents no results so far — as of writing of this thesis, the project is still undergoing.

Understanding current practices in habituation research is an important topic, but solely analysing already existing practices may miss opportunities to change the practices entirely. Chapter 7 provided a critical overview of the current standards in habituation research and discussed potential alternatives. Key points included the use of modeling to understand individual differences in infants' habituation patterns, the possibility of multi-population (mixture) models to represent 'habituators' and 'non-habituators', and the flexibility of statistical models in representing various habituation processes. We emphasized the need for empirical validation of novel statistical models and suggest embracing a more collaborative approach in research to enhance the understanding of habituation processes.

It is crucial to recognize that the challenges highlighted in the three chapters on the habituation paradigm are not unique to this area alone but are pervasive across various research fields within and without of developmental psychology. The underlying message can be succinctly described by the somewhat cliché statement that "science is hard". Despite these complexities, advancements in data collection methodologies, study design, and data analysis techniques are progressively equipping researchers to address many of the prevalent issues in the field. Further, we now explicitly understand what we know and what we do not know. This work represents a step forward, yet it does not resolve or address all the challenges, and there remains a substantial need for continued collaborative efforts and innovations in future research endeavors.

Learning under Uncertainty

Uncertainty is an inherent part of every aspect of science. Learning how to deal with uncertainty is therefore an important aspect of any scientific endeavor. Bayesian reasoning provides a coherent description of rational reasoning process of dealing with uncertainty. Much of this work relied on Bayesian reasoning and Bayesian statistics. However, it can be argued that Bayesian methods are still underused in the scientific literature. A common barrier for a more widespread adoption of Bayesian methods is the relative difficulty of implementing these methods manually, and a lack of training and available methods ready to be used by researchers.

The focus of the third part of the thesis was on making statistical methods more accessible and sustainable through the use of JASP, a free and open-source software designed for both classical and Bayesian analyses. This work emphasized the importance of statistical literacy and the need for tools that are both user-friendly and robust. By utilizing JASP, we aim to bridge the gap between complex statistical theory and practical application, making advanced statistical techniques more approachable for researchers, students, and practitioners across various fields.

This part of the book collects the work dedicated to (1) developing novel Bayesian methods to analyze common designs, (2) providing educational material to make Bayesian reasoning more intuitive, and (3) implementing Bayesian solutions in user friendly statistical software packages.

Chapter 8 extended the arsenal of statistical tools by Bayesian inference for partial correlations. The work presented analytic derivations for the Bayes factor test as well as the posterior distribution of the partial correlation coefficient, and proved that the resulting inference satisfied several desiderata for optimal inference in situations where prior knowledge is limited or absent, which makes the approach appealing as the *default* choice for analysing partial correlations.

Chapter 9 is a tutorial aimed at medical students and professionals, as well as students of statistics, and discussed interpreting the results of a fallible medical test; a standard example for explaining the Bayes theorem using *prevalence*, *sensitivity*, and *specificity*. However, a less well known fact is that *prevalence*, *sensitivity*, and *specificity* are usually associated with considerable uncertainty. Bayesian reasoning dictates that such uncertainty is taken into account. There-

fore, we delved deeper than discussing Bayes' theorem by discussing how allowing uncertainty in these parameters affects our conclusions. The examples were accompanied with examples from the `Binary classification` module in JASP whose implementation was done as part of the present work as well.

Chapter 10 presented a high level overview of JASP, a user friendly statistical software that makes many of the novel Bayesian methods available to the practical researcher. We demonstrated how JASP can be effectively used for a wide range of statistical analyses, highlighting its intuitive interface, which allows for easy navigation and understanding of statistical concepts. The chapter explained components of the software as well as its underlying philosophy, stressing the value of open-source, transparency, and results verification. We also addressed the challenge of maintaining sustainability in statistical methods, advocating for open-source solutions like JASP that ensure ongoing development and adaptation to the evolving needs of the scientific community.

In summary, this part of the book described comprehensive additions and maintenance of JASP as a tool for accessible statistical analysis of data. This work underscores the importance of making statistical methods more user-friendly and widely available, thereby fostering a deeper understanding and broader application of statistics in various disciplines.

Overall picture

Bespoke vs. generic models

This thesis presented a seemingly dichotomous approach to statistical modeling, balancing between the development of bespoke models tailored for specific applications and the creation of generic statistical procedures suitable for a wide range of scenarios. At first glance, this dual approach might have appeared to reflect contradictory philosophies. However, upon closer examination, it became evident that this dichotomy was not a contradiction but rather a comprehensive strategy that leveraged the strengths of both bespoke and generic models, acknowledging their respective utility in different contexts.

Custom-built, or bespoke, models are designed with specific theoretical assumptions and empirical contexts in mind. They excel in addressing nuanced

and detailed aspects of a particular research question, offering deep insights that were closely aligned with the theoretical framework of the study. These models, by virtue of their specificity, provide a high degree of relevance and precision for the targeted research scenarios. A significant portion of this thesis was devoted to creating tailored models designed to address distinct empirical questions within particular contexts and research paradigms. The advantage of these 'bespoke' models lies in their incorporation of theoretical assumptions, which are often only implied in traditional methodologies, directly into the model's structure. This integration simplifies the interpretation of results, as the model's parameters are closely linked to the theoretical concepts of interest to researchers. This also allows leveraging expert judgement in setting informed priors, which can lead to improved inferences. Additionally, the use of generative models enables the verification of a model's capability to replicate observed data patterns. Looking ahead, these models offer the potential to simulate outcomes under different experimental conditions, paving the way for innovative research questions and experimental design exploration. On the other hand, implementing bespoke models may take considerable time, effort, and expertise. Further, it may not be possible to build such models when there is little theoretical understanding in the first place behind the process that generates the data

While custom-built models offer valuable specificity, they can sometimes become too narrowly focused, making adaptation to new contexts challenging. Additionally, their initial implementation might be difficult. A segment of this thesis was dedicated to developing models that, while somewhat context-dependent, can be more readily applied or adapted to various scenarios. For instance, the WALD-EM model in Chapter 2 is versatile, allowing for the addition, removal, or alteration of different factors influencing eye movements, making it adaptable beyond its initial application. This flexibility is even more pronounced in Chapter 3, which employs a straightforward two-step process: (1) categorizing groups of eye movement behaviors, and (2) associating these groups with other variables. This chapter could have focused on a joint model specifically designed for its primary application, like the Mastermind game. However, such a model would likely be tied to the theoretical assumptions of that particular game, potentially limiting its applicability to other tasks, like

the Matrix reasoning task. Therefore, the chapter adopted a methodology that strikes a balance between bespoke and generic statistical analyses. This approach aligns the data analysis more closely with the theoretical underpinnings of deriving cognitive strategies from eye-tracking data, while maintaining flexibility for specific applications. Due to its simplicity, it also presents a more accessible implementation option for researchers using familiar statistical software, as opposed to complex models requiring specialized tools.

On the other hand, generic statistical procedures were developed with versatility and broad applicability in mind. These models were designed to perform robustly across a variety of scenarios, especially in situations where specific models might not have been available or applicable. They served as valuable tools for researchers who needed reliable and accessible methods for a wide range of applications, particularly when bespoke models were not feasible or necessary. This thesis also emphasized the creation and implementation of these generic statistical procedures The emphasis on generic methods extends to the thesis's third part, which is dedicated to the implementation of these models.

A notable example of this approach is found in Chapter 8, which introduced a Bayesian method for analyzing partial correlations. This chapter not only demonstrated the practical application of the procedure but also provided theoretical validation, showcasing its suitability as a default Bayes factor in various research contexts.

Furthermore, many of these generic statistical methods are being developed for integration into JASP, as discussed in Chapter 10. JASP is designed to be user-friendly and accessible, promoting the adoption of Bayesian methods in the scientific community. By focusing on the general applicability and ease of use, this thesis contributes to the broader goal of encouraging widespread adoption of Bayesian methods in scientific research. This endeavor reflects a commitment to enhancing the statistical toolkit available to researchers, facilitating more rigorous and reliable scientific inquiry across various fields.

The thesis thus embraced a holistic view of statistical modeling, recognizing that both bespoke and generic models had their place in scientific research. This balanced approach underscored the importance of having a diverse toolkit of statistical methods, each suited to different types of research questions and contexts. By integrating both bespoke and generic models, this thesis contributed

to a more flexible and comprehensive approach to statistical analysis in various fields of scientific inquiry.

Statistical modeling as software development

The process of building statistical models, whether tailored for specific scenarios or designed for general application, plays a crucial role in scientific research. This task often involves the creation of custom code in various programming languages, drawing parallels with the principles of software development. However, a key distinction lies in the traditional perception of code in academic research versus that in software development. Academic code is typically crafted to function effectively for a single purpose—generating results for a specific study or publication. In contrast, software development prioritizes repeatability and reproducibility, ensuring that the software operates consistently across different systems and use cases.

This thesis advocates for a broader perspective on the role of academic code. It posits that statistical models are not merely instruments for generating academic results but are subjects of scientific investigation in their own right. This approach entails viewing the development of these models as more than just reaching an end result; it includes preparing them for scrutiny, adaptation, and utilization by other researchers. Such an approach necessitates rigorous validation of models through comprehensive simulation studies, practical applications to real-world datasets, and continual testing for code accuracy during development. Additionally, this perspective underscores the importance of publicly releasing the code, complete with thorough documentation, thereby enabling other researchers to access, assess, and adapt it for their respective purposes.

Taking this concept further, some of the methodologies developed in this thesis were integrated into JASP, a user-friendly, point-and-click software. This integration blurs the lines between statistical modeling and software development even more. It raises the bar for consistency and correctness in implementation, aligning more closely with the standards of professional software development. This shift signifies an evolving landscape in academic research, where statistical modeling is increasingly recognized and treated as a sophisticated form of software development, adhering to high standards of reliability,

usability, and accessibility.

Science as a collective endeavor

While the thesis covered a diverse array of topics, a recurring theme emerged: the increasing complexity and specialization required in empirical research. Modern statistical modeling demands highly skilled professionals, each bringing specialized expertise within the broader field of statistical analysis. This specialization extends beyond mere theoretical knowledge, encompassing technical proficiency in sophisticated statistical tools and methodologies.

Similarly, the design and execution of research studies call for a specialized set of skills. This specialization is not just theoretical but also practical, involving the intricate setup and management of laboratory environments and their equipment. Successful research design and data collection demand a nuanced understanding of theoretical concepts and the practical know-how to implement them effectively in a laboratory setting.

Moreover, drawing robust theoretical conclusions from the gathered evidence is another area where specific domain knowledge and experience become crucial. This aspect of research underscores the importance of deep subject-matter expertise to interpret data correctly and to draw meaningful inferences.

Finally, it is vital to recognize that the elements of research design, data collection, data analysis, and synthesis are intricately interconnected. The development of complex statistical models necessitates a deep understanding of their theoretical foundations. Similarly, the process of data collection may involve the real-time application of statistical models. Yet, requiring individual scientists to master all of the steps in the empirical research cycle becomes increasingly unrealistic under the assumption that the knowledge required to conduct each individual step continues to become more specialised. This interdependence highlights the necessity for more integrated and collaborative approaches in scientific research.

Given these complexities and the need for diverse skill sets, high-quality research increasingly relies on collaborative efforts. Building interdisciplinary teams that bring together researchers with varied expertise is becoming essential. These teams, by combining the strengths of scientists from different domains, can navigate the multifaceted challenges of modern research more ef-

fectively. Such collaborations facilitate the sharing of unique insights and techniques, leading to more comprehensive and well-rounded research outcomes. In this evolving research landscape, fostering strong collaborative networks is not just beneficial but may become a necessity for advancing scientific knowledge and achieving breakthroughs in various fields.

List of Contributions

Chapter 1: Characterising Eye Movement Events with an Unsupervised Hidden Markov Model

This chapter is published as Lüken, M., Kucharský, Š., and Visser, I. (2022). Characterising eye movement events with an unsupervised hidden Markov model. *Journal of Eye Movement Research*, *15*(1). doi: 10.16910/jemr.15.1.4
Malte Lüken and Šimon Kucharský share the first authorship.

Ingmar Visser and Šimon Kucharský provided the concept of this article. Malte Lüken implemented the model, conducted the simulation and validation study, and drafted the initial version of the manuscript. Šimon Kucharský checked the implementation for correctness, revised and finalized the manuscript. Šimon Kucharský and Ingmar Visser provided the final editing.

We would like to thank Daan van Renswoude, Maartje Raijmakers, and Maximilian Maier for their helpful comments on earlier versions of this paper. Furthermore, we would like to acknowledge feedback by Karel Veldkamp and Phil Norberts in the early stage of this project.

