UNIVERSITY OF CALIFORNIA,
IRVINE

Using Hierarchical Bayesian Models to Test Complex Theories About the Nature of Latent Cognitive Processes

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Sciences

by

Beth Baribault

Dissertation Committee:
Associate Professor Joachim Vandekerckhove, Chair
Professor Jeffrey L. Krichmar
Professor Michael D. Lee

2019

DEDICATION

This dissertation is dedicated to my parents:

Mark Baribault
who fostered my love of science with trips to AMNH, "science stories," and water rockets,
and who pushed me to develop the coding skills that made this dissertation possible

&

Millie Baribault
who appreciated and loved that I "march to the beat of my own drum,"
who always knew I could do it even when I didn't, and who never, ever lost faith in me.

As I have said to you many times before, this was a family effort.
I absolutely could not have done it without you.
I love you.

This dissertation is furthermore dedicated to my husband
Gregory I. Telian
whom I love more than anything.

You quite literally made this dissertation possible, and I do not have the words to express how deeply, eternally grateful I am for your support the past few months (and years).

I am so lucky to have you in my life. You are my favorite person. I love you.

TABLE OF CONTENTS

Page

LIST OF FIGURES iv

LIST OF TABLES vi

ACKNOWLEDGMENTS vii

CURRICULUM VITAE viii

ABSTRACT OF THE DISSERTATION x

INTRODUCTION 1

CHAPTER 1: An Overview of Bayesian Statistics and Cognitive Modeling 7

CHAPTER 2: Using Hierarchical Modeling to Quantify the Robustness of Experimentally Observed Effects 50

CHAPTER 3: Cognitive Latent Variable Modeling: A Technique for Hierarchically Combining Cognitive and Psychometric Models 72

CHAPTER 4: A Hierarchical Bayesian Approach to Joint Modeling of Neural and Behavioral Data 132

LIST OF FIGURES

		Page
Figure 1.1	A reproduction of Figure 2 from Lindley (1993)	12
Figure 1.A1	An overview of focus and difficulty ratings for all sources included in the present paper	42
Figure 2.1	Level-2 variability	66
Figure 2.2	The sensitivity and robustness of the effect to two moderators	67
Figure 3.1	Attention Network Test	81
Figure 3.2	Continuous Performance Test	81
Figure 3.3	Number-Letter	82
Figure 3.4	Local-Global	82
Figure 3.5	Example items from each survey task	83
Figure 3.6	Schematic representation of the cognitive components	86
Figure 3.7	Cognitive model components	92
Figure 3.8	Psychometric model component	94
Figure 3.9	Covariate data model	95
Figure 3.10	Graphical models for each version of the CLVM	97
Figure 3.B1	Model specification for the LATER model version of the CLVM	112
Figure 3.B2	Model specification for the diffusion model version of the CLVM	113
Figure 4.1	Spatial discrimination task	137
Figure 4.2	Graphical model representation of the rate code model	140
Figure 4.3	Graphical model representation of the time-dependent rate code model	144
Figure 4.4	Raw data for units with different inferred preferences	148
Figure 4.5	Posterior distributions of preference parameters for all units in each of three session	149

Page

Figure 4.6 Proportion of units that are selective during different trial epochs 150

Figure 4.7 Posterior distributions of baseline bias parameters for each mouse 151

LIST OF TABLES

		Page
Table 2.1	Heterogeneity over facets	63
Table 3.1	Summary of theories under consideration	99
Table 3.2	Model comparison results	101
Table 3.A1	Descriptive statistics and Bayes factors	111

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my committee chair, Professor Joachim Vandekerckhove. One day in early 2014, I showed up in your office and all but demanded to be your grad student. That you accepted me into your lab that day changed the course of my career, and is the first of many, many things you have done for which I am tremendously grateful. Thank you for giving me the freedom to do the research I most wanted to do, and to go at my own pace when I needed it. Thank you for your support, for your guidance, and for giving me so many opportunities that I would never have had otherwise. I thank you for all this and for so, so much more — far more than can be contained here — but most of all, I thank you for being such a wonderful friend. I look forward to continuing our friendship — and hopefully, our collaborations — for many years yet to come. Thank you for everything.

I would also like to thank my committee members, Professors Jeffrey L. Krichmar and Michael D. Lee. I especially thank Dr. Krichmar, who taught me the core concepts and techniques of computational neuroscience, and in doing so, helped shape my current research interests. It was during your computational neuroscience course that I started work on the research presented in Chapter 4. I truly appreciate your feedback and continued support over the past few years.

I also thank my coauthors: Chris Donkin, Daniel R. Little, Zita Oravecz, Don van Ravenzwaaij, Paul De Boeck, Alexander Etz, Quentin F. Gronau, Fabian Dablander, and Peter A. Edelsbrunner. I would like to thank three coauthors in particular, Jennifer R. McCullen, Corey N. White, and Jennifer S. Trueblood, for their support and guidance at critical junctures in my career, and for their valued friendship.

Finally, I would like to thank many of my professors at SUNY Purchase College, who shaped me as a scientist in immeasurable ways. I thank Dr. Linda Bastone, who, through her passionate teaching of the first psychology course I ever took, inspired me to begin my career in this field; Dr. Bill Needham, who gave me my first exposure to research, my first opportunity to present research, and, unintentionally, a frisbee (sorry!); Dr. Karen Singer-Freeman, who taught me how to design exceptionally rigorous experiments; Dr. Nancy Zook, who guided me through the combined joy and horror of running my first independent research project, and who taught me about psychological testing, neuropsychology, and other concepts that inspired parts of the research described here; and and Dr. Lynn Winters, who taught me the most essential skill of all: how to write.

I thank the UC open access policy for ensuring I have the right to my own research, and therefore the ability to include Chapters 1 and 2 in my dissertation, which have previously been published as articles in *Psychonomics Bulletin & Review* and the *Proceedings of the National Academy of Sciences*, respectively. Financial support was provided by the University of California, Irvine and NSF Grants #1230118, #1534472, #1754205, and #1658303.

CURRICULUM VITAE
Beth Baribault

Education

PhD **Cognitive Sciences**	*June 2019*	University of California, Irvine
MA **Psychology**	*2017*	University of California, Irvine
OIST Computational Neuroscience Course	*2016*	Okinawa, Japan
Computational Modeling Summer School	*2014*	Laufen, Germany
BA **Psychology** Minor in Mathematics/Computer Science	*2011*	SUNY Purchase College

Publications

Peer-reviewed papers:

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., ... Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences.* doi:10.1073/pnas.1708285114

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P., & **Baribault, B.** (2017). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin and Review.* doi:10.3758/s13423-017-1317-5

In preparation:

Baribault, B. & Vandekerckhove, J. (*in prep*). A tutorial on cognitive latent variable modeling with a case example from attention research.

Baribault, B.*, Wilson, J.R.*, & Vandekerckhove, J. (*in prep*). Recommendations for designing rigorous tests of controversial claims: Lessons from a Bayesian analysis of applied kinesiology.