Chapter 2: WALD-EM: Wald Accumulation of Locations and Durations of Eye Movements

This chapter is published as Kucharský, Š., van Renswoude, D., Raijmakers, M., and Visser, I. (2021). WALD-EM: Wald accumulation for locations and durations of eye movements. *Psychological Review*, *128*(4), 667-689. doi:

10.1037/rev0000292

All authors contributed to the initial concept of the article. Šimon Kucharský developed, implemented and validated the model, and analyzed the data provided by Daan van Renswoude, while others provided feedback. Šimon Kucharský drafted the initial version of the manuscript. All authors contributed to final editing.

Chapter 3: Cognitive Strategies Revealed by Clustering Eye Movement Transitions

This chapter is published as Kucharský, Š., Visser, I., Truțescu, G.-O., Laurence, P. G., Zaharieva, M., and Raijmakers, M. E. (2020). Cognitive strategies revealed by clustering eye movement transitions. *Journal of Eye Movement Research*, *13*(1). doi: 10.16910/jemr.13.1.1

Maartje Raijmakers, Ingmar Visser, and Šimon Kucharský provided the original idea of this article. Šimon Kucharský conducted all analyses and drafted the initial version of the manuscript. Martina Zaharieva verified that the associated code is correct and reproducible. Gabriela-Olivia Truțescu collected and cleaned the Mastermind data. Paulo G. Laurence cleaned and processed data for the Progressive Matrices example and provided feedback to the manuscript and analyses. All authors contributed to final editing.

Chapter 4: Hidden Markov Models of Evidence Accumulation in Speeded Decision Tasks

This chapter is published as Kucharský, Š., Tran, N.-H., Veldkamp, K., Raijmakers, M., and Visser, I. (2021). Hidden Markov models of evidence accumulation in speeded decision tasks. *Computational Brain & Behavior*, *4*, 416–441. doi: 10.1007/s42113-021-00115-0

Ingmar Visser provided the concept of the article. Karel Veldkamp, Šimon Kucharský, and Ingmar Visser conducted initial feasibility study that provided insights into the issues associated with this topic. Šimon Kucharský and N.-Han Tran developed the model presented in this article and drafted the

initial manuscript. Šimon Kucharský implemented the model, conducted the simulation study and analysed the data. N.-Han Tran checked the correctness and reproducibility of the code. All authors contributed to the final version of the manuscript.

Chapter 5: Bayesian Sample Size Planning for Developmental Studies

This chapter is published as Visser, I., Kucharský, Š., Levelt, C., Stefan, A. M., Wagenmakers, E.-J., and Oakes, L. (2023). Bayesian sample size planning for developmental studies. *Infant and Child Development*, e2412. doi: 10.1002/icd.2412

Ingmar Visser and Šimon Kucharský share the first authorship.

Ingmar Visser and Šimon Kucharský provided the initial concept of the article. Lisa Oakes and Angelika M. Stefan were involved in further elaborating the concept of the article. Angelika M. Stefan and E.-J. Wagenmakers wrote the theoretical introduction to Bayesian sequential testing. Šimon Kucharský and Ingmar Visser wrote the practical part on planning and conducting sequential designs. Claartje Levelt provided the developmental example of rule learning. Šimon Kucharský conducted the analysis, Ingmar Visser and Angelika M. Stefan verified the correctness of the code. Lisa Oakes heavily edited the manuscript. All authors contributed to final editing.

Chapter 6: Habituation, Part I. Design Choices in the Infant Habituation Paradigm: A Preregistered Crowd-Sourced Systematic Review and Meta-Analysis

This chapter is a Stage I. Registered Report accepted at *Infant & Child Development* and preprinted as Zaharieva, M., Kucharský, Š., Colonnesi, C., Gu, T., Jo, S., Luttenbacher, I., … Visser, I. (2022). Habituation, Part I. Design choices

in the infant habituation paradigm: A pre-registered crowd-sourced systematic review and meta-analysis. *PsyArXiv*. doi: 10.31234/osf.io/bdtx9
Martina Zaharieva and Šimon Kucharský share the first authorship.

Ingmar Visser and Maartje Raijmakers provided the initial concept of this project. Šimon Kucharský and Martina Zaharieva coordinate this project. All authors were involved in drafting the manuscript. Šimon Kucharský and Martina Zaharieva finalized the manuscript.

Chapter 7: Habituation, Part II. Rethinking the Habituation Paradigm

This chapter is published as Kucharský, Š., Zaharieva, M., Raijmakers, M., and Visser, I. (2022). Habituation, Part II. Rethinking the habituation paradigm. *Infant and Child Development*, e2383. doi: 10.1002/icd.2383

Ingmar Visser and Maartje Raijmakers provided the concept of the article. Šimon Kucharský and Martina Zaharieva conducted initial feasibility investigation and identified the general challenges. Šimon Kucharský worked out the remedies presented in this article and drafted the article. All authors contributed to the final manuscript.

We would like to thank Yong-Qi Cong, Lisa Oakes, and Christina Bergmann for helpful insights into the current practices in habituation research and providing some habituation data sets. Julia Haaf, Angelika Stefan, N.-Han Tran, and E.-J. Wagenmakers for discussions about the statistical issues.

This article is dedicated in memoriam to Ágnes Hoffmann who worked on this topic previously.

Chapter 8: Analytic Posterior Distribution and Bayes Factor for Pearson Partial Correlations

This chapter is preprinted as Kucharský, Š., Wagenmakers, E.-J., van den Bergh, D., and Ly, A. (2023). Analytic posterior distribution and Bayes factor for Pearson partial correlations. *PsyArXiv*. doi: 10.31234/osf.io/6muwy

E.-J. Wagenmakers provided the initial concept of this article. Šimon Kucharský and Alexander Ly spent countless of hours working on the analytical solutions, while Don van den Bergh provided feedback. Šimon Kucharský authored the final proofs and Alexander Ly checked them for correctness. Šimon Kucharský drafted the initial version of the manuscript. All authors contributed to final editing.

We would like to thank Max Hinne, Maarten Marsman, Johnny van Doorn, and Fabian Dablander for early discussions on this project.

Chapter 9: Correct Conclusions from Fallible Medical Tests: A Tutorial with JASP

This chapter is published as Kucharský, Š., and Wagenmakers, E.-J. (2023). Correct conclustions from fallible medical tests: A tutorial with JASP. *OSF Preprints*. doi: 10.31219/osf.io/jksz6

E.-J. Wagenmakers provided the initial concept of this article. Šimon Kucharský implemented the functionality in JASP and drafted the initial version of the manuscript. E.-J. Wagenmakers revised the manuscript. Both authors contributed to final editing.

We would like to thank Emir Erhan for his contribution in the early stage of this project.

Chapter 10: Accessible and Sustainable Statistics with JASP

This chapter is preprinted as Wagenmakers, E.-J., Kucharský, Š., van den Bergh, D., and van Doorn, J. (2023). Accessible and Sustainable Statistics with JASP. *PsyArXiv*. doi: 10.31234/osf.io/ud2vj

E.-J. Wagenmakers provided the initial concept of the article. E.-J. Wagenmakers, Don van den Bergh, and Johnny van Doorn wrote the Dutch version of the article (Wagenmakers et al., in press). Šimon Kucharský translated the Dutch article into English using DeepL (`deepl.com`), and corrected the automatic translation. Šimon Kucharský expanded the introduction, wrote section

about integration with R and added examples on the Learn Bayes, Distributions, and Time Series modules. All authors contributed to final editing.

We are grateful to all members of the JASP team for their insight and commitment over the years.

Additional publications

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., … others (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*(1), 4–6.

Tran, N.-H., Kucharský, Š., Waring, T. M., Atmaca, S., and Beheim, B. A. (2021). Limited scope for group coordination in stylistic variations of kolam art. *Frontiers in Psychology*, *12*, 742577.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., … others (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*, 813–826.

van den Bergh, D., Van Doorn, J., Marsman, M., Draws, T., Van Kesteren, E.-J., Derks, K., … others (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année psychologique*, *120*(1), 73–96.

Wagenmakers, E.-J., and Kucharský, Š. (2020). *The JASP Data Library*. Amsterdam. Retrieved from https://jasp-stats.org/wp-content/uploads/2020/05/The_JASP_Data_Library_1st_Edition.pdf

A Word of Thanks

I would like to say thanks to my wife Han for providing never ending love and support, and for letting me be myself around you. I would be never able to do what I did without your help, and I would never get so much happiness in life without you.

A huge thanks to my family, especially my parents Anna and Milan, and my brother Jonáš. You always supported me, cheered me on, and encouraged me to pursue happiness and purpose even though it meant a lot of sacrifice and even though I had to leave you far away. Without you I would have never even thought about going abroad, yet here we are.

To my supervisors, Ingmar, Maartje, and E.-J. Thank you for helping me navigate my career through all these years, for providing advice, for encouragement, and for patience, but first and foremost for believing in me and giving me a chance. You were always there when I needed support and guidance, but gave me freedom to explore my interests. I learned so much from you and with you and I will never forget about that.

Last but not least, to my colleagues and friends who supported me throughout these years. A special shout-out to Martina from the Baby Lab and to all my team mates from JASP. It was always a pleasure working with you. I am privileged to be surrounded by such amazing and brilliant people. Thank you.

References

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., … others (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*(1), 4–6.

Addabbo, M., Longhi, E., Marchis, I., Tagliabue, P., & Turati, C. (2018). Dynamic facial expressions of emotions are discriminated at birth. *PloS One*, *13*(3), e0193868. doi: 10.1371/journal.pone.0193868

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River, N.J: Pearson Prentice Hall.

Ahmed, K., & Al Dhubaib, B. (2011). Zotero: A bibliographic assistant to researcher. *Journal of Pharmacology and Pharmacotherapeutics*, *2*(4), 303–305.

Albert, J. (2009). *Bayesian Computation with R* (2nd ed.). Dordrecht: Springer.

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: A reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 997–1010.

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, *311*(7003), 485. doi: 10.1136/bmj.311.7003.485

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307.

Anders, R., Alario, F. X., & van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychological Methods*, *21*(3), 309–327.

Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., & Nyström, M. (2017). One algorithm to rule them all? An evaluation and discussion of ten

eye movement event-detection algorithms. *Behavior Research Methods*, *49*(2), 616–637. doi: 10.3758/s13428-016-0738-9

Anzures, G., Quinn, P., Pascalis, O., Slater, A., & Lee, K. (2009). Categorization, categorical perception, and asymmetry in infants' representation of face race: Infants' representation of face race. *Developmental Science*, *13*(4), 553–564. doi: 10.1111/j.1467-7687.2009.00900.x

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), 1890–1900.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society. Series A (General)*, *132*(2), 235–244. doi: 10.2307/2343787

Arnold, B. F., Hogan, D. R., Colford, J. M., & Hubbard, A. E. (2011). Simulation methods to estimate design power: An overview for applied research. *BMC medical research methodology*, *11*(1), 1–10. doi: 10.1186/1471-2288-11-94

Ashmead, D. H., & Davis, D. L. (1996). Measuring habituation in infants: An approach using regression analysis. *Child Development*, *67*(6), 2677–2690.

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

Aslin, R. N., & McMurray, B. (2004). Automated corneal-reflection eye tracking in infancy: Methodological Developments and Applications to Cognition. *Infancy*, *6*(2), 155–163.

Azevedo-Filho, A., & Shachter, R. D. (1994). Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In R. L. de Mantaras & D. Poole (Eds.), *Uncertainty Proceedings 1994* (pp. 28–36). Elsevier.

Baba, K., Shibata, R., & Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, *46*(4), 657–664.

Baba, K., & Sibuya, M. (2005). Equivalence of partial and conditional correla-

tion coefficients. *Journal of the Japan Statistical Society*, *35*(1), 1–19.

Balakrishnan, N., & Lai, C.-D. (2009). *Continuous bivariate distributions* (2nd ed.). New York: Springer.

Baron, J. A. (1994). Uncertainty in Bayes. *Medical Decision Making*, *14*(1), 46–51.

Barthelme, S. (2020). *imager: Image Processing Library Based on 'CImg'*. Retrieved from `https://CRAN.R-project.org/package=imager`

Barthelmé, S., Trukenbrod, H., Engbert, R., & Wichmann, F. (2013). Modeling fixation locations using spatial point processes. *Journal of Vision*, *13*(12), 1–34. doi: 10.1167/13.12.1

Barzilai, J., & Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, *8*, 141–148. doi: 10.1093/imanum/8.1.141

Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, *40*, 1550–1577.