Honors & Awards

UC Irvine Associate Dean Fellowship	*2018*
Leamer-Rosenthal Prize for Open Social Science, Emerging Researcher ($10,000)	*2016*
Young Scientist Travel Award	*2016*
Women of Math Psych Travel & Networking Award	*2015*
UC Irvine Associated Graduate Students Travel Award	*2015*
SUNY Purchase Award for Outstanding Senior in Psychology	*2012*
SUNY Purchase Department of Psychology Travel Grant	*2012*
SUNY Purchase Undergraduate Research Support Award	*2011*

Presentations

Talks: (⋆ = Invited)

Psychonomic Society (New Orleans, LA)	*November, 2018*	
Brown University, Michael J. Frank Lab (Providence, RI)	*November, 2016*	⋆
Psychonomic Society, Mathematical Psychology workshop (Boston, MA)	*November, 2016*	⋆
Society for Mathematical Psychology (New Brunswick, NJ)	*August, 2016*	
UC Irvine, Cognitive Sciences Colloquium (Irvine, CA)	*January, 2016*	
UC Berkeley, Adnesnik Lab (Berkeley, CA)	*June, 2014*	⋆
Natural & Social Sciences Student Symposium (Purchase, NY)	*May, 2012*	

Presentations (continued)

Posters:	
Society for Neuroscience (San Diego, CA)	*November, 2016*
Psychonomic Society (Boston, MA)	*November, 2016*
Society for Mathematical Psychology (Newport Beach, CA)	*July, 2015*
Association for Psychological Science (New York, NY)	*May, 2015*
Association for Psychological Science (Chicago, IL)	*May, 2012*
Association for Psychological Science (Boston, MA)	*May, 2010*

Research Experience

Lead Researcher *UC Irvine*
Advisor: Dr. Joachim Vandekerckhove *2014–present*

- Initiated five lines of research and personally conducted all phases of the research including:
  - Hypothesis generation, experimental design, experiment programming, data preprocessing, data analysis, and manuscript writing
- Managed and mentored up to four Research Assistants at a time
  - Trained RAs in participant recruitment (from inside and outside UCI) and in data collection
- *Research topics included:* Models of decision-making; model-based comparative evaluation of theories of attention; joint models of behavioral and neural data; developing new methods for increasing replicability and quantifying robustness

Lead Researcher *SUNY Purchase*
Advisor: Dr. Nancy A. Zook *Jan.–Dec. 2011*

- Year-long independent research project on the role of attention in insight problem solving

Research Assistant *SUNY Purchase*
Advisor: Dr. Bill Needham *Aug.–Dec. 2009*

- Research on schematic memory as it applies to the reporting of psychological research

Professional Experience

Teaching Assistant *UC Irvine*

Experimental Psychology I, II, & III *2013–2018*

- Taught experimental design, data analysis, and scientific communication in an independent lab section
- Oversaw five undergraduate research groups each term in applying these principles to their own projects

Introduction to Psychology *2014*
Introduction to Human Memory *2015*

Skills

Languages:	MATLAB, Python
Packages & Programs:	JAGS, PsychoPy, OpenSesame, JASP, SPSS/PASW
OS:	Linux, Windows
Statistical Paradigms:	Classical/frequentist statistics, Bayesian statistics

Professional Memberships

Society for Neuroscience	*2016–present*
Psychonomic Society	*2016–present*
Society for Mathematical Psychology	*2014–present*
Association for Psychological Science	*2010–present*

ABSTRACT OF THE DISSERTATION

Using Hierarchical Bayesian Models to Test Complex Theories About the Nature of Latent Cognitive Processes

By

Beth Baribault

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2019

Professor Joachim Vandekerckhove, Chair

From a computational perspective, the primary goal of cognitive science is to infer the influence of unobservable psychological constructs on observed behavioral data. Cognitive models can facilitate this inference by more directly expressing a theoretical cognitive processes through relationships among psychologically interpretable parameters. If cognitive models are developed in a hierarchical Bayesian framework, significantly more nuanced and complex theories may be expressed, thereby allowing for deeper insight into the nature of latent cognitive processes. Here, I highlight three specific benefits of a hierarchical Bayesian approach to cognitive modeling, with a special emphasis on model-based theory testing. First, after a brief introduction to Bayesian methods, I discuss how hierarchical modeling permits one to coherently analyze data from highly complex experimental designs. To demonstrate this first benefit, I describe how the development of a hierarchical Bayesian metaregression model inspired a new technique for quantitative assessment of the robustness of a psychological theory. Next, I describe how hierarchical modeling enables simultaneous implementation of multiple different styles of cognitive models, which permits theory testing at a higher level of abstraction. I demonstrate this through a cognitive latent variable model-based comparison of theories of attention ability. Finally, I discuss how hierarchical modeling can be used to express highly complex neurocognitive processes, as demonstrated through a new approach to joint modeling of neural and behavioral data that is better suited to hypothesis testing than previous approaches. Ultimately, I contend that hierarchical Bayesian cognitive modeling is an ideal way to perform more powerful and informative analyses of behavioral data by radically expanding the scope of questions we may ask about the nature of latent cognitive processes.

INTRODUCTION

A primary goal of cognitive science research is to infer the influence of unobservable psychological constructs on human behavior. Often, verbal theories about the nature of these constructs will include a proposed mechanism through which the construct is expected to systematically shape behavior. In other words, many psychological theories propose a latent *cognitive process* and, from a computational perspective, the goal is to infer whether there is evidence of this process in the observed data.

While analysis of behavioral data using the standard toolbox of general linear models (such as ANOVA and regression) is common, this approach is a rather indirect way of testing process-based psychological theories. A more direct approach is to employ formal mathematical models of cognition, which are used to express theoretical cognitive processes in a quantitative way. In a *cognitive model*, a latent process is expressed through dependencies among model parameters, which typically have meaningful psychological interpretations. For example, diffusion models are designed to quantify a hypothetical decision-making process in which evidence is accumulated over time toward either of two possible choices. If the accumulated evidence crosses one of the two decision thresholds, the corresponding behavior is executed (see Chapter 3 for a more detailed description). This process is expressed through a likelihood function and four parameters, each of which captures a different dynamic of the evidence accumulation process.

Whether a cognitive model captures the data well can suggest whether the theory is a good approximation of the actual cognitive mechanism underlying people's behavior. This is one of many ways that cognitive models can aid in theory testing. Another way is to use cognitive model parameter estimates as the basis for inference: Assuming the model provides a good fit to the observed data, examining the inferred parameter values in the context of their semantic interpretations can provide additional insights about the nature of the latent cognitive process. Yet another way is to use cognitive models as proxies for verbal theories. If multiple cognitive models are developed, each of which expresses a different theoretical account of the latent cognitive process, performing a model comparison may serve as as a quantitative comparison of the theoretical accounts. If cognitive models are implemented in a Bayesian framework, these analyses also inherit the multitude of benefits offered by Bayesian methods, as outlined in Chapter 1.

If a Bayesian cognitive model is extended *hierarchically*, a much wider variety of potential theories may be expressed quantitatively, and therefore tested empirically. Hierarchical models are distinguished by their effort to mimic the structure of the data or of latent psychological constructs in the structure of the model. The simplest example of a hierarchical model is one that accounts for individual differences by introducing a hierarchy over participants. Rather than assuming that all data points were generated from the exact same latent process, a model that is hierarchical over participants posits that the data from each participant were generated from slightly different versions of the process. Consider the application of a diffusion model to data from participants who completed a simple 2AFC task. Rather than assuming that all data points were generated from the exact same Wiener process, a hierarchical diffusion model would now permit a different parameterization of the process for each participant, where each drift rate parameter, for example, is drawn from a hierarchical distribution of rates, which may have associated hyperparameters. By accounting for this structural aspect of the data, the model both becomes more flexible and offers a more realistic description of how people make decisions. In a similar fashion, models may be made hierarchical over conditions, sessions, or other known structural aspects of the data.