Bellet, M. E., Bellet, J., Nienborg, H., Hafed, Z. M., & Berens, P. (2019). Human-level saccade detection performance using deep neural networks. *Journal of Neurophysiology*, *121*, 646–661. doi: 10.1152/jn.00601.2018

Berchtold, A., & Raftery, A. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, *17*(3), 328–356.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402. doi: 10.1214/06-BA115

Bergmann, C., & Cristia, A. (2018). Environmental influences on infants' native vowel discrimination: The case of talker number in daily life. *Infancy*, *23*(4), 484–501.

Bergmann, C., Rabagliati, H., & Tsuji, S. (2019). What's in a looking time preference? *PsyArXiv*. doi: 10.31234/osf.io/6u453

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009.

Berzosa, P., de Lucio, A., Romay-Barja, M., Herrador, Z., González, V., García, L., … Benito, A. (2018). Comparison of three diagnostic methods (microscopy, RDT, and PCR) for the detection of malaria parasites in representative samples from Equatorial Guinea. *Malaria Journal*, *17*(1), 1–12.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205–238.

Bianchi, M. T., Alexander, B. M., & Cash, S. S. (2009). Incorporating uncertainty into medical decision making: An approach to unexpected test results. *Medical Decision Making*, *29*(1), 116–124.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–726. doi: 10.1109/34.865189

Bijeljac-Babic, R., Höhle, B., & Nazzi, T. (2016). Early prosodic acquisition in bilingual infants: The case of the perceptual trochaic bias. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.00210

Birgin, E. G., Martinez, J. M., & Raydan, M. (2000). Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal of Optimization*, *10*, 1196–1211. doi: 10.1137/S1052623497330963

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Blignaut, P. (2009). Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics*, *71*(4), 881–895. doi: 10.3758/APP.71.4.881

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in neurosciences*, *33*(1), 10–16.

Bogartz, R. S. (1965). The criterion method: Some analyses and remarks. *Psychological Bulletin*, *64*(1), 1–14.

Boisvert, J. F., & Bruce, N. D. (2016). Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing*, *207*, 653–668.

Bolstad, W. M. (2007). *Introduction to Bayesian Statistics (2nd ed.)*. Hoboken, NJ: Wiley.

Borchers, H. W. (2019). *pracma: Practical Numerical Math Functions*. Retrieved from https://CRAN.R-project.org/package=pracma

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Bornstein, M. H. (1985). Habituation of attention as a measure of visual information processing in human infants: Summary, systematization, and synthesis. In G. Gottlieb & N. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life: A methodological overview* (pp. 253–300). Ablex Publishing.

Bornstein, M. H., Colombo, J., & Pauen, S. (2012). Infant cognitive functioning and mental development. In *Early childhood development and later outcome* (pp. 118–147). Cambridge University Press.

Borsboom, D., van der Maas, H., Dalege, J., Kievit, R., & Haig, B. (2020). Theory construction methodology: A practical framework for theory formation in psychology. *PsyArXiv*. doi: 10.31234/osf.io/w5tp8

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766.

Bosker, R., & Snijders, T. A. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, *143*(4), 383–404.

Brannon, E. (2002). The development of ordinal numerical knowledge in infancy. *Cognition*, *83*(3), 223–240. doi: 10.1016/S0010-0277(02)00005-7

Brenna, V., Proietti, V., Montirosso, R., & Turati, C. (2013). Positive, but not negative, facial expressions facilitate 3-month-olds' recognition of an individual face. *International Journal of Behavioral Development*, *37*(2), 137–142. doi: 10.1177/0165025412465363

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a

binomial proportion. *Statistical science*, *16*(2), 101–117.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153–178. doi: 10.1016/j.cogpsych.2007.12.002

Brunson, J. C. (2020). Ggalluvial: Layered grammar for alluvial plots. *Journal of Open Source Software*, *5*(49), 1–6.

Bulf, H., Brenna, V., Valenza, E., Johnson, S., & Turati, C. (2015). Many faces, one rule: The role of perceptual expertise in infants' sequential rule learning. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01595

Bulf, H., Johnson, S., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–132. doi: 10.1016/j.cognition.2011.06.010

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M., Hamlin, J., Kline, M., & others. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 349–363.

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable infant research. *Infant and Child Development*, *31*(5), e2296.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 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.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. doi: 10.18637/jss.v076.i01

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven progressive matrices test. *Psychological review*, *97*(3), 404.

Carpenter, R. H. S. (1981). Oculomotor procrastination. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye Movements: Cognition and Visual Perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carpenter, R. H. S., & Williams, M. (1995). Neural computation of log likeli-

hood in control of saccadic eye movements. *Nature*, *377*(6544), 59–62. doi: 10.1038/377059a0

Chen, S.-C., She, H.-C., Chuang, M.-H., Wu, J.-Y., Tsai, J.-L., & Jung, T.-P. (2014). Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities. *Computers & Education*, *74*, 61–72. doi: 10.1016/j.compedu.2013.12.012

Chhikara, R., & Folks, L. J. (1988). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. CRC Press.

Chormunge, S., & Jena, S. (2018). Correlation based feature selection with clustering for high dimensional data. *Journal of Electrical Systems and Information Technology*, *5*(3), 542–549.

Chouinard, B., Scott, K., & Cusack, R. (2019). Using automatic face analysis to score infant behaviour from video collected online. *Infant Behavior and Development*, *54*, 1–12.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *Journal of vision*, *14*(11), 1–14.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2017). Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling. *Vision research*, *141*, 204–216.

Chuk, T., Crookes, K., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2017). Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures. *Cognition*, *169*, 102–117.

Clarke, A. D. F., Stainer, M. J., Tatler, B. W., & Hunt, A. R. (2017). The saccadic flow baseline: Accounting for image-independent biases in fixation behavior. *Journal of Vision*, *17*(11), 1–19. doi: 10.1167/17.11.12

Clarke, A. D. F., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41–51. doi: 10.1016/j.visres.2014.06.016

Clohessy, A., Posner, M., & Rothbart, M. (2001). Development of the functional visual field. *Acta Psychologica*, *106*(1-2), 51–68.

Coccia, M. (2021). Effects of the spread of COVID-19 on public health of pol-

luted cities: Results of the first wave for explaining the dejà vu in the second wave of COVID-19 pandemic and epidemics of future vital agents. *Environmental Science and Pollution Research*, *28*(15), 19147–19154.

Cohen, A. S., Kane, M. T., & Kim, S.-H. (2001). The precision of simulation study results. *Applied Psychological Measurement*, *25*(2), 136–145.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. NJ: Lawrence Earlbaum Associates.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101.

Cohen, L. B. (2004). Uses and misuses of habituation and related preference paradigms. *Infant and Child Development*, *13*(4), 349–352.

Cohen, L. B., & Menten, T. G. (1981). The rise and fall of infant habituation. *Infant Behavior and Development*, *4*, 269–280.

Colombo, J. (2002). Infant attention grows up: The emergence of a developmental cognitive neuroscience perspective. *Current Directions in Psychological Science*, *11*(6), 196–200.

Colombo, J., Mitchell, D., O'Brien, M., & Horowitz, F. (1987). The stability of visual habituation during the first year of life. *Child Development*, *58*(2), 474–487.

Colombo, J., & Mitchell, D. W. (1990). Individual differences in early visual attention: Fixation time and information processing. In J. Colombo & J. W. Fagen (Eds.), *Individual differences in infancy: Reliability, stability, prediction* (pp. 193–227). Lawrence Erlbaum Associates.

Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of Learning and Memory*, *92*(2), 225–234.

Colombo, J., Shaddy, D., Richman, W., Maikranz, J., & Blaga, O. (2004). The developmental course of habituation in infancy and preschool outcome. *Infancy*, *5*(1), 1–38.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*, 140216. doi: 10.1098/rsos.140216

Cong, Y., Junge, C., Aktar, E., Raijmakers, M. E. J., Franklin, A., & Sauter, D. (2019). Pre-verbal infants perceive emotional facial expressions categorically. *Cognition and Emotion*, *33*(3), 391–403.

Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *The Stata Journal*, *20*(1), 131–148.

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29.

Coutrot, A., Binetti, N., Harrison, C., Mareschal, I., & Johnston, A. (2016). Face exploration dynamics differentiate men and women. *Journal of vision*, *16*(14), 16–16.

Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods*, *50*(1), 362–379. doi: 10.3758/s13428-017-0876-8

Crawford, J. R., Garthwaite, P. H., & Betkowska, K. (2009). Bayes' theorem and diagnostic tests in neuropsychology: Interval estimates for post-test probabilities. *The Clinical Neuropsychologist*, *23*(4), 624–644.

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, *42*(3), 692–700.

Crosby, M. E., & Peterson, W. W. (1991). Using eye movements to classify search strategies. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 35, pp. 1476–1480). Los Angeles, CA: Sage.

Crüwell, S., Stefan, A. M., & Evans, N. J. (2019). Robust standards in cognitive science. *Computational Brain & Behavior*, *2*(3), 255–265. doi: 10.1007/s42113-019-00049-8

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521–536.

Dahlin, M. P. (2004). *Infant Visual Habituation Modeled by Non-linear Regression* (PhD Thesis, Pennsylvania State University). Retrieved from https://etda.libraries.psu.edu/catalog/6208

Dalrymple, K. A., Manner, M. D., Harmelink, K. A., Teska, E. P., & Elison, J. T. (2018). An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Frontiers in psychology*, *9*, 803.

Damon, F., Quinn, P., Heron-Delaney, M., Lee, K., & Pascalis, O. (2016). Development of category formation for faces differing by age in 9- to 12-month-olds: An effect of experience with infant faces. *British Journal of Developmental Psychology*, *34*(4), 582–597.

Dannemiller, J. L. (1984). Infant habituation criteria: I. A Monte Carlo study of the 50% decrement criterion. *Infant Behavior & Development*, *7*(2), 147–166.

Davis-Kean, P. E., & Ellis, A. (2019). An overview of issues in infant and developmental research for the creation of robust and replicable science. *Infant Behavior and Development*, *57*, 101339. doi: 10.1016/j.infbeh.2019.101339

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, *111*(3), 378–382.

DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2020, July). Robust data and power in infant research: A case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy*, *25*(4), 393–419. doi: 10.1111/infa.12337

de Groot, A. D. (2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas] (E.-J. Wagenmakers et al., Trans.). *Acta psychologica*, *148*, 188–194.

De Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., & Gegenfurtner, K. R. (2019). Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences*, *116*(24), 11687–11692. doi: 10.1073/pnas.1820553116

de Heide, R., & Grünwald, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review*, *28*, 795–812. doi: 10.3758/s13423-020-01803-x

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *19*(1), 1–38. doi: 10.1111/j.2517-6161.1977.tb01600.x

Diaconis, P., & Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, *7*(2), 269–281.

Dickey, J. M., & Lientz, B. P. (1970). The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Domsch, H., Lohaus, A., & Thomas, H. (2009). Learning and retention in 3-and 6-month-old infants: A comparison of different experimental paradigms. *European Journal of Developmental Psychology*, *6*(3), 396–407.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic bulletin & review*, *18*(1), 61–69.

Duchowski, A. T. (2017). *Eye tracking methodology* (3rd ed.). London: Springer.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177.

Dupierrix, E., Hillairet de Boisferon, A., Barbeau, E., & Pascalis, O. (2015). Memory for complex visual objects but not for allocentric locations during the first year of life. *International Journal of Behavioral Development*, *39*(4), 332–338.

Dutilh, G., Wagenmakers, E.-J., Visser, I., & van der Maas, H. L. (2011). A phase transition model for the speed-accuracy trade-off in response time experiments. *Cognitive Science*, *35*(2), 211–250.

Eddelbuettel, D., Francois, R., Allaire, J. J., Ushey, K., Kou, Q., Russell, N., … Chambers, J. (2020). *Rcpp: Seamless R and C++ Integration*. New York: Springer.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, *40*(8), 1–18. doi: 10.18637/jss.v040.i08

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242.

Ehinger, B. V., Groß, K., Ibs, I., & Peter, K. (2019). A new comprehensive

eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. *BioRxiv*. doi: 10.1101/536243

Einhäuser, W., Atzert, C., & Nuthmann, A. (2020). Fixation durations in natural scene viewing are guided by peripheral scene content. *Journal of Vision*, *20*(4), 1–15. doi: 10.1167/jov.20.4.15

Ellis, S. R., & Stark, L. (1986). Statistical dependency in visual scanning. *Human factors*, *28*(4), 421–438.

Epskamp, S. (2017). *Network Psychometrics* (PhD Thesis). Retrieved from http://sachaepskamp.com/dissertation/EpskampDissertation.pdf

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 1–5.

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788.

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, *1*(1), 60–69.

Evans, J. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Evans, M., & Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, *1*(4), 893–914.

Evans, N. (2019). A method, framework, and tutorial for efficiently simulating models of decision-making. *Behavior research methods*, *51*(5), 2390–2404.

Evans, N. (2020). Same model, different conclusions: An identifiability issue in the linear ballistic accumulator model of decision-making. *PsyArXiv*. doi: 10.31234/osf.io/2xu7f

Evans, N., & Brown, S. D. (2018). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior Research Methods*, *50*(2), 589–603.

Evans, N., & Wagenmakers, E.-J. (2019). Evidence accumulation models: Current limitations and future directions. *The Quantitative Methods for Psychology*, *16*(2), 73–90.

Fagan III, J. F. (1970). Memory in the infant. *Journal of experimental child*

psychology, *9*(2), 217–226.

Fantz, R. (1957). Form preferences in newly hatched chicks. *Journal of Comparative and Physiological Psychology*, *50*(5), 422.

Fantz, R. (1958). Visual discrimination in a neonate chimpanzee. *Perceptual and Motor Skills*, *8*(3), 59–66.

Fantz, R. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668–670.

Faraggi, D., & Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, *21*(20), 3093–3106.

Fennell, C., & Werker, J. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech*, *46*(2–3), 245–264.

Field, A. (2017). *Discovering statistics using IBM SPSS Statistics* (5th ed.). London: Sage.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing* (No. 37). 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. doi: 10.1017/s0140525x99002150

Fiser, J., & Aslin, R. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826.

Flom, R., & Pick, A. (2012). Dynamics of infant habituation: Infants' discrimination of musical excerpts. *Infant Behavior and Development*, *35*(4), 697–704.

Formann, A. K., & Piswanger, K. (1979). Wiener Matrizen-Test (WMT). *Weinheim: Beltz*.

Formann, A. K., Waldherr, K., & Piswanger, K. (2011). *Wiener Matrizen-Test 2: WMT-2; ein Rasch-skalierter sprachfreier Kurztest zu Erfassung der Intelligenz*. Hogrefe.

Forney Jr, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268–278. doi: 10.1109/PROC.1973.9030

Forstmann, B. U., & Wagenmakers, E.-J. (2015). *An Introduction to Model-based Cognitive Neuroscience*. New York: Springer.

Foulsham, T., Frost, E., & Sage, L. (2018). Stable individual differences predict

eye movements to the left, but not handedness or line bisection. *Vision Research*, *144*, 38–46. doi: 10.1016/j.visres.2018.02.002

Foulsham, T., Gray, A., Nasiopoulos, E., & Kingstone, A. (2013). Leftward biases in picture scanning and line bisection: A gaze-contingent window study. *Vision Research*, *78*, 14–25. doi: 10.1016/j.visres.2012.12.001

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … others (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435.

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental Science*, *12*(4), 504–509.

Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal*, *48*(4), 537–555. doi: 10.1002/bimj.200510238

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, *7*(1), 143–167.

Frühwirth-Schnatter, S. (2019). Keeping the balance—Bridge sampling for marginal likelihood estimation in finite mixture, mixture of experts and Markov mixture models. *Brazilian Journal of Probability and Statistics*, *33*(4), 706–733.

Gabry, J., & Češnovar, R. (2020). *cmdstanr: R Interface to 'CmdStan'*. Retrieved from `https://CRAN.R-project.org/package=cmdstanr`

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, *28*(2), 210–216.

Gasparini, L., Iverson, E., El-Shawa, S., Tsuji, S., Frank, M., & Bergmann, C. (2021). MetaLab and metalabR: Facilitating dynamic meta-analyses in developmental psychology. *Research Synthesis & Big Data 2021 Virtual Conference*.

Geambasu, A., van Renswoude, D., Visser, I., Raijmakers, M. E. J., & Levelt, C. (2021). Marcus et al. (1999) revisited: Which mechanism underlies

infants' abstraction of algebraic rules? *in preparation*.

Geambașu, A., Spit, S., van Renswoude, D., Blom, E., Fikkert, P. J., Hunnius, S., … others (2022). Robustness of the rule-learning effect in 7-month-old infants: A close, multicenter replication of Marcus et al.(1999). *Developmental Science*, e13244.

Geisser, S., & Johnson, W. O. (2006). *Modes of parametric statistical inference* (Vol. 529). John Wiley & Sons.

Gelman, A., & Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, *13*(2), 163–185. doi: 10.1214/ss/1028905934

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. doi: 10.1016/j.jmp.2016.01.006

Gervain, J., Berent, I., & Werker, J. F. (2012). Binding at birth: The newborn brain detects identity relations and sequential position in speech. *Journal of Cognitive Neuroscience*, *24*(3), 564–574.

Gierasimczuk, N., van der Maas, H. L., & Raijmakers, M. E. J. (2013). An analytic tableaux model for deductive mastermind empirically tested with a massively used online learning system. *Journal of Logic, Language and Information*, *22*(3), 297–314.

Gigerenzer, G. (1996). The psychology of good judgment: Frequency formats and simple algorithms. *Medical Decision Making*, *16*(3), 273–280.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704.

Gilmore, R. O., & Thomas, H. (2002). Examining individual differences in infants' habituation patterns using objective quantitative techniques. *Infant Behavior and Development*, *25*(4), 399–412.

Glady, Y., Thibaut, J.-P., & French, B. (2013). Visual strategies in analogical rea-

soning development: A new method for classifying scanpaths. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35).

Goldberg, J. H., & Helfman, J. I. (2010a). Identifying Aggregate Scanning Strategies to Improve Usability Evaluations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 54, pp. 590–594). Los Angeles, CA: Sage.

Goldberg, J. H., & Helfman, J. I. (2010b). Scanpath clustering and aggregation. In *Proceedings of the 2010 Symposium on Eye-tracking Research & Applications* (pp. 227–234). ACM.

Gredebäck, G., Johnson, S., & von Hofsten, C. (2009). Eye tracking in infancy research. *Developmental Neuropsychology*, *35*(1), 1–19.

Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*, *62*, 1–8.

Gronau, Q., Heathcote, A., & Matzke, D. (2019). Computing Bayes factors for evidence-accumulation models using Warp-III bridge sampling. *Behavior Research Methods*, *52*, 918–937.

Gronau, Q., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-tests. *The American Statistician*, *74*, 137–143.

Gronau, Q., Singmann, H., & Wagenmakers, E.-J. (2017). bridgesampling: An R package for estimating normalizing constants. *arXiv*. doi: 10.48550/arXiv.1710.08162

Gronau, Q., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R Package for Estimating Normalizing Constants. *Journal of Statistical Software*, *92*(10), 1–29.

Groner, R., & Groner, M. (1982). Towards a hypothetico-deductive theory of cognitive activity. *Cognition and Eye Movements*, *10*, 100–121.

Groner, R., Walder, F., & Groner, M. (1984). Looking at faces: Local and global aspects of scanpaths. In A. G. Gale & F. Johnson (Eds.), *Theoretical and Applied Aspects of Eye Movement Research* (Vol. 22, pp. 523–533). North Holland: Elsevier.

Groves, P. M., & Thompson, R. F. (1970). Habituation: a dual-process theory. *Psychological review*, *77*(5), 419–450.

Guo, J., Gabry, J., & Goodrich, B. (2020). *rstan: R Interface to Stan*. Retrieved

from https://CRAN.R-project.org/package=rstan

Gureckis, T., & Love, B. (2004). Common mechanisms in infant and adult category learning. *Infancy*, *5*(2), 173–198.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, *3*(10), e189.

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*(3), 772–789.

Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision Research*, *103*, 127–142.

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2021). *Doing Meta-Analysis with R: A Hands-On Guide*. London: Chapman & Hall/CRC Press.

Hartigan, J. A., & Hartigan, J. (1975). *Clustering algorithms* (Vol. 209). New York: Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning* (2nd ed.). Springer. Retrieved from https://web.stanford.edu/~hastie/ElemStatLearn/

Hayes, T. R., & Henderson, J. M. (2017). Scan patterns during real-world scene viewing predict individual differences in cognitive capacity. *Journal of Vision*, *17*(5), 23–23. doi: 10.1167/17.5.23

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision*, *11*(10), 10–10. doi: 10.1167/11.10.10

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence*, *48*, 1–14. doi: 10.1016/j.intell.2014.10.005

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience* (pp. 25–48). New York: Springer.

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in psychology*, *3*, 292.

Heck, D., Gronau, Q., & Wagenmakers, E.-J. (2019). *metaBMA: Bayesian model averaging for random and fixed effects meta-analysis [Computer software manual]*.

Hein, O., & Zangemeister, W. H. (2017). Topology for gaze analyses - Raw data segmentation. *Journal of Eye Movement Research*, *10*(1), 1–25. doi: 10.16910/jemr.10.1.1

Henderson, A., Wang, Y., Matz, L., & Woodward, A. (2013). Active experience shapes 10-month-old infants' understanding of collaborative goals: Experience and collaboration. *Infancy*, *18*(1), 10–39.

Henderson, A., & Woodward, A. (2011). "Let's work together": What do infants understand about collaborative goals? *Cognition*, *121*(1), 12–21.

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 1–18. doi: 10.1167/18.6.10

Henderson, J. M., Nuthmann, A., & Luke, S. G. (2013). Eye movement control during scene viewing: Immediate effects of scene luminance on fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 318–322. doi: 10.1037/a0031224

Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PloS One*, *8*(5), e64937.

Hendriksen, A., de Heide, R., & Grünwald, P. (2021). Optional stopping with Bayes factors: A categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis*, *16*(3), 961–989. doi: 10.1214/20-BA1234

Hessels, R., Andersson, R., Hooge, I. T. C., Nyström, M., & Kemner, C. (2015). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, *20*(6), 601–633.

Hessels, R., & Hooge, I. T. C. (2019). Eye tracking in developmental cognitive neuroscience–The good, the bad and the ugly. *Developmental Cognitive Neuroscience*, *40*, 100710.

Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. C. (2016).

The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods*, *48*(4), 1694–1712.

Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. C. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behavior Research Methods*, *49*, 1802–1823. doi: 10.3758/s13428-016-0822-1

Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. T. C. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society Open Science*, *5*(8), 180502.

Hild, J., Voit, M., Kühnle, C., & Beyerer, J. (2018). Predicting observer's task from eye movement patterns during motion image analysis. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (p. 58). ACM.

Hinne, M., Gronau, Q., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. doi: 10.1177/2515245919898657

Hock, A., Oberst, L., Jubran, R., White, H., Heck, A., & Bhatt, R. (2017). Integrated emotion processing in infancy: Matching of faces and bodies. *Infancy*, *22*(5), 608–625.

Hoffrage, U., & Gigerenzer, G. (2004). How to improve the diagnostic inferences of medical experts. In Kurz-Milke, E. & Gigerenzer, G. (Eds.), *Experts in science and society* (pp. 249–268). Berlin, Germany: Springer Science & Business Media.

Hofman, A. D., Visser, I., Jansen, B. R., Marsman, M., & van der Maas, H. L. (2018). Fast and slow strategies in multiplication. *Learning and Individual Differences*, *68*, 30–40.

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal*, *34*(6), 357–359.

Hood, B., Murray, L., King, F., Hooper, R., Atkinson, J., & Braddick, O. (1996). Habituation changes in early infancy: Longitudinal measures from birth to 6 months. *Journal of Reproductive and Infant Psychology*,

14(3), 177–185.

Hooge, I. T. C., Niehorster, D. C., Nyström, M., Andersson, R., & Hessels, R. S. (2018). Is human classification by experienced untrained observers a gold standard in fixation detection? *Behavior Research Methods*, *50*, 1864–1881. doi: 10.3758/s13428-017-0955-x

Horowitz, F. D., Paden, L., Bhana, K., Aitchison, R., & Self, P. (1972). Developmental changes in infant visual fixation to differing complexity levels among cross-sectionally and longitudinally studied infants. *Developmental Psychology*, *7*(1), 88–89.

Horowitz, F. D., Paden, L., Bhana, K., & Self, P. (1972). An infant-control procedure for studying infant visual fixations. *Developmental Psychology*, *7*(1), 90. doi: 10.1037/h0032855

Houpt, J. W., Frame, M. E., & Blaha, L. M. (2018). Unsupervised parsing of gaze data with a beta-process vector auto-regressive hidden Markov model. *Behavior Research Methods*, *50*, 2074–2096. doi: 10.3758/s13428-017-0974-7

Hubach, R. D., Mahaffey, C., Rhoads, K., O'Neil, A. M., Ernst, C., Bui, L. X., … Giano, Z. (2021). Rural college students' amenability toward using at-home human immunodeficiency virus and sexually transmitted infection testing kits. *Sexually Transmitted Diseases*, *48*(8), 583–588.