In Chapter 2, I introduce the concept of a hierarchy over planned experiments, and discuss how this hierarchical extension permits one to make novel qualitative conclusions about the robustness of a psychological theory. Specifically, I describe the *metastudy* approach, in which the same theory or hypothesis is tested in a large number of small experiments or *micro-experiments.* Unlike a meta-analysis of previously published research testing the same theory, which is likely to be a biased sample, in a metastudy, micro-experiments are systematically sampled without bias from a predefined space of possible experiments, where each dimension of the space is a possible moderating variable or *facet*. With this technique, the robustness of a theory to a large number of facets may be assessed simultaneously by observing whether the effect sizes are consistent across micro-experiments (i.e., are consistent across the space of possible experiments). If a theory is not robust, there may be a subset of micro-experiments in which the effect vanishes; in this case, the hierarchical nature of the model allows for the effect of each facet to be observed and genuine moderators to be identified, in a similar fashion to how one might analyze the strength of individual differences. Thus, the hierarchical extension over micro-experiments not only is a principled way to analyze data from highly complex experimental designs, but also allows for a satisfyingly direct test of a theory's generalizability.

A yet more interesting approach to introducing hierarchy in Bayesian cognitive models is to permit the generating process of a cognitive model parameter to be yet another model. Adding hierarchical structure in this way permits an increased depth of theoretical abstraction that is highly useful for expressing more nuanced theories. For example, this approach could be used to express a theory that describes nested cognitive processes. This general approach could also be used to describe how a small number of large-scale cognitive constructs exert a common effect on multiple different cognitive processes and thereby shape multiple different observed behaviors.

In Chapter 3, I offer a tutorial on cognitive latent variable models, which serve to quantify theories of exactly this type. Cognitive latent variables models (CLVMs) are a new class of model that join a cognitive process model with a psychometric model hierarchically to offer a single, unified account of heterogeneous data. In the case application of this approach described in Chapter 3, I use this technique to analyze data collected using a battery of established attention measures, including response time tasks and survey tasks. This procedure resulted in two qualitatively different types of data (response times for binary decisions, and Likert-scale responses in self-report scales). At a shallower level of the model, the data are described by either a formal cognitive model or a simpler likelihood distribution, depending on the data type. By imposing hierarchical structure (over tasks and participants) on one selected parameter type in each process, all of the parameters most closely related to attention ability are able to be collectively explained with a psychometric model. At this deepest level of abstraction, different theories of the nature of the psychological construct of attention may be expressed. As each theory under consideration was expressed in a different CLVM, comparing the fit across models effectively compared the viability of the theories, while still accounting for the latent processes that generated each type of data.

In this dissertation, the purpose of Chapter 3 is to emphasize the wide range of possibilities for theory testing that are created by hierarchically joining cognitive process models with latent variable models. A CLVM approach places a strong emphasis on the latent structure of cognitive constructs as they are expressed in cognitive process model parameters. Because the exact choice of model components is up to the user, this technique is a flexible approach to model construction. Through the choice of cognitive model components, CLVMs may be tailored to address different research contexts. Similarly, how one chooses to distinguish the different CLVMs under consideration, such as through the specification of the latent variable model component, or in the nature of the connection between model components, can enable different highly abstract theories about latent cognitive processes to be expressed. In this way, a CLVM-based model comparison may be used to address a multitude of different research questions.

I continued my exploration of methods for novel model-based comparisons of competing theories in my joint modeling work, as described in Chapter 4. Similar to the work described in Chapter 3, I fuse two different types of models — namely, a neural model and a behavioral model — in a hierarchical Bayesian framework. However, in Chapter 4, my goal is to develop models that each describe a single complex *neurocognitive* process in a comprehensive way. I call the theorized neurocognitive process *complex*, because it incorporates significant domain knowledge and multiple theoretical concepts in the structure of the model at different hierarchical levels. I demonstrated my approach by analyzing a published dataset from an experiment in which mice performed a behavioral task while electrophysiological data was simultaneously recorded. While the primary theory being tested was whether neurons encode the planned behavior in their average firing rate or in a time-dependent rate, the neurocognitive process in both accounts also incorporated abstract concepts including population-based encoding, individual differences across mice, and the binary nature of the behavioral response. Including this complexity in the analysis led to a conclusion about the proportion of neurons encoding the behavioral response that is qualitatively different from the conclusions of the dataset authors. This approach also allowed for entirely new types of conclusions to be made about the joint dataset. For example, I was able to infer individual biases in behavior across mice beyond that which was accounted for by the neural data. These biases were significant for individual mice, even though there was no purely behavioral bias evident across the population of mice.

I have called my approach *neurocognitive process modeling* to emphasize that the approach is designed to capture both neural and behavioral data with a single, unified latent process. This is unlike other hierarchical Bayesian approaches to joint modeling, where a neural model is the generative process for the neural data, a behavioral model is the generative process for the behavioral data, and both are influenced by a small number of latent abstract constructs. (Described thus, it is clear that these other approaches have a model architecture that is more similar to the work discussed in Chapter 3.) Because these other approaches place a strong emphasis on capturing abstract hierarchical correlations, they are an excellent approach for theory generation. However, if the goal is to perform confirmatory analyses, as is more often the goal of empirical research, I argue that the neurocogntive process modeling approach is preferable because it is more naturally suited to testing competing theories about hypothetical mechanisms and latent processes.

Ultimately, the unifying goal of the research described in this dissertation was to push the boundaries of what theories can be expressed in and tested using computational models. It is my contention that carefully designed hierarchical Bayesian cognitive models will become a dominant method for conducting theory testing in future psychological research due to the exceptional diversity of theories that this general approach is capable of capturing. In the chapters that follow, I attempt to make the case through example that we should favor the hierarchical Bayesian approach, as it radically expands the scope of questions we may ask and hypotheses that we may test concerning the nature of latent cognitive processes.

CHAPTER 1: AN OVERVIEW OF BAYESIAN STATISTICS AND COGNITIVE MODELING

This chapter was published in June 2017 as a peer-reviewed article in *Psychonomics Bulletin & Review*.[1] This invited paper was part of a special issue on Bayesian inference in psychology. As the senior author, my primary role in this work was to determine the goals, structure, and tone of the paper. I wrote all sections that serve to frame the paper (Abstract, Introduction, each section's introductory text, Conclusion), summarized the final source, and edited the paper.

How to become a Bayesian in eight easy steps: An annotated reading list

Alexander Etz[a], Quentin F. Gronau[b], Fabian Dablander[c], Peter A. Edelsbrunner[d], & Beth Baribault[a]

[a]University of California, Irvine

[b]University of Amsterdam

[c]University of Tübingen

[d]ETH Zürich

Abstract

In this guide, we present a reading list to serve as a concise introduction to Bayesian data analysis. The introduction is geared toward reviewers, editors, and interested researchers who are new to Bayesian statistics. We provide commentary for eight recommended sources, which together cover the theoretical and practical cornerstones of Bayesian statistics in psychology and

[1] `https://link.springer.com/content/pdf/10.3758%2Fs13423-017-1317-5.pdf`

related sciences. The resources are presented in an incremental order, starting with theoretical foundations and moving on to applied issues. In addition, we outline an additional 32 articles and books that can be consulted to gain background knowledge about various theoretical specifics and Bayesian approaches to frequently used models. Our goal is to offer researchers a starting point for understanding the core tenets of Bayesian analysis, while requiring a low level of time commitment. After consulting our guide, the reader should understand how and why Bayesian methods work, and feel able to evaluate their use in the behavioral and social sciences.