Hudson, T. E. (2021). *Bayesian Data Analysis for the Behavioral and Neural Sciences: Non-Calculus Fundamentals*. Cambridge University Press.

Hunnius, S. (2007). The early development of visual attention and its implications for social and cognitive development. *Progress in Brain Research*, *164*, 187–209.

Hunter, M., Ames, E., & Koopman, R. (1983). Effects of stimulus complexity and familiarization time on infant preferences for novel and familiar stimuli. *Developmental Psychology*, *19*(3), 338.

Hunter, M., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, *5*, 69–95.

Huth, K., de Ron, J., Luigjes, J., Goudriaan, A., Mohammadi, R., van Holst, R., … Marsman, M. (in press). Bayesian analysis of cross-sectional networks: A tutorial in R and JASP. *Advances in Methods and Practices in Psychological Science*.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., … Yasmeen, F. (2023). *forecast: Forecasting functions for time series and linear models*. Retrieved from https://pkg.robjhyndman.com/forecast/

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*(3), 1–22. doi: 10.18637/jss.v027.i03

IBM Corp. (2017). *IBM SPSS Statistics [Computer Software]*. Armonk, NY: IBM Corp. Retrieved from https://www.ibm.com/products/spss-statistics

Ichikawa, H., Kanazawa, S., & Yamaguchi, M. (2014). Infants recognize the subtle happiness expression. *Perception*, *43*(4), 235–248.

Itti, L., & Borji, A. (2014). Computational models: Bottom-up and top-down aspects. In A. C. Nobre & S. Kastner (Eds.), *Oxford Handbook of Attention*. Oxford University Press.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506. doi: 10.1016/S0042-6989(99)00163-7

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. doi: 0.1038/35058500

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. doi: 10.1109/34.730558

Jaeger, R. G., & Halliday, T. R. (1998). On confirmatory versus exploratory research. *Herpetologica*, S64–S66.

JASP Team. (2021). *JASP (Version 0.15)[Computer Software]*. Retrieved from https://jasp-stats.org/

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, *20*(1), 50–67. doi: 10.1214/088342305000000016

Jean, S., Burnham, C.-A. D., Chapin, K., Garner, O. B., Pant Pai, N., Turabelidze, G., & Butler-Wu, S. (2022). At-home testing for infectious diseases: The laboratory where you live. *Clinical Chemistry*, *68*(1), 19–26.

Jeffrey, W. (1968). The orienting reflex and attention in cognitive development. *Psychological Review*, *75*(4), 323–334. doi: 10.1037/h0025898

Jeffreys, H. (1935). Some Tests of Significance, Treated by the Theory of Probability. *Proceedings of the Cambridge Philosophy Society*, *31*, 203–222.

Jeffreys, H. (1939). *Theory of Probability* (First ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1961). *Theory of Probability* (Third ed.). Oxford, UK: Oxford University Press.

Johnson, M., & Tucker, L. (1996). The development and temporal dynamics of spatial orienting in infants. *Journal of Experimental Child Psychology*, *63*(1), 171–188.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N., Marcus, G., Rabagliati, H., & Slemmer, J. A. (2009). Abstract rule learning for visual sequences in 8-and 11-month-olds. *Infancy*, *14*(1), 2–18.

Johnson, V. E., & Rossell, D. (2010). On the Use of Non-Local Prior Densities in Bayesian Hypothesis Tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(2), 143–170. doi: 10.1111/j.1467-9868.2009.00730.x

Jones, E., Pascalis, O., Eacott, M., & Herbert, J. (2011). Visual recognition memory across contexts. *Developmental Science*, *14*(1), 136–147.

Kagan, J., & Lewis, M. (1965). Studies of attention in the human infant. *Merrill-Palmer Quarterly of Behavior and Development*, *11*(2), 95–127.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251.

Kajikawa, S., Fais, L., Mugitani, R., Werker, J., & Amano, S. (2006). Cross-language sensitivity to phonotactic patterns in infants. *The Journal of the Acoustical Society of America*, *120*(4), 2278–2284.

Kanan, C., Ray, N. A., Bseiso, D. N., Hsiao, J. H., & Cottrell, G. W. (2014). Predicting an observer's task using multi-fixation pattern analysis. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 287–290).

Kasneci, E., Kasneci, G., Kübler, T. C., & Rosenstiel, W. (2014). The applicability of probabilistic methods to the online recognition of fixations and saccades in dynamic scenes. In *Proceedings of the Sympo-*

sium on Eye Tracking Research and Applications (pp. 323–326). doi: 10.1145/2578153.2578213

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.1080/01621459.1995.10476572

Kattari, S. K., Gross, E. B., Harner, V., Andrus, E., Stroumsa, D., Moravek, M. B., & Brouwer, A. (in press). "Doing it on my own terms": Transgender and nonbinary adults' experiences with HPV self-swabbing home testing kits. *Women's Reproductive Health*, 1–17.

Kavšek, M. (2004). Predicting later IQ from infant visual habituation and dishabituation: A meta-analysis. *Journal of Applied Developmental Psychology*, *25*(3), 369–393.

Kay, M. (2020). *tidybayes: Tidy Data and 'Geoms' for Bayesian Models*. Retrieved from `https://CRAN.R-project.org/package=tidybayes`

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–385.

Kellman, P. J., & Arterberry, M. E. (2000). *The cradle of knowledge: Development of perception in infancy*. MIT Press.

Kelly, D., Quinn, P., Slater, A., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychological Science*, *18*(12), 1084–1089.

Kennedy, L., Simpson, D., & Gelman, A. (2019). The Experiment is just as Important as the Likelihood in Understanding the Prior: a Cautionary Note on Robust Cognitive Modeling. *Computational Brain & Behavior*, *2*(3), 210–217. doi: 10.1007/s42113-019-00051-0

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes Factor Hypothesis Testing in Neuroscience to Establish Evidence of Absence. *Nature Neuroscience*, *23*(7), 788–799. doi: 10.1038/s41593-020-0660-4

Kidd, C., Piantadosi, S., & Aslin, R. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*(5), e36399.

Kim, S. (2015). ppcor: An R Package for a Fast Calculation to Semi-partial

Correlation Coefficients. *Communications for Statistical Applications and Methods*, *22*(6), 665–674. doi: 10.5351/CSAM.2015.22.6.665

Kirkham, N., Slemmer, J., Richardson, D., & Johnson, S. (2007). Location, location, location: Development of spatiotemporal sequence learning in infancy. *Child Development*, *78*(5), 1559–1571. doi: 10.1111/j.1467-8624.2007.01083.x

Kit, D., & Sullivan, B. (2016). Classifying mobile eye tracking data with hidden Markov models. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (pp. 1037–1040). ACM.

Klein, R. M. (2000). Inhibition of return. *Trends in cognitive sciences*, *4*(4), 138–147.

Klement, R. J., & Bandyopadhyay, P. S. (2021). The epistemology of a positive SARS-CoV-2 test. *Acta Biotheoretica*, *69*(3), 359–375.

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, *57*(11), 2635–2645. doi: 10.1109/TBME.2010.2057429

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. doi: 10.3758/s13423-016-1221-4

Kucharský, Š., Tran, N.-H., Veldkamp, K., Raijmakers, M., & Visser, I. (2021). Hidden Markov models of evidence accumulation in speeded decision tasks. *Computational Brain & Behavior*, *4*, 416–441. doi: 10.1007/s42113-021-00115-0

Kucharský, Š., van Renswoude, D., Raijmakers, M., & Visser, I. (2021). WALD-EM: Wald accumulation for locations and durations of eye movements. *Psychological Review*, *128*(4), 667-689. doi: 10.1037/rev0000292

Kucharský, Š., Visser, I., Truțescu, G.-O., Laurence, P. G., Zaharieva, M., & Raijmakers, M. E. (2020). Cognitive strategies revealed by clustering eye movement transitions. *Journal of Eye Movement Research*, *13*(1). doi: 10.16910/jemr.13.1.1

Kucharský, Š., & Wagenmakers, E.-J. (2023). Correct conclustions from fallible medical tests: A tutorial with JASP. *OSF Preprints*. doi: 10.31219/osf.io/jksz6

Kucharský, Š., Wagenmakers, E.-J., van den Bergh, D., & Ly, A. (2023). Analytic posterior distribution and Bayes factor for Pearson partial correlations. *PsyArXiv*. doi: 10.31234/osf.io/6muwy

Kucharský, Š., Zaharieva, M., Raijmakers, M., & Visser, I. (2022). Habituation, Part II. Rethinking the habituation paradigm. *Infant and Child Development*, e2383. doi: 10.1002/icd.2383

Kuijpers, R. E., Visser, I., & Molenaar, D. (2021). Testing the within-state distribution in mixture models for responses and response times. *Journal of Educational and Behavioral Statistics*, *46*(3), 348–373.

Kübler, T. C., Rothe, C., Schiefer, U., Rosenstiel, W., & Kasneci, E. (2017). SubsMatch 2.0: Scanpath comparison and classification based on subsequence frequencies. *Behavior Research Methods*, *49*(3), 1048–1064.

Landis, R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Langrock, R., Kneib, T., Sohn, A., & Deruiter, S. L. (2015). Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, *71*(2), 520–528. doi: 10.1111/biom.12282

Larsson, L., Nystrom, M., & Stridh, M. (2013). Detection of saccades and postsaccadic oscillations in the presence of smooth pursuit. *IEEE Transactions on Biomedical Engineering*, *60*(9), 2484–2493. doi: 10.1109/TBME.2013.2258918

Larsson, L., Nyström, M., Andersson, R., & Stridh, M. (2015). Detection of fixations and smooth pursuit movements in high-speed eye-tracking data. *Biomedical Signal Processing and Control*, *18*, 145–152. doi: 10.1016/j.bspc.2014.12.008

Lasky, R. E. (1979). Serial habituation or regression to the mean? *Child Development*, *50*(2), 568–570.

Lau, C.-S., & Aw, T.-C. (2021). Disease prevalence matters: Challenge for SARS-CoV-2 testing. *Antibodies*, *10*(4), 50.

Laurence, P. G., Mecca, T. P., Serpa, A., Martin, R., & Macedo, E. C. (2018). Eye Movements and Cognitive Strategy in a Fluid Intelligence Test: Item

Type Analysis. *Frontiers in psychology*, *9*, 380.

Lauritzen, S. L. (1996). *Graphical models* (No. 17). Oxford, UK: Oxford University Press.

Lawrance, A. (1976). On conditional and partial correlation. *The American Statistician*, *30*(3), 146–149.

Leahy, R. L. (1976). Development of preferences and processes of visual scanning in the human infant during the first 3 months of life. *Developmental Psychology*, *12*(3), 250–254.

Lee, M. D. (2008). BayesSDT: Software for Bayesian inference with signal detection theory. *Behavior Research Methods*, *40*(2), 450–456.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., … others (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*(3-4), 141–153.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, UK: Cambridge University Press.

Leigh, R. J., & Zee, D. S. (2015). *The neurology of eye movements* (5th ed.). Oxford, UK: Oxford University Press.

Le Meur, O., & Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision Research*, *116*, 152–164. doi: 10.1016/j.visres.2014.12.026

Lemon, J. (2006). Plotrix: a package in the red light district of R. *R-News*, *6*(4), 8–12.

Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson, B., Wickham, H., … Venables, B. (2019). *plotrix: Various Plotting Functions*. Retrieved from https://CRAN.R-project.org/package=plotrix

Lenhard, W., & Lenhard, A. (2014). *Calculation of test quality criteria for screenings.* Retrieved 2023-10-18, from https://www.psychometrica.de/test_criteria_en.html

Leppänen, J., Forssman, L., Kaatiala, J., Yrttiaho, S., & Wass, S. (2015). Widely applicable MATLAB routines for automated analysis of saccadic reaction times. *Behavior Research Methods*, *47*(2), 538–548.

Letham, B., & Taylor, S. (2017). Forecasting at scale. *PeerJ*, *5*. doi: 10.7287/peerj.preprints.3190v2

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, inser-

tions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

Li, M., & Bolker, B. M. (2017). Incorporating periodic variability in hidden Markov models for animal movement. *Movement Ecology*, *5*(1), 1–12. doi: 10.1186/s40462-016-0093-6

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423.

Ligges, U., Short, T., & Kienzle, P. (2015). *signal: Signal processing*. Retrieved from `http://r-forge.r-project.org/projects/signal/`

Lindley, D. V. (1972). *Bayesian statistics: A review*. Philadelphia, PA: SIAM.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*(1), 22–25.

Lipton, J., & Spelke, E. (2003). Origins of number sense: large-number discrimination in human infants. *Psychological Science*, *14*(5), 396–401.

Lipton, J., & Spelke, E. (2004). Discrimination of large and small numerosities by human infants. *Infancy*, *5*(3), 271–290.

Liu, K. (1988). Measurement error and its impact on partial correlation and multiple linear regression analyses. *American Journal of Epidemiology*, *127*(4), 864–874.

Liu, Y., Hsueh, P.-Y., Lai, J., Sangin, M., Nussli, M.-A., & Dillenbourg, P. (2009). Who is the expert? Analyzing gaze data to predict expertise level in collaborative applications. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on* (pp. 898–901). IEEE.

Lleras, A., Porporino, M., Burack, J. A., & Enns, J. T. (2011). Rapid resumption of interrupted search is independent of age-related improvements in visual search. *Journal of Experimental Child Psychology*, *109*(1), 58–72.

Lloyd-Fox, S., Blasi, A., McCann, S., Rozhko, M., Katus, L., Mason, L., … Team, B. P. (2019). Habituation and novelty detection fNIRS brain responses in 5-and 8-month-old infants: The Gambia and UK. *Developmental Science*, *22*(5), e12817.

Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven Advanced Progressive Matrices Test. *Intelligence*, *48*, 58–75.

Lomax, R. G., & Hahs-Vaughn, D. L. (2012). *Statistical concepts: a second course*

(4th ed.). New York: Routledge.

Luce, R. D. (1991). *Response times: Their role in inferring elementary mental organization* (2nd ed.) (No. 8). Oxford University Press.

Lüken, M., Kucharský, Š., & Visser, I. (2022). Characterising eye movement events with an unsupervised hidden Markov model. *Journal of Eye Movement Research*, *15*(1). doi: 10.16910/jemr.15.1.4

Lund, U., & Agostinelli, C. (2018). *CircStats: Circular statistics, from "Topics in circular statistics" (2001).* Retrieved from https://cran.r-project.org/web/packages/CircStats/

Ly, A., Marsman, M., Verhagen, A. J., Grasman, R. P. P. P., & Wagenmakers, E. (2017). A Tutorial on Fisher Information. *Journal of Mathematical Psychology*, *80*, 40–55.

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*, *72*(1), 4–13. doi: 10.1111/stan.12111

Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., Sarafoglou, A., … others (2020). The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p value hypothesis test. *Computational Brain & Behavior*, *3*(2), 153–161.

Ly, A., van den Bergh, D., Bartoš, F., & Wagenmakers, E.-J. (2021). Bayesian Inference With JASP. *The ISBA Bulletin*, *28*, 7–15.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016b). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. doi: 10.1016/j.jmp.2015.06.004

Maier, M., Bartoš, F., & Wagenmakers, E.-J. (2020). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *PsyArXiv*. doi: 10.31234/osf.io/u4cns

Malem-Shinitski, N., Opper, M., Reich, S., Schwetlick, L., Seelig, S. A., & Engbert, R. (2020). A mathematical model of exploration and exploitation in natural scene viewing. *bioRxiv*. doi: 10.1371/journal.pcbi.1007880

Mani, N., Schreiner, M. S., Brase, J., Köhler, K., Strassen, K., Postin, D., & Schultze, T. (2021). Sequential Bayes Factor designs in developmental research: Studies on early word learning. *Developmental Science*, *24*(4), e13097. doi: 10.1111/desc.13097

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Mardia, K. V., & Jupp, P. E. (2009). *Directional statistics* (Vol. 494). John Wiley & Sons.

Mareschal, D., French, R., & Quinn, P. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, *36*(5), 635–645. doi: 10.1037/0012-1649.36.5.635

Marewski, J. N., & Gigerenzer, G. (2022). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, *14*(1), 77–89.

Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, *14*(5), 545–555.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.

McCall, R. B., & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development*, *64*(1), 57–79.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC press.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, *6*(4), 831–860.

Mercaldo, N. D., Lau, K. F., & Zhou, X. H. (2007). Confidence intervals for predictive values with an emphasis to case–control studies. *Statistics in Medicine*, *26*(10), 2170–2183.

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.

Mihali, A., van Opheusden, B., & Ma, W. J. (2017). Bayesian microsaccade detection. *Journal of Vision*, *17*(1), 1–23. doi: 10.1167/17.1.13

Miller, D. J. (1972). Visual habituation in the human infant. *Child Development*, *43*(2), 481–493.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., & others. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1–9.

Molenaar, D., & de Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 279–297.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral research*, *51*(5), 606–626.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2012). *Introduction to the practice of statistics: with unique international edition exercises* (7th ed.). New York: Freeman and Company.

Morey, R., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs.* Retrieved from `https://cran.r-project.org/web/packages/BayesFactor/index.html`

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125.

Mortensen, M. L., Adam, G. P., Trikalinos, T. A., Kraska, T., & Wallace, B. C. (2017). An exploration of crowdsourcing citation screening for systematic reviews. *Research Synthesis Methods*, *8*(3), 366–386.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … others (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network.

Advances in Methods and Practices in Psychological Science, *1*(4), 501–515.

Mossman, D., & Berger, J. O. (2001). Intervals for posttest probabilities: A comparison of 5 methods. *Medical Decision Making*, *21*(6), 498–507.

Mouselimis, L. (2019). *OpenImageR: An Image Processing Toolkit*. Retrieved from https://CRAN.R-project.org/package=OpenImageR

Mulder, J., & Pericchi, L. R. (2018). The matrix-F prior for estimating and testing covariance matrices. *Bayesian Analysis*, *13*(4), 1193–1214.

Mulder, K., Klugkist, I., van Renswoude, D. R., & Visser, I. (2020). Mixtures of peaked power Batschelet distributions for circular data with application to saccade directions. *Journal of Mathematical Psychology*, *95*, 102309. doi: 10.1016/j.jmp.2019.102309

Müller, K. (2017). *here: A Simpler Way to Find Your Files*. Retrieved from https://CRAN.R-project.org/package=here

Nakahara, H., Nakamura, K., & Hikosaka, O. (2006). Extended LATER model can account for trial-by-trial variability of both pre-and post-processes. *Neural Networks*, *19*(8), 1027–1046.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453.

Nelson, D. G. K., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, *18*(1), 111–116.

Noorani, I., & Carpenter, R. H. S. (2016). The LATER model of reaction time and decision. *Neuroscience & Biobehavioral Reviews*, *64*, 229–251.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The Preregistration Revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. doi: 10.1073/pnas.1708274114

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, *171*(3968), 308–311.

Nuthmann, A. (2017). Fixation durations in scene viewing: Modeling the effects of local image features, oculomotor parameters, and task. *Psycho-*

nomic Bulletin & Review, *24*(2), 370–392. doi: 10.3758/s13423-016-1124-4

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, *117*(2), 382–405. doi: 10.1037/a0018924

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*(7487), 150.

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*(1), 188–204. doi: 10.3758/BRM.42.1.188

Oakes, L. M. (2010). Using habituation of looking time to assess mental processes in infancy. *Journal of Cognition and Development*, *11*(3), 255–268.

Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, *17*(1), 1–8.

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469.

Oakes, L. M., & Kovack-Lesh, K. (2013). Infants' visual recognition memory for a series of categorically related items. *Journal of Cognition and Development*, *14*(1), 63–86.

Oakes, L. M., Madole, K. L., & Cohen, L. B. (1991). Infants' object examining: Habituation and categorization. *Cognitive Development*, *6*(4), 377–392.

Oakes, L. M., Sperka, D., DeBolt, M. C., & Cantrell, L. M. (2019). Habit2: A stand-alone software solution for presenting stimuli and recording infant looking times in order to study infant development. *Behavior Research Methods*, *51*(5), 1943–1952.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*(3), 549–556. doi: 10.2307/2530245

Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science*, *6*(4), 155–156.

Olsson, P. (2007). *Real-time and offline filters for eye tracking* (Unpublished master's thesis). Royal Institute of Technology Stockholm.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation*, *8*(1), 1–15. doi: 10.7275/r222-hv23

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan - A web and mobile app for systematic reviews. *Systematic Reviews*, *5*, 210. doi: 10.1186/s13643-016-0384-4

Page, M., Moher, D., Fidler, F., Higgins, J., Brennan, S., Haddaway, N., & others. (2021). The REPRISE project: protocol for an evaluation of REProducibility and Replicability In Syntheses of Evidence. *Systematic Reviews*, *10*(1), 1–13.

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58–71. doi: 10.1037/a0020747

Pamminger, C., & Frühwirth-Schnatter, S. (2010). Model-based clustering of categorical time series. *Bayesian Analysis*, *5*(2), 345–368.

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.

Paulewicz, B., & Blaut, A. (2020). The BHSDTR package: A general-purpose method of Bayesian inference for Signal Detection Theory models. *Behavior Research Methods*, *52*, 2122–2141.

Pavlov, I. (1927). *Conditioned Reflexes*. New York: Dover Books.

Pedersen, T. L. (2019a). *ggforce: Accelerating 'ggplot2'*. Retrieved from `https://CRAN.R-project.org/package=ggforce`

Pedersen, T. L. (2019b). *patchwork: The Composer of Plots*. Retrieved from `https://CRAN.R-project.org/package=patchwork`

Pekkanen, J., & Lappi, O. (2017). A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Scientific Reports*, *7*, 6. doi: 10.1038/s41598-017-17983-x

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford, UK: Oxford University Press.

Peterson, D. (2016). The Baby Factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius*, *2*, 1–10. doi: 10.1177/2378023115625071

Plude, D., Enns, J., & Brodeur, D. (1994). The development of selective attention: A life-span overview. *Acta Psychologica*, *86*(2-3), 227–272.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, pp. 1–10). Vienna, Austria.

Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*, *64*(2), 191–199. doi: 10.1093/biomet/64.2.191

Pohle, J., Langrock, R., van Beest, F. M., & Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological, and Environmental Statistics*, *22*(3), 270–293. doi: 10.1007/s13253-017-0283-8

Polonio, L., & Coricelli, G. (2018). Testing the level of consistency between choices and beliefs in games using eye-tracking. *Games and Economic Behavior*, *113*, 566–586.

Polonio, L., Di Guida, S., & Coricelli, G. (2015). Strategic sophistication and attention in games: An eye-tracking study. *Games and Economic Behavior*, *94*, 80–96. doi: 10.1016/j.geb.2015.09.003

Ponsoda, V., Scott, D., & Findlay, J. M. (1995). A probability vector and transition matrix analysis of eye movements during visual search. *Acta Psychologica*, *88*(2), 167–185.

Pontius, R. G., & Si, K. (2014). The total operating characteristic to measure diagnostic ability for multiple thresholds. *International Journal of Geographical Information Science*, *28*(3), 570–583.

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*(2), 160–174.

Primates, M., Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., … others (2019). Establishing an infrastructure for collaboration in primate cognition research. *PLoS One*, *14*(10), e0223675.

Quade, D. (2017). Nonparametric partial correlation. In *Measurement in the social sciences* (pp. 369–398). Routledge.

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, *22*(4), 463–475.

Quinn, P. C., Eimas, P. D., & Tarr, M. J. (2001). Perceptual categorization of cat and dog silhouettes by 3-to 4-month-old infants. *Journal of Experimental Child Psychology*, *79*(1), 78–94.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science*, *22*(1), e12704.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. New York, NY: Wiley.

Raijmakers, M. E. J., Schmittmann, V. D., & Visser, I. (2014). Costs and benefits of automatization in category learning of ill-defined rules. *Cognitive Psychology*, *69*, 1–24.

Rankin, C., Abrams, T., Barry, R., Bhatnagar, S., Clayton, D., Colombo, J., & others. (2009). Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory*, *92*(2), 135–138.

Rasti, R. (2022). *Point-of-care diagnostics of childhood central nervous system infections, with a focus on usability in low-resource settings* (PhD Thesis). Karolinska Institutet (Sweden).

Ratcliff, R. (2001). Putting noise into neurophysiological models of simple decision making. *Nature neuroscience*, *4*(4), 336–336.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi: 10.1162/neco.2008.12-06-420

Reichle, E. D., & Sheridan, H. (2015). EZ Reader: An overview of the model and two recent applications. In A. Pollatsek & R. Treiman (Eds.), *The Oxford Handbook of Reading*. Oxford University Press.

Richards, J. (2010). The development of attention to simple and complex visual stimuli in infants: Behavioral and psychophysiological measures. *Developmental Review*, *30*(2), 203–219.