Introduction

In recent decades, significant advances in computational software and hardware have allowed Bayesian statistics to rise to greater prominence in psychology (Van de Schoot, Winder, Ryan, Zondervan-Zwijnenburg, & Depaoli, in press). In the past few years, this rise has accelerated as a result of increasingly vocal criticism of *p*-values in particular (Nickerson, 2000; Wagenmakers, 2007), and classical statistics in general (Trafimow & Marks, 2015). When a formerly scarcely used statistical method rapidly becomes more common, editors and peer reviewers are expected to master it readily, and to adequately evaluate and judge manuscripts in which the method is applied. However, many researchers, reviewers, and editors in psychology are still unfamiliar with Bayesian methods.

We believe that this is at least partly due to the perception that a high level of difficulty is associated with proper use and interpretation of Bayesian statistics. Many seminal texts in Bayesian statistics are dense, mathematically demanding, and assume some background in mathematical statistics (e.g., Gelman et al., 2013). Even texts that are geared toward psychologists (e.g., Lee & Wagenmakers, 2014; Kruschke, 2015), while less mathematically difficult, require a radically different way of thinking than the classical statistical methods most researchers are familiar with. Furthermore, transitioning to a Bayesian framework requires a level of time commitment that is not feasible for many researchers. More approachable sources that survey the core tenets and reasons for using Bayesian methods exist, yet identifying these sources can prove difficult for researchers with little or no previous exposure to Bayesian statistics.

In this guide, we provide a small number of primary sources that editors, reviewers, and other interested researchers can study to gain a basic understanding of Bayesian statistics. Each of these sources was selected for their balance of accessibility with coverage of essential Bayesian topics. By focusing on interpretation, rather than implementation, the guide is able to provide an introduction to core concepts, from Bayes' theorem through to Bayesian cognitive models, without getting mired in secondary details.

This guide is divided into two primary sections. The first, *Theoretical sources*, includes commentaries on three articles and one book chapter that explain the core tenets of Bayesian methods as well as their philosophical justification. The second, *Applied sources*, includes commentaries on four articles that cover the most commonly used methods in Bayesian data analysis at a primarily conceptual level. This section emphasizes issues of particular interest to reviewers, such as basic standards for conducting and reporting Bayesian analyses.

We suggest that for each source, readers first review our commentary, then consult the original source. The commentaries not only summarize the essential ideas discussed in each source, but also give a sense of how those ideas fit into the bigger picture of Bayesian statistics. This guide is part of a larger special issue in *Psychonomic Bulletin & Review* on the topic of Bayesian inference that contains articles which elaborate on many of the same points we discuss here, so we will periodically point to these as potential next steps for the interested reader. For those who would like to delve further into the theory and practice of Bayesian methods, the Appendix provides a number of supplemental sources that would be of interest to researchers and reviewers. To facilitate readers' selection of additional sources, each source is briefly described and has been given a rating by the authors that reflects its level of difficulty and general focus (i.e., theoretical versus applied; see Figure 1.A1). It is important to note that our reading list covers sources published up to the time of this writing (August, 2016).

Overall, the guide is designed such that a researcher might be able to read all eight of the highlighted articles[2] and some supplemental readings within a week. After readers acquaint themselves with these sources, they should be well-equipped both to interpret existing research and to evaluate new research that relies on Bayesian methods.

Theoretical sources

In this section, we discuss the primary ideas underlying Bayesian inference in increasing levels of depth. Our first source introduces *Bayes' theorem* and demonstrates how Bayesian statistics are based on a different conceptualization of probability than classical, or *frequentist*, statistics (Lindley, 1993). These ideas are extended in our second source's discussion of Bayesian inference as a reallocation of credibility (Kruschke, 2015) between possible states of nature. The third source demonstrates how the concepts established in the previous sources lead to many practical benefits for experimental psychology (Dienes, 2011). The section concludes with an in-depth review of Bayesian hypothesis testing using Bayes factors with an emphasis on this technique's theoretical benefits (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

1. Conceptual introduction: What is Bayesian inference?

Source: Lindley (1993) — The analysis of experimental data: The appreciation of tea and wine

Lindley leads with a story in which renowned statistician Ronald A. Fisher is having his colleague, Dr. Muriel Bristol, over for tea. When Fisher prepared the tea—as the story goes—Dr. Bristol protested that Fisher had made the tea all wrong. She claims that tea tastes better when milk is added first and infusion second,[3] rather than the other way around;

[2] Links to freely available versions of each article are provided in the *References* section.

[3] As a historical note: Distinguishing milk-first from infusion-first tea preparation was not a particular

she furthermore professes her ability to tell the difference. Fisher subsequently challenged Dr. Bristol to prove her ability to discern the two methods of preparation in a perceptual discrimination study. In Lindley's telling of the story, which takes some liberties with the actual design of the experiment in order to emphasize a point, Dr. Bristol correctly identified 5 out of 6 cups where the tea was added either first or second. This result left Fisher faced with the question: Was his colleague merely guessing, or could she really tell the difference? Fisher then proceeded to develop his now classic approach in a sequence of steps, recognizing at various points that tests that seem intuitively appealing actually lead to absurdities, until he arrived at a method that consists of calculating the total probability of the observed result plus the probability of any more extreme results possible under the null hypothesis (i.e., the probability that she would correctly identify 5 *or 6* cups by sheer guessing). This probability is the p-value. If it is less than .05, then Fisher would declare the result significant and reject the null hypothesis of guessing.

Lindley's paper essentially continues Fisher's work, showing that Fisher's classic procedure is inadequate and itself leads to absurdities because it hinges upon the nonexistent ability to define what other unobserved results would count as "more extreme" than the actual observations. That is, if Fisher had set out to serve Dr. Bristol 6 cups (and only 6 cups) and she is correct 5 times, then we get a p-value of .1, which is not statistically significant. According to Fisher, in this case we should not reject the null hypothesis that Dr. Bristol is guessing. But had he set out to keep giving her additional cups until she was correct 5 times, which incidentally required 6 cups, we get a p-value of .03, which is statistically significant. According to Fisher, we should now reject the null hypothesis. Even though the data observed in both cases are exactly the same, we reach different conclusions because our definition of "more extreme" results (that did not occur) changes depending on which sampling plan we use. Absurdly, the p-value, and with it our conclusion about Dr. Bristol's ability, depends on how we think about results that might have occurred but never

affectation of Dr. Bristol's, but a cultural debate that has persisted for over three centuries (e.g.; Orwell, 1946).

actually did, and that in turn depends on how we planned the experiment (rather than only on how it turned out).

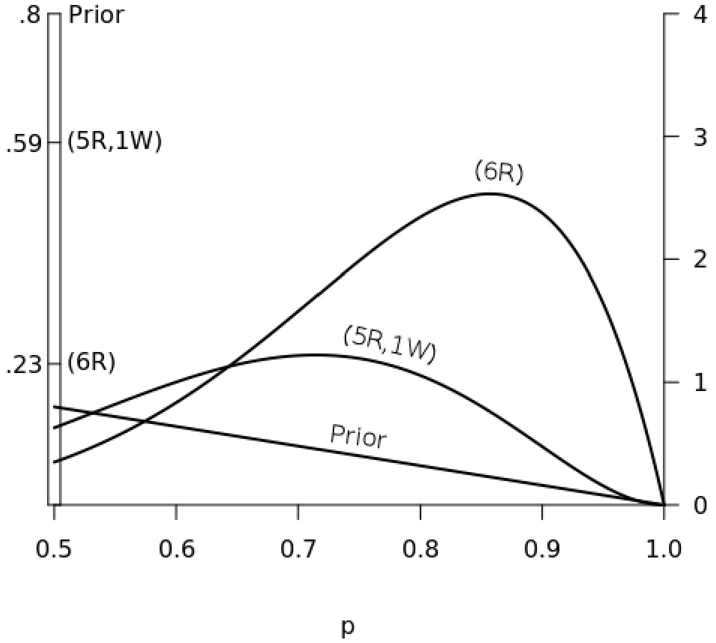


Figure 1.1. A reproduction of Figure 2 from Lindley (1993). The left bar indicates the probability that Dr. Bristol is guessing prior to the study (.8), if 5 right and 1 wrong are observed (.59), and if 6 right and 0 wrong are observed (.23). The lines represents Lindley's corresponding beliefs about Dr. Bristol's accuracy if she is not guessing.

Lindley's Bayesian solution to this problem considers only the probability of observations actually obtained, avoiding the problem of defining more extreme, unobserved results. The observations are used to assign a probability to each possible value of Dr. Bristol's success rate. Lindley's Bayesian approach to evaluating Dr. Bristol's ability to discriminate between the differently made teas starts by assigning a priori probabilities across the range of values of her success rate. If it is reasonable to consider that Dr. Bristol is simply guessing the outcome at random (i.e., her rate of success is .5), then one must assign an a priori probability to this null hypothesis (see our Figure 1, and note the separate amount of probability assigned to $p = .5$). The remaining probability is distributed among the range of other plausible values of Dr. Bristol's success rate (i.e., rates that do not assume that she is guessing at random)[4]. Then the observations are used to update these probabilities using *Bayes' rule* (this is derived in detail in Etz & Vandekerckhove, this issue). If the observations better fit with the null hypothesis (pure guessing), then the probability assigned to the null hypothesis will increase; if the data better fit the alternative hypothesis, then the probability assigned to the alternative hypothesis will increase, and subsequently the probability attached to the null hypothesis will decrease (note the decreasing probability of the null hypothesis on the left axis of Figure 2). The factor by which the data shift the balance of the hypotheses' probabilities is the *Bayes factor* (Kass & Raftery, 1995; see also Rouder et al., 2009, and Dienes, 2011, below).

A key takeaway from this paper is that Lindley's Bayesian approach depends only on the observed data, so the results are interpretable regardless of whether the sampling plan was rigid or flexible or even known at all. Another key point is that the Bayesian approach is inherently *comparative*: Hypotheses are tested against one another and never in isolation. Lindley further concludes that, since the posterior probability that the null is true will often be higher than the p-value, the latter metric will discount null hypotheses more easily in general.

2. Bayesian credibility assessments

Source: Kruschke (2015, Chapter 2) — Introduction: Credibility, models, and parameters

> "How often have I said to you that when all other θ yield $P(x|\theta)$ of 0, whatever remains, however low its $P(\theta)$, must have $P(\theta|x) = 1$?"
>
> – Sherlock Holmes, paraphrased

[4] If the null hypothesis is not initially considered tenable, then we can proceed without assigning separate probability to it and instead focus on estimating the parameters of interest (e.g., the taster's accuracy in distinguishing wines, as in Lindley's second example; see Lindley's Figure 1, and notice that the amount of probability assigned to $p = .5$ is gone). Additionally, if a range of values of the parameter is considered impossible—such as rates that are below chance—then this range may be given zero prior probability.

In this book chapter, Kruschke explains the fundamental Bayesian principle of *reallocation of probability*, or “credibility,” across possible states of nature. Kruschke uses an example featuring Sherlock Holmes to demonstrate that the famous detective essentially used Bayesian reasoning to solve his cases. Suppose that Holmes has determined that there exist only four different possible causes (A, B, C, and D) of a committed crime which, for simplicity in the example, he holds to be equally credible at the outset. This translates to equal *prior* probabilities for each of the four possible causes (i.e., a prior probability of 1/4 for each). Now suppose that Holmes gathers evidence that allows him to rule out cause A with certainty. This development causes the probability assigned to A to drop to zero, and the probability that used to be assigned to cause A to be then redistributed across the other possible causes. Since the probabilities for the four alternatives need to sum to one, the probability for each of the other causes is now equal to 1/3 (Figure 2.1, p. 17). What Holmes has done is reallocate credibility across the different possible causes based on the evidence he has gathered. His new state of knowledge is that only one of the three remaining alternatives can be the cause of the crime and that they are all equally plausible. Holmes, being a man of great intellect, is eventually able to completely rule out two of the remaining three causes, leaving him with only one possible explanation—which has to be the cause of the crime (as it now must have probability equal to 1), no matter how improbable it might have seemed at the beginning of his investigation.

The reader might object that it is rather unrealistic to assume that data can be gathered that allow a researcher to completely rule out contending hypotheses. In real applications, psychological data are noisy, and outcomes are only probabilistically linked to the underlying causes. In terms of reallocation of credibility, this means that possible hypotheses can rarely be ruled out completely (i.e., reduced to zero probability), however, their credibility can be greatly diminished, leading to a substantial increase in the credibility of other possible hypotheses. Although a hypothesis has not been eliminated, something has been learned: Namely, that one or more of the candidate hypotheses has had their probabilities reduced and are now less likely than the others.

In a statistical context, the possible hypotheses are parameter values in mathematical models that serve to describe the observed data in a useful way. For example, a scientist could assume that their observations are normally distributed and be interested in which range of values for the mean is most credible. Sherlock Holmes only considered a set of discrete possibilities, but in many cases it would be very restrictive to only allow a few alternatives (e.g., when estimating the mean of a normal distribution). In the Bayesian framework one can easily consider an infinite continuum of possibilities, across which credibility may still be reallocated. It is easy to extend this framework of reallocation of credibility to hypothesis testing situations where one parameter value is seen as "special" and receives a high amount of prior probability compared to the alternatives (as in Lindley's tea example above).

Kruschke (2015) serves as a good first introduction to Bayesian thinking, as it requires only basic statistical knowledge (a natural follow-up is Kruschke & Liddell, this issue). In this chapter, Kruschke also provides a concise introduction to mathematical models and parameters, two core concepts which our other sources will build on. One final key takeaway from this chapter is the idea of sequential updating from prior to posterior (Figure 2.1, p. 17) as data are collected. As Dennis Lindley famously said: "Today's posterior is tomorrow's prior" (Lindley, 1972, p. 2).

3. Implications of Bayesian statistics for experimental psychology

Source: Dienes (2011) — Bayesian versus orthodox statistics: Which side are you on?

Dienes explains several differences between the frequentist (which Dienes calls *orthodox* and we have called *classical*; we use these terms interchangeably) and Bayesian paradigm which have practical implications for how experimental psychologists conduct experiments, analyze data, and interpret results (a natural follow-up to the discussion in this section is available in Dienes & McLatchie, this issue). Throughout the paper, Dienes also discusses *subjective* (or context-dependent) Bayesian methods which allow for inclusion of relevant problem-specific knowledge in to the formation of one's statistical model.

The probabilities of data given theory and of theory given data. When testing a theory, both the frequentist and Bayesian approaches use probability theory as the basis for inference, yet in each framework, the interpretation of probability is different. It is important to be aware of the implications of this difference in order to correctly interpret frequentist and Bayesian analyses. One major contrast is a result of the fact that frequentist statistics only allow for statements to be made about $P(\text{data} \mid \text{theory})$[5]: Assuming the theory is correct, the probability of observing the obtained (or more extreme) data is evaluated. Dienes argues that often the probability of the data assuming a theory is correct is not the probability the researcher is interested in. What researchers typically want to know is $P(\text{theory} \mid \text{data})$: Given that the data were those obtained, what is the probability that the theory is correct? At first glance, these two probabilities might appear similar, but Dienes illustrates their fundamental difference with the following example: The probability that a person is dead (i.e., *data*) given that a shark has bitten the person's head off (i.e., *theory*) is 1. However, given that a person is dead, the probability that a shark has bitten this person's head off is very close to zero (see Senn, 2013, for an intuitive explanation of this distinction). It is important to keep in mind that a p-value does *not* correspond to $P(\text{theory} \mid \text{data})$; in fact, statements about this probability are only possible if one is willing to attach prior probabilities (degrees of plausibility or credibility) to theories—which can only be done in the Bayesian paradigm.

In the following sections, Dienes explains how the Bayesian approach is more liberating than the frequentist approach with regard to the following concepts: *stopping rules*, *planned versus post hoc comparisons*, and *multiple testing.* For those new to the Bayesian paradigm, these proposals may seem counterintuitive at first, but Dienes provides clear and accessible explanations for each.

[5] The conditional probability (P) of data given ($\mid$) theory.

Stopping rules. In the classical statistical paradigm, it is necessary to specify in advance how the data will be collected. In practice, one usually has to specify how many participants will be collected; stopping data collection early or continuing after the pre-specified number of participants has been reached is not permitted. One reason why collecting additional participants is not permitted in the typical frequentist paradigm is that, given the null hypothesis is true, the *p*-value is not driven in a particular direction as more observations are gathered. In fact, in many cases the distribution of the *p*-value is uniform when the null hypothesis is true, meaning that every *p*-value is equally likely under the null. This implies that even if there is no effect, a researcher is guaranteed to obtain a statistically significant result if they simply continue to collect participants and stop when the *p*-value is sufficiently low. In contrast, the Bayes factor, the most common Bayesian method of hypothesis testing, will approach infinite support in favor of the null hypothesis as more observations are collected if the null hypothesis is true. Furthermore, since Bayesian inference obeys the *likelihood principle*, one is allowed to continue or stop collecting participants at any time while maintaining the validity of one's results (p. 276; see also Cornfield, 1966, Rouder, 2014, and Royall, 2004 in the appended *Further Reading* section).

Planned versus post hoc comparisons. In the classical hypothesis-testing approach, a distinction is made between planned and post hoc comparisons: It matters whether the hypothesis was formulated before or after data collection. In contrast, Dienes argues that adherence to the likelihood principle entails that a theory does not necessarily need to precede the data when a Bayesian approach is adopted; since this temporal information does not enter into the likelihood function for the data, the evidence for or against the theory will be the same no matter its temporal relation to the data.

Multiple testing. When conducting multiple tests in the classical approach, it is important to correct for the number of tests performed (see Gelman & Loken, 2014). Dienes points out that within the Bayesian approach, the number of hypotheses tested does not matter—it is not the number of tests that is important, but the evaluation of how accurately each hypothesis predicts the observed data. Nevertheless, it is crucial to consider all relevant evidence, including so-called “outliers,” because “cherry picking is wrong on all statistical approaches” (Dienes, 2011, p. 280).

Context-dependent Bayes factors. The last part of the article addresses how problem-specific knowledge may be incorporated in the calculation of the Bayes factor. As is also explained in our next highlighted source (Rouder et al., 2009), there are two main schools of Bayesian thought: default (or *objective*) Bayes and context-dependent (or *subjective*) Bayes. In contrast to the default Bayes factors for general application that are designed to have certain desirable mathematical properties (e.g., Jeffreys, 1961; Rouder et al., 2009; Rouder & Morey, 2012; Rouder, Morey, Speckman, & Province, 2012; Ly, Verhagen, & Wagenmakers, 2016), Dienes provides an online calculator[6] that enables one to obtain context-dependent Bayes factors that incorporate domain knowledge for several commonly used statistical tests. In contrast to the default Bayes factors, which are typically designed to use standardized effect sizes, the context-dependent Bayes factors specify prior distributions in terms of the raw effect size. Readers who are especially interested in prior elicitation should see the appendix of Dienes’ article for a short review of how to appropriately specify prior distributions that incorporate relevant theoretical information (and Dienes, 2014, for more details and worked examples).

4. Structure and motivation of Bayes factors

Source: Rouder et al. (2009) — Bayesian t-tests for accepting and rejecting the null hypothesis

In many cases, a scientist’s primary interest is in showing evidence for an *invariance*, rather than a difference. For example, researchers may want to conclude that experimental and control groups do not differ in performance on a task (e.g., van Ravenzwaaij, Boekel, Forstmann, Ratcliff, & Wagenmakers, 2014), that participants were performing at chance

[6] `http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm`

(Dienes & Overgaard, 2015), or that two variables are unrelated (Rouder & Morey, 2012). In classical statistics this is generally not possible as significance tests are asymmetric; they can only serve to reject the null hypothesis and never to affirm it. One benefit of Bayesian analysis is that inference is perfectly symmetric, meaning evidence can be obtained that favors the null hypothesis as well as the alternative hypothesis (see Gallistel, 2009, as listed in our *Further Reading* appendix). This is made possible by the use of *Bayes factors*.[7] The section covering the shortcomings of classical statistics ("Critiques of Inference by Significance Tests") can safely be skipped, but readers particularly interested in the motivation of Bayesian inference are advised to read it.

What is a Bayes factor? The Bayes factor is a representation of the relative predictive success of two or more models, and it is a fundamental measure of relative evidence. The way Bayesians quantify predictive success of a model is to calculate the probability of the data given that model—also called the *marginal likelihood* or sometimes simply the *evidence*. The ratio of two such probabilities is the Bayes factor. Rouder and colleagues (2009) denote the probability of the data given some model, represented by H_i, as $f(\text{data} \mid H_i)$.[8] The Bayes factor for H_0 versus H_1 is simply the ratio of $f(\text{data} \mid H_0)$ and $f(\text{data} \mid H_1)$ written B_{01} (or BF_{01}), where the B (or BF) indicates a Bayes factor, and the subscript indicates which two models are being compared (see p. 228). If the result of a study is $B_{01} = 10$ then the data are ten times more probable under H_0 than under H_1. Researchers should report the exact value of the Bayes factor since it is a continuous measure of evidence, but various benchmarks have been suggested to help researchers interpret Bayes factors, with values between 1 and 3, between 3 and 10, and greater than 10 generally taken to indicate inconclusive, weak, and strong evidence, respectively (see Jeffreys, 1961; Wagenmakers, 2007; Etz & Vandekerckhove, 2016), although different researchers may set different benchmarks.

[7] Readers for whom Rouder and colleagues' (2009) treatment is too technical could focus on Dienes' conceptual ideas and motivations underlying the Bayes factor.

[8] The probability (f) of the observed data given ($|$) hypothesis i (H_i), where i indicates one of the candidate hypotheses (e.g., 0, 1, A, etc.). The null hypothesis is usually denoted H_0 and the alternative hypothesis is usually denoted either H_1 or H_A.

Care is need when interpreting Bayes factors against these benchmarks, as they are not meant to be bright lines against which we judge a study's success (as opposed to how a statistical significance criterion is sometimes treated); the difference between a Bayes factor of, say, 8 and 12 is more a difference of degree than of category. Furthermore, Bayes factors near 1 indicate the data are uninformative, and should not be interpreted as even mild evidence for either of the hypotheses under consideration.

Readers who are less comfortable with reading mathematical notation may skip over most of the equations without too much loss of clarity. The takeaway is that to evaluate which model is better supported by the data, we need to find out which model has done the best job predicting the data we observe. To a Bayesian, the probability a model assigns to the observed data constitutes its predictive success (see Morey, Romeijn, & Rouder, 2016); a model that assigns a high probability to the data relative to another model is best supported by the data. The goal is then to find the probability a given model assigns the data, $f(\text{data} \mid H_i)$. Usually the null hypothesis specifies that the true parameter is a particular value of interest (e.g., zero), so we can easily find $f(\text{data} \mid H_0)$. However, we generally do not know the value of the parameter if the null model is false, so we do not know what probability it assigns the data. To represent our uncertainty with regard to the true value of the parameter if the null hypothesis is false, Bayesians specify a range of plausible values that the parameter might take under the alternative hypothesis. All of these parameter values are subsequently used in computing an average probability of the data given the alternative hypothesis, $f(\text{data} \mid H_1)$ (for an intuitive illustration, see Gallistel, 2009 as listed in our *Further Reading* appendix). If the prior distribution gives substantial weight to parameter values that assign high probability to the data, then the average probability the alternative hypothesis assigns to the data will be relatively high—the model is effectively rewarded for its accurate predictions with a high value for $f(\text{data} \mid H_1)$.

The role of priors. The form of the prior can have important consequences on the resulting Bayes factor. As discussed in our third source (Dienes, 2011), there are two primary schools of Bayesian thought: default (objective) Bayes (Berger, 2006) and context-dependent (subjective) Bayes (Goldstein et al., 2006; Rouder, Morey, & Wagenmakers, 2016). The default Bayesian tries to specify prior distributions that convey little information while maintaining certain desirable properties. For example, one desirable property is that changing the scale of measurement should not change the way the information is represented in the prior, which is accomplished by using standardized effect sizes. Context-dependent prior distributions are often used because they more accurately encode our prior information about the effects under study, and can be represented with raw or standardized effect sizes, but they do not necessarily have the same desirable mathematical properties (although sometimes they can).

Choosing a prior distribution for the standardized effect size is relatively straightforward for the default Bayesian. One possibility is to use a normal distribution centered at 0 and with some standard deviation (i.e., spread) σ. If σ is too large, the Bayes factor will always favor the null model, so such a choice would be unwise (see also DeGroot, 1982; Robert, 2014). This happens because such a prior distribution assigns weight to very extreme values of the effect size, when in reality, the effect is most often reasonably small (e.g., almost all psychological effects are smaller than Cohen's $d = 2$). The model is penalized for low predictive success. Setting σ to 1 is reasonable and common—this is called the *unit information prior*. However, using a Cauchy distribution (which resembles a normal distribution but with less central mass and fatter tails) has some better properties than the unit information prior, and is now a common default prior on the alternative hypothesis, giving rise to what is now called the *default Bayes factor* (see Rouder & Morey, 2012 for more details; see also Wagenmakers, Love, et al., this issue and Wagenmakers, Marsman, et al., this issue). To use the Cauchy distribution, like the normal distribution, again one must specify a scaling factor. If it is too large, the same problem as before occurs where the null model will always be favored. Rouder and colleagues suggest a scale of 1, which implies that the effect size has a prior probability of 50% to be between $d = -1$ and $d = 1$. For some areas, such as social psychology, this is not reasonable, and the scale should be reduced. However, slight changes to the scale often do not make much difference in the qualitative conclusions one draws.

Readers are advised to pay close attention to the sections “Subjectivity in priors” and “Bayes factors with small effects.” The former explains how one can tune the scale of the default prior distribution to reflect more contextually relevant information while maintaining the desirable properties attached to prior distributions of this form, a practice that is a reasonable compromise between the default and context-dependent schools. The latter shows why the Bayes factor will often show evidence in favor of the null hypothesis if the observed effect is small and the prior distribution is relatively diffuse.

Applied sources

At this point, the essential concepts of Bayesian probability, Bayes’ theorem, and the Bayes factor have been discussed in depth. In the following four sources, these concepts are applied to real data analysis situations. Our first source provides a broad overview of the most common methods of model comparison, including the Bayes factor, with a heavy emphasis on its proper interpretation (Vandekerckhove, Matzke, & Wagenmakers, 2015). The next source begins by demonstrating Bayesian estimation techniques in the context of developmental research, then provides some guidelines for reporting Bayesian analyses (van de Schoot et al., 2014). Our final two sources discuss issues in Bayesian cognitive modeling, such as the selection of appropriate priors (Lee & Vanpaemel, this issue), and the use of cognitive models for theory testing (Lee, 2008).

Before moving on to our final four highlighted sources, it will be useful if readers consider some differences in perspective among practitioners of Bayesian statistics. The application of Bayesian methods is very much an active field of study, and as such, the literature contains a multitude of deep, important, and diverse viewpoints on how data analysis should be done, similar to the philosophical divides between Neyman–Pearson and Fisher concerning proper application of classical statistics (see Lehmann, 1993). The divide between subjective Bayesians, who elect to use priors informed by theory, and objective Bayesians, who instead prefer "uninformative" or default priors, has already been mentioned throughout the *Theoretical sources* section above.

A second division of note exists between Bayesians who see a place for hypothesis testing in science, and those who see statistical inference primarily as a problem of estimation. The former believe statistical models can stand as useful surrogates for theoretical positions, whose relative merits are subsequently compared using Bayes factors and other such "scoring" metrics (as reviewed in Vandekerckhove et al., 2015, discussed below; for additional examples, see Jeffreys, 1961 and Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016). The latter would rather delve deeply into a single model or analysis and use point estimates and credible intervals of parameters as the basis for their theoretical conclusions (as demonstrated in Lee, 2008, discussed below; for additional examples, see Gelman & Shalizi, 2013 and McElreath, 2016).[9]

Novice Bayesians may feel surprised that such wide divisions exist, as statistics (of any persuasion) is often thought of as a set of prescriptive, immutable procedures that can be only right or wrong. We contend that debates such as these should be expected due to the wide variety of research questions—and diversity of contexts—to which Bayesian methods are applied. As such, we believe that the existence of these divisions speaks to the intellectual vibrancy of the field and its practitioners. We point out these differences here so that readers might use this context to guide their continued reading.

5. Bayesian model comparison methods

Source: Vandekerckhove et al. (2015) — Model comparison and the principle of parsimony

John von Neumann famously said: "With four parameters I can fit an elephant, and

[9] This divide in Bayesian statistics may be seen as a parallel to the recent discussions about use of classical statistics in psychology (e.g., Cumming, 2014), where a greater push has been made to adopt an estimation approach over null hypothesis significance testing (NHST). Discussions on the merits of hypothesis testing have been running through all of statistics for over a century, with no end in sight.

with five I can make him wiggle his trunk" (as quoted in Mayer, Khairy, & Howard, 2010, p. 698), pointing to the natural tension between model parsimony and goodness of fit. The tension occurs because it is always possible to decrease the amount of error between a model's predictions and the observed data by simply adding more parameters to the model. In the extreme case, any data set of N observations can be reproduced perfectly by a model with N parameters. Such practices, however, termed *overfitting*, result in poor generalization and greatly reduce the accuracy of out-of-sample predictions. Vandekerckhove and colleagues (2015) take this issue as a starting point to discuss various criteria for model selection. How do we select a model that both fits the data well and generalizes adequately to new data?

Putting the problem in perspective, the authors discuss research on recognition memory that relies on multinomial processing trees, which are simple, but powerful, cognitive models. Comparing these different models using only the likelihood term is ill-advised, because the model with the highest number of parameters will—all other things being equal—yield the best fit. As a first step to addressing this problem, Vandekerckhove et al. (2015) discuss the popular Akaike information criterion (AIC) and Bayesian information criterion (BIC).

Though derived from different philosophies (for an overview, see Aho, Derryberry, & Peterson, 2014), both AIC and BIC try to solve the trade-off between goodness-of-fit and parsimony by combining the likelihood with a penalty for model complexity. However, this penalty is solely a function of the number of parameters and thus neglects the functional form of the model, which can be informative in its own right. As an example, the authors mention Fechner's law and Steven's law. The former is described by a simple logarithmic function, which can only ever fit negatively accelerated data. Steven's law, however, is described by an exponential function, which can account for both positively and negatively accelerated data. Additionally, both models feature just a single parameter, nullifying the benefit of the complexity penalty in each of the two aforementioned information criteria.

The Bayes factor yields a way out. It extends the simple likelihood ratio test by integrating the likelihood with respect to the prior distribution, thus taking the predictive success of the prior distribution into account (see also Gallistel, 2009, in the *Further Reading* appendix). Essentially, the Bayes factor is a likelihood ratio test averaged over all possible parameter values for the model, using the prior distributions as weights: It is the natural extension of the likelihood ratio test to a Bayesian framework. The net effect of this is to penalize complex models. While a complex model can predict a wider range of possible data points than a simple model can, each individual data point is less likely to be observed under the complex model. This is reflected in the prior distribution being more spread out in the complex model. By weighting the likelihood by the corresponding tiny prior probabilities, the Bayes factor in favor of the complex model decreases. In this way, the Bayes factor instantiates an automatic Ockham's Razor (see also Myung & Pitt, 1997, in the appended *Further Reading* section).

However, the Bayes factor can be difficult to compute because it often involves integration over very many dimensions at once. Vandekerckhove and colleagues (2015) advocate two methods to ease the computational burden: importance sampling and the Savage-Dickey density ratio (see also Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010 in our in our *Further reading* appendix); additional common computational methods include the Laplace approximation (Kass & Raftery, 1995), bridge sampling (Meng & Wong, 1996; Gronau et al., 2017), and the encompassing prior approach (Hoijtink, Klugkist, & Boelen, 2008). They also provide code to estimate parameters in multinomial processing tree models and to compute the Bayes factor to select among them. Overall, the chapter provides a good overview of different methods used to tackle the tension between goodness-of-fit and parsimony in a Bayesian framework. While it is more technical then the sources reviewed above, this article can greatly influence how one thinks about models and methods for selecting among them.

6. Bayesian estimation

Source: van de Schoot et al. (2014) — A gentle introduction to Bayesian analysis: Applications to developmental research

This source approaches practical issues related to parameter estimation in the context of developmental research. This setting offers a good basis for discussing the choice of priors and how those choices influence the posterior estimates for parameters of interest. This is a topic that matters to reviewers and editors alike: How does the choice of prior distributions for focal parameters influence the statistical results and theoretical conclusions that are obtained? The article discusses this issue on a basic and illustrative level.

At this point we feel it is important to note that the difference between hypothesis testing and estimation in the Bayesian framework is much greater than it is in the frequentist framework. In the frequentist framework there is often a one-to-one relationship between the null hypothesis falling outside the sample estimate's 95% confidence interval and rejection of the null hypothesis with a significance test (e.g., when doing a *t*-test). This is not so in the Bayesian framework; one cannot test a null hypothesis by simply checking if the null value is inside or outside a credible interval. A detailed explanation of the reason for this deserves more space than we can afford to give it here, but in short: When testing hypotheses in the Bayesian framework one should calculate a model comparison metric. See Rouder and Vandekerckhove (this issue) for an intuitive introduction to (and synthesis of) the distinction between Bayesian estimation and testing.

Van de Schoot and colleagues (2014) begin by reviewing the main differences between frequentist and Bayesian approaches. Most of this part can be skipped by readers who are comfortable with basic terminology at that point. The only newly introduced term is *Markov chain Monte Carlo (MCMC)* methods, which refers to the practice of drawing samples from the posterior distribution instead of deriving the distribution analytically (which may not be feasible for many models; see also van Ravenzwaaij, Cassey, & Brown, this issue and Matzke, Boehm, & Vandekerckhove, this issue). After explaining this alternative approach (p. 848), Bayesian estimation of focal parameters and the specification of prior distributions is discussed with the aid of two case examples.

The first example concerns estimation of an ordinary mean value and the variance of reading scores and serves to illustrate how different sources of information can be used to inform the specification of prior distributions. The authors discuss how expert domain knowledge (e.g., reading scores usually fall within a certain range), statistical considerations (reading scores are normally distributed), and evidence from previous studies (results obtained from samples from similar populations) may be jointly used to define adequate priors for the mean and variance model parameters. The authors perform a prior sensitivity analysis to show how using priors based on different considerations influence the obtained results. Thus, the authors examine and discuss how the posterior distributions of the mean and variance parameters are dependent on the prior distributions used.

The second example focuses on a data set from research on the longitudinal reciprocal associations between personality and relationships. The authors summarize a series of previous studies and discuss how results from these studies may or may not inform prior specifications for the latest obtained data set. Ultimately, strong theoretical considerations are needed to decide whether data sets that were gathered using slightly different age groups can be used to inform inferences about one another.

The authors fit a model with data across two time points and use it to discuss how convergence of the MCMC estimator can be supported and checked. They then evaluate overall model fit via a posterior predictive check. In this type of model check, data simulated from the specified model are compared to the observed data. If the model is making appropriate predictions, the simulated data and the observed data should appear similar. The article concludes with a brief outline of guidelines for reporting Bayesian analyses and results in a manuscript. Here, the authors emphasize the importance of the specification of prior distributions and of convergence checks (if MCMC sampling is used) and briefly outline how both might be reported. Finally, the authors discuss the use of default priors and various options for conducting Bayesian analyses with common software packages (such as Mplus and WinBUGS).

The examples in the article illustrate different considerations that should be taken into account for choosing prior specifications, the consequences they can have on the obtained results, and how to check whether and how the choice of priors influenced the resulting inferences.

7. Prior elicitation

Source: Lee and Vanpaemel (this issue) — Determining priors for cognitive models

Statistics does not operate in a vacuum, and often prior knowledge is available that can inform one's inferences. In contrast to classical statistics, Bayesian statistics allows one to formalize and use this prior knowledge for analysis. The paper by Lee and Vanpaemel (this issue) fills an important gap in the literature: What possibilities are there to formalize and uncover prior knowledge?

The authors start by noting a fundamental point: Cognitive modeling is an extension of general purpose statistical modeling (e.g., linear regression). Cognitive models are designed to instantiate theory, and thus may need to use richer information and assumptions than general purpose models (see also Franke, 2016). A consequence of this is that the prior distribution, just like the likelihood, should be seen as an integral part of the model. As Jaynes (2003) put it: "If one fails to specify the prior information, a problem of inference is just as ill-posed as if one had failed to specify the data" (p. 373).

What information can we use to specify a prior distribution? Because the parameters in such a cognitive model usually have a direct psychological interpretation, theory may be used to constrain parameter values. For example, a parameter interpreted as a probability of correctly recalling a word must be between 0 and 1. To make this point clear, the authors discuss three cognitive models and show how the parameters instantiate relevant information about psychological processes. Lee and Vanpaemel also discuss cases in which all of the theoretical content is carried by the prior, while the likelihood does not make any strong assumptions. They also discuss the principle of *transformation invariance*, that is, prior distributions for parameters should be invariant to the scale they are measured on (e.g., measuring reaction time using seconds versus milliseconds).

Lee and Vanpaemel also discuss specific methods of prior specification. These include the maximum entropy principle, the prior predictive distribution, and hierarchical modeling. The prior predictive distribution