Robinaugh, D., Haslbeck, J., Ryan, O., Fried, E. I., & Waldorp, L. (2020). Invisible hands and fine calipers: A call to use formal theory as a toolkit

for theory construction. *PsyArXiv*. doi: 10.31234/osf.io/ugz7y

Roder, B., Bushnell, E., & Sasseville, A. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy*, *1*(4), 491–507.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *American Psychologist*, *65*(1), 1–12.

Rose, S., Gottfried, A., Melloy-Carminar, P., & Bridger, W. (1982). Familiarity and novelty preferences in infant recognition memory: Implications for information processing. *Developmental Psychology*, *18*(5), 704.

Rothstein, H., Suton, A., & Borenstein, M. (2006). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester, UK: John Wiley & Sons.

Rouder, J. N. (2014). Optional Stopping: No Problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. doi: 10.3758/s13423-014-0595-4

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Saayman, G., Ames, E. W., & Moffett, A. (1964). Response to novelty as an indicator of visual discrimination in the human infant. *Journal of Experimental Child Psychology*, *1*(2), 189–198.

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, *105*(3), 669–680.

Sakuta, Y., Sato, K., Kanazawa, S., & Yamaguchi, M. (2014). The effect of eye size on discriminating faces: Can infants recognize facial uncanniness? *Japanese Psychological Research*, *56*(4), 331–339.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Symposium on Eye Tracking Research and Applications*, 71–78. doi: 10.1145/355017.355028

Sanborn, A. N., & Hills, T. T. (2014). The Frequentist Implications of Optional Stopping on Bayesian Hypothesis Tests. *Psychonomic Bulletin & Review*, *21*, 283–300.

Sanderson, G. (2019). *Bayes theorem, the geometry of changing beliefs*. Retrieved

from https://youtu.be/HZGCoVF3YvM?si=eBkLa1hvYCtt53hm (Publisher: 3blue1brown channel on YouTube)

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639. doi: 10.1021/ac60214a047

Schad, D. J., Betancourt, M., & Vasishth, S. (2019). Toward a principled Bayesian workflow in cognitive science. *arXiv*. Retrieved from https://arxiv.org/abs/1904.12765

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021). Workflow techniques for the robust use of Bayes factors. *arXiv*. doi: 10.48550/arXiv.2103.08744

Schlingloff, L., Csibra, G., & Tatone, D. (2020). Do 15-month-old infants prefer helpers? A replication of Hamlin et al. (2007). *Royal Society Open Science*, *7*, 191795.

Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., & Willemsen, M. C. (2017). Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science*, *26*(5), 442–450. doi: 10.1177/0963721417708229

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

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*(1), 1–11.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813.

Schönbrodt, F. D., & Stefan, A. M. (2018). *BFDA: An R Package for Bayes Factor Design Analysis.*

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently

testing mean differences. *Psychological Methods*, *22*(2), 322–339. doi: 10.1037/met0000061

Schöner, G., & Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychological Review*, *113*(2), 273.

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

Scott, S. L., & Varian, H. R. (2014). Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, *5*(1-2), 4–23.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Shic, F., Scassellati, B., & Chawarska, K. (2008). The incomplete fixation measure. *Proceedings of the Symposium on Eye Tracking Research and Applications*, 111–114. doi: 10.1145/1344471.1344500

Sirois, S., & Mareschal, D. (2002). Models of habituation in infancy. *Trends in Cognitive Sciences*, *6*(7), 293–298.

Sirois, S., & Mareschal, D. (2004). An interacting systems model of infant habituation. *Journal of Cognitive Neuroscience*, *16*(8), 1352–1362.

Slater, A. (1997). Can measures of infant habituation predict later intellectual ability? *Archives of Disease in Childhood*, *77*(6), 474–476.

Slater, A., Brown, E., Mattock, A., & Bornstein, M. (1996). Continuity and change in habituation in the first 4 months from birth. *Journal of Reproductive and Infant Psychology*, *14*(3), 187–194.

Slaughter, V., & Suddendorf, T. (2007). Participant loss due to "fussiness" in infant visual paradigms: A review of the last 20 years. *Infant Behavior and Development*, *30*(3), 505–514.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology*. Routledge.

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. NY, New York: Macmillan.

Sokolov, E. N. (1977). Brain functions: neuronal mechanisms of learning and memory. *Annual Review of Psychology*, *28*, 85–112.

Spakov, O. (2012). Comparison of eye movement filters used in HCI. *Proceedings of the Symposium on Eye Tracking Research and Applications*, 281–284. doi: 10.1145/2168556.2168616

Spezia, L. (2009). Reversible jump and the label switching problem in hidden Markov models. *Journal of Statistical Planning and Inference*, *139*(7), 2305–2315.

Spierings, M. J., & Ten Cate, C. (2016). Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment. *Proceedings of the National Academy of Sciences*, *113*(27), E3977–E3984. doi: 10.1073/pnas.1600483113

Stallard, N., Todd, S., Ryan, E. G., & Gates, S. (2020). Comparison of Bayesian and frequentist group-sequential clinical trial designs. *BMC Medical Research Methodology*, *20*(1), 1–14.

Stampe, D. M. (1993). Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavior Research Methods, Instruments, & Computers*, *25*(2), 137–142. doi: 10.3758/BF03204486

Stan Development Team. (2020). *CmdStan: the command-line interface to Stan*. Retrieved from `https://github.com/stan-dev/cmdstan/releases/tag/v2.24.0-rc1`

Startsev, M., Agtzidis, I., & Dorr, M. (2019). 1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits. *Behavior Research Methods*, *51*(2), 556–572. doi: 10.3758/s13428-018-1144-2

Stefan, A. M., Gronau, Q., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058.

Stefan, A. M., Schönbrodt, F. D., Evans, N. J., & Wagenmakers, E.-J. (2022). Efficiency in sequential testing: Comparing the sequential probability ratio test and the sequential Bayes factor test. *Behavior Research Methods*, 1–18. doi: 10.31234/osf.io/ry4fw

Steingroever, H., Jepma, M., Lee, M. D., Jansen, B. R., & Huizenga, H. M.

(2019). Detecting strategies in developmental psychology. *Computational Brain & Behavior*, *2*, 128–140. doi: 10.1007/s42113-019-0024-x

Stewart, N., Gächter, S., Noguchi, T., & Mullett, T. L. (2016). Eye movements in strategic choice. *Journal of Behavioral Decision Making*, *29*(2-3), 137–156.

Steyvers, M., & Benjamin, A. S. (2019). The joint contribution of participation and performance to learning functions: Exploring the effects of age in large-scale data sets. *Behavior Research Methods*, *51*(4), 1531–1543.

Strang, L., & Simmons, R. K. (2018). *Citizen science: Crowdsourcing for systematic reviews*. THIS Institute. Retrieved from https://www.thisinstitute.cam.ac.uk/research-articles/citizen-science-crowdsourcing-systematic-reviews/

Streri, A., & Pêcheux, M. (1986). Tactual habituation and discrimination of form in infancy: A comparison with vision. *Child Development*, *57*(1), 100–104.

Stuart, S., Hickey, A., Vitorio, R., Welman, K., Foo, S., Keen, D., & Godfrey, A. (2019). Eye-tracker algorithms to detect saccades during static and dynamic tasks: A structured review. *Physiological Measurement*, *40*(2), 1–26. doi: 10.1088/1361-6579/ab02ab

Sun, W., Wang, J., Fang, Y., & others. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, *6*, 148–167.

Suojanen, J. N. (1999). False false positive rates. *New England Journal of Medicine*, *341*(2), 131–132.

Tafaj, E., Kasneci, G., Rosenstiel, W., & Bogdan, M. (2012). Bayesian online clustering of eye movement data. *Proceedings of the Symposium on Eye Tracking Research and Applications*, 285–288. doi: 10.1145/2168556.2168617

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv*. doi: 10.48550/arXiv.1804.06788

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 1–17. doi: 10.1167/7.14.4

Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. S. (2017). LATEST: A model of saccadic decisions in space and time. *Psychological Review*, *124*(3), 267–300. doi: 10.1037/rev0000054

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 1–23. doi: 10.1167/11.5.5

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, *2*(2), 1–18. doi: 10.16910/jemr.2.2.5

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus, eye movements, and vision. *i-Perception*, *1*(1), 7–27.

Thomas, H., & Gilmore, R. O. (2004). Habituation assessment in infancy. *Psychological Methods*, *9*(1), 70–92.

Thompson, R. F., & Spencer, W. A. (1966). Habituation: A model phenomenon for the study of neuronal substrates of behavior. *Psychological Review*, *73*(1), 16–43.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tidy, E. J., Shine, B., Oke, J., & Hayward, G. (2018). Home self-testing kits: Helpful or harmful? *British Journal of General Practice*, *68*(673), 360–361.

Tillman, G., & Evans, N. J. (2020). Redefining Qualitative Benchmarks of Theories and Models: An Empirical Exploration of Fast and Slow Errors in Speeded Decision-making. *PsyArXiv*. doi: 10.31234/osf.io/ze5ns

Tillman, G., Van Zandt, T., & Logan, G. D. (2020). Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. *Psychonomic Bulletin & Review*, *27*, 935.

Timmers, B. (2019). *Mixture Components in Response Times: A Hidden Markov Modeling Approach for Evidence Accumulation Models* (Master's thesis, University of Amsterdam). Retrieved from `https://osf.io/mjpzt/`

Topor, M., Pickering, J., Mendes, A., Bishop, D., Büttner, F., Henderson, E., & others. (2020). An integrative framework for planning and conduct-

ing Non-Interventional, Reproducible, and Open Systematic Reviews (NIRO-SR). *MetaArXiv*. doi: 10.31222/osf.io/8gu5z

Tran, N.-H., Kucharský, Š., Waring, T. M., Atmaca, S., & Beheim, B. A. (2021). Limited scope for group coordination in stylistic variations of kolam art. *Frontiers in Psychology*, *12*, 742577.

Tran, N.-H., van Maanen, L., Heathcote, A., & Matzke, D. (2020). Systematic parameter reviews in cognitive modeling: Towards a robust and cumulative characterization of psychological processes in the diffusion decision model. *Frontiers in Psychology*, *11*, 608287. doi: 10.3389/fpsyg.2020.608287

Trukenbrod, H. A., & Engbert, R. (2014). ICAT: A computational model for the adaptive control of fixation durations. *Psychonomic Bulletin & Review*, *21*(4), 907–934. doi: 10.3758/s13423-013-0575-0

Trutescu, G.-O., & Raijmakers, M. E. J. (2016). *Logical reasoning in a deductive version of the Mastermind game* (Doctoral dissertation). Retrieved from https://doi.org/10.31237/osf.io/hzqx3

Tsang, C. D. (2012). Habituation in Infant Cognition. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning*. New York: Springer.

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7), 1–16. doi: 10.1167/9.7.4

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science*, *9*(6), 661–665.

Tsuji, S., Bergmann, C., Lewis, M., Braginsky, M., Piccinini, P., Frank, M., & Cristia, A. (2017). MetaLab: A Repository for Meta-Analyses on Language Development, and More. In *Interspeech* (pp. 2038–2039).

Turk-Browne, N. B., Scholl, B. J., & Chun, M. M. (2008). Babies and brains: habituation in infant cognition and functional neuroimaging. *Frontiers in Human Neuroscience*, *2*, 16. doi: 10.3389/neuro.09.016.2008

Tuyl, F., Gerlach, R., & Mengersen, K. (2008). A comparison of Bayes–Laplace, Jeffreys, and other priors: The case of zero events. *The American Statistician*, *62*(1), 40–44.

Vakil, E., & Lifshitz-Zehavi, H. (2012). Solving the Raven Pro-

gressive Matrices by adults with intellectual disability with/without Down syndrome: Different cognitive patterns as indicated by eye-movements. *Research in Developmental Disabilities*, *33*(2), 645–654. doi: 10.1016/j.ridd.2011.11.009

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*(1), 1–4. doi: 10.3758/s13423-018-1443-8

van den Bergh, D., Van Doorn, J., Marsman, M., Draws, T., Van Kesteren, E.-J., Derks, K., … others (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année psychologique*, *120*(1), 73–96.

van der Maas, H. L., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*(2), 141–177.

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological review*, *118*(2), 339.

van der Maas, H. L., & Straatemeier, M. (2008). How to detect cognitive strategies: Commentary on 'Differentiation and integration: guiding principles for analyzing cognitive change'. *Developmental science*, *11*(4), 449–453.

van der Stigchel, S., & Theeuwes, J. (2007). The relationship between covert and overt attention in endogenous cuing. *Perception & Psychophysics*, *69*(5), 719–731.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child development*, *85*(3), 842–860.

van Doorn, J., Aust, F., Haaf, J. M., & Wagenmakers, E.-J. (2021). Bayes Factors for Mixed Models. *Computational Brain & Behavior*, *6*, 1–13. doi: 10.1007/s42113-021-00113-2

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*, *72*(4), 303–308. doi: 10.1080/00031305.2016.1264998

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2019). Bayesian estimation of Kendall's tau using a latent normal approach. *Statistics and Probability Letters*, *145*, 268–272.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., … others (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*, 813–826.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., … others (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813–826.

van Maanen, L., Couto, J., & Lebreton, M. (2016). Three boundary conditions for computing the fixed-point property in binary mixture data. *PloS One*, *11*(11), e0167377.

van Maanen, L., Taatgen, N., van Vugt, M., Borst, J., & Mehlhorn, K. (2015). Speed-accuracy trade-off behavior: Response caution adjustment or mixing task strategies? In *Proceedings of ICCM 2015: 13th International Conference on Cognitive Modeling*. University of Groningen.

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*, *55*(1), 106–117. doi: https://doi.org/10.1016/j.jmp.2010.08.005

van Renswoude, D. R., Johnson, S., Raijmakers, M. E. J., & Visser, I. (2016). Do infants have the horizontal bias? *Infant Behavior and Development*, *44*, 38–48. doi: 10.1016/j.infbeh.2016.05.005

van Renswoude, D. R., Raijmakers, M. E. J., Koornneef, A., Johnson, S. P., Hunnius, S., & Visser, I. (2018). Gazepath: An eye-tracking analysis tool that accounts for individual differences and data quality. *Behavior Research Methods*, *50*(2), 834–852.

van Renswoude, D. R., van den Berg, L., Raijmakers, M. E. J., & Visser, I. (2019). Infants' center bias in free viewing of real-world scenes. *Vision Research*, *154*, 44–53. doi: 10.1016/j.visres.2018.10.003

van Renswoude, D. R., Voorvaart, R. E., van den Berg, L., Raijmakers, M. E. J., & Visser, I. (in prep). Object familiarity influences infant gaze control

during free scene viewing. *Manuscript in preparation*.

Varadhan, R., & Gilbert, P. (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software*, *32*(4), 1–26. Retrieved from http://www.jstatsoft.org/

Vargha, A., Bergman, L. R., & Delaney, H. D. (2013). Interpretation problems of the partial correlation with nonnormally distributed variables. *Quality & Quantity*, *47*, 3391–3402.

Veldkamp, K. (2020). *Fitting Mixtures of Linear Ballistic Accumulation models.* University of Amsterdam. Retrieved from https://github.com/Kucharssim/hmm_lba

Venker, C. E., Pomper, R., Mahr, T., Edwards, J., Saffran, J., & Ellis Weismer, S. (2020). Comparing automatic eye tracking and manual gaze coding methods in young children with autism spectrum disorder. *Autism Research*, *13*(2), 271–283.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian Tests to Quantify the Result of a Replication Attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*(3), 261–272. doi: 10.1016/j.intell.2005.11.003

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition*, *30*(2), 171–178.

Visser, I. (2011). Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. *Journal of Mathematical Psychology*, *55*(6), 403–415. doi: 10.1016/j.jmp.2011.08.002

Visser, I., Kucharský, Š., Levelt, C., Stefan, A. M., Wagenmakers, E.-J., & Oakes, L. (2023). Bayesian sample size planning for developmental studies. *Infant and Child Development*, e2412. doi: 10.1002/icd.2412

Visser, I., & Poessé, R. (2017). Parameter recovery, bias and standard errors in the linear ballistic accumulator model. *British Journal of Mathematical*

and Statistical Psychology, *70*(2), 280–296.

Visser, I., Raijmakers, M. E. J., & van der Maas, H. L. (2009). Hidden Markov models for individual time series. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 269–289). New York: Springer.

Visser, I., & Speekenbrink, M. (2010). depmixS4: An R package for hidden Markov models. *Journal of Statistical Software*, *36*(7), 1–21. doi: 10.18637/jss.v036.i07

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269. doi: 10.1109/TIT.1967.1054010

von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, *65*(2), 109–127.

Wabersich, D., & Vandekerckhove, J. (2014). The RWiener Package: an R Package Providing Distribution Functions for the Wiener Diffusion Model. *R Journal*, *6*(1), 49–56.

Wadehn, F., Weber, T., Mack, D. J., Heldt, T., & Loeliger, H.-A. (2020). Model-based separation, detection, and classification of eye movements. *IEEE Transactions on Biomedical Engineering*, *67*(2), 588–600. doi: 10.1109/TBME.2019.2918986

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi: 10.3758/BF03194105

Wagenmakers, E.-J. (2020). Bayesian Thinking for Toddlers. *PsyArXiv*. doi: 10.31234/osf.io/w5vbp

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q., Acosta, A., Adams, R. B., Jr., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917–928.

Wagenmakers, E.-J., Dutilh, G., & Sarafoglou, A. (2018). The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science*, *13*(4), 418–427. doi: 10.1177/1745691618771357

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*, *11*(1), 192–196. doi: 10.3758/BF03206482

Wagenmakers, E.-J., & Kucharský, Š. (2020). *The JASP Data Library*. Amsterdam. Retrieved from https://jasp-stats.org/wp-content/uploads/2020/05/The_JASP_Data_Library_1st_Edition.pdf

Wagenmakers, E.-J., Kucharský, Š., van den Bergh, D., & van Doorn, J. (2023). Accessible and Sustainable Statistics with JASP. *PsyArXiv*. doi: 10.31234/osf.io/ud2vj

Wagenmakers, E.-J., & Ly, A. (2023). History and nature of the Jeffreys–Lindley paradox. *Archive for History of Exact Sciences*, *77*, 25–72.

Wagenmakers, E.-J., & Matzke, D. (2023). *Bayesian Inference from the Ground Up: The Theory of Common Sense*. Amsterdam. Retrieved from https://www.bayesianspectacles.org/free-course-book/

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176.

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahnik, S., … Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*, 1473–1480.

Wagenmakers, E.-J., van Doorn, J., & van den Bergh, D. (in press). Toegankelijke en duurzame statistiek met JASP. *STAtOR*.

Wald, A., & Wolfowitz, J. (1948). Optimal Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, *19*(3), 326–339. doi: 10.1214/aoms/1177730197

Wang, M., Chen, F., Lu, T., & Dong, J. (2019). Bayesian t-tests for correlations and partial correlations. *Journal of Applied Statistics*, *47*(19), 1820–1832. doi: 10.1080/02664763.2019.1695760

Wang, X.-S., & Wong, R. (2007). Discrete analogues of Laplace's approximation. *Asymptotic Analysis*, *54*(3-4), 165–180.

Warnes, G. R., Bolker, B., & Lumley, T. (2020). *gtools: Various R Programming Tools*. Retrieved from https://CRAN.R-project.org/package=

gtools

Wass, S., Forssman, L., & Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy*, *19*(5), 427–460.

Wass, S., Smith, T., & Johnson, M. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229–250.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*, 129–133.

Waterman, M. (1981). Identification of common molecular subsequence. *Journal of Molecular Biology*, *147*, 195–197.

Watson, J., Richter, A., & Deeks, J. (2020). Testing for SARS-CoV-2 antibodies. *BMJ*, *370*, m3325.

Watson, J., Whiting, P. F., & Brush, J. E. (2020). Interpreting a COVID-19 test result. *BMJ*, *369*, m1808.

Weatherburn, C. E. (1961). *A First Course In Mathematical Statistics*. London: Cambridge University Press.

Welsh, M. B., & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes*, *119*(1), 1–14.

West, J. M., Haake, A. R., Rozanski, E. P., & Karn, K. S. (2006). EyePatterns: Software for identifying patterns and similarities across fixation sequences. In *Proceedings of the 2006 symposium on Eye tracking research & applications* (pp. 149–154). ACM.

Wetherford, M. J., & Cohen, L. B. (1973). Developmental changes in infant visual preferences for novelty and familiarity. *Child Development*, 416–424.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064. doi: 10.3758/s13423-012-0295-x

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85.

Wickham, H. (2011). testthat: Get Started with Testing. *The R Journal*, *3*, 5–10. Retrieved from https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *tidyverse: Easily Install and Load the 'Tidyverse'*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H., Hester, J., Chang, W., & Bryan, J. (2022). *devtools: Tools to Make Developing R Packages Easier*. Retrieved from https://CRAN.R-project.org/package=devtools

Williams, D. R. (2021). Bayesian estimation for Gaussian graphical models: Structure learning, predictability, and network comparisons. *Multivariate Behavioral Research*, *56*(2), 336–352.

Williams, D. R., & Mulder, J. (2019). BGGM: Bayesian Gaussian Graphical Models in R. *PsyArXiv*. doi: 10.31234/osf.io/t2cn7

Williams, D. R., & Mulder, J. (2020). Bayesian hypothesis testing for Gaussian graphical models: Conditional independence and order constraints. *Journal of Mathematical Psychology*, *99*, 102441.

Winkler, R. L., & Smith, J. E. (2004). On uncertainty in medical testing. *Medical Decision Making*, *24*(6), 654–658.

Xiao, N., Quinn, P., Liu, S., Ge, L., Pascalis, O., & Lee, K. (2015). Eye tracking reveals a crucial role for facial motion in recognition of faces by infants. *Developmental Psychology*, *51*(6), 744–757.

Xie, Y. (2014). knitr: A Comprehensive Tool for Reproducible Research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing Reproducible Computational Research*. Chapman and Hall/CRC.

Xie, Y. (2020). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. Retrieved from https://CRAN.R-project.org/package=knitr

Xu, F., Spelke, E., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, *8*(1), 88–101.

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predict-

ing human gaze beyond pixels. *Journal of Vision*, *14*(1), 28–28. doi: 10.1167/14.1.28

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision* (pp. 171–211). New York: Springer.

Yarkoni, T. (2019). The generalizability crisis. *PsyArXiv*. doi: 10.31234/osf.io/jqw35

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Young, D. S., & Hunter, D. R. (2015). Random effects regression mixtures for analyzing infant habituation. *Journal of Applied Statistics*, *42*(7), 1421–1441.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*(2), 268–282. doi: 10.3758/s13423-013-0495-z

Zacks, S. (2014). *Parametric statistical inference: Basic theory and modern approaches* (Vol. 4). Elsevier.

Zaharieva, M., Kucharský, Š., Colonnesi, C., Gu, T., Jo, S., Luttenbacher, I., … Visser, I. (2022). Habituation, Part I. Design choices in the infant habituation paradigm: A pre-registered crowd-sourced systematic review and meta-analysis. *PsyArXiv*. doi: 10.31234/osf.io/bdtx9

Zelinsky, G. J., Adeli, H., Peng, Y., & Samaras, D. (2013). Modelling eye movements in a categorical search task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1628), 20130058. doi: 10.1098/rstb.2013.0058

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (pp. 233–243). New York: Elsevier Science Publishers.

Zemblys, R., Niehorster, D. C., & Holmqvist, K. (2019). gazeNet: End-to-end eye-movement event detection with deep neural networks. *Behavior Research Methods*, *51*(2), 840–864. doi: 10.3758/s13428-018-1133-5

Zemblys, R., Niehorster, D. C., Komogortsev, O., & Holmqvist, K. (2018). Using machine learning to detect events in eye-tracking data. *Behavior*

Research Methods, *50*, 160–181. doi: 10.3758/s13428-017-0860-3

Zosh, J., Halberda, J., & Feigenson, L. (2011). Memory for multiple visual ensembles in infancy. *Journal of Experimental Psychology: General*, *140*(2), 141–158.

Zou, G. (2004). From diagnostic accuracy to accurate diagnosis: Interpreting a test result with confidence. *Medical Decision Making*, *24*(3), 313–318.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov models for time series - An introduction using R* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Zwet, E. v., & Gelman, A. (2022). A proposal for informative default priors scaled by the standard error of estimates. *The American Statistician*, *76*(1), 1–9. doi: 10.1080/00031305.2021.1938225

GAZING INTO A DISCRETE WORLD

MIXTURE MODELS OF COGNITION & BEHAVIOR

This book presents perspectives on modeling human behavior, focusing on alternative sources of data such as eye-tracking and response times, with a special focus dedicated to substantive questions about qualitative patterns in individual differences and development. Further, it advocates for a closer alignment between design and analysis of experiments, and their theoretical underpinnings.

The thesis contributes with advancements in integration of eye-tracking into cognitive-behavioral modeling, improvements in developmental psychology research, and provides openly available Bayesian tools that help researchers learn in the face of uncertainty.

ISBN: 978-94-93330-78-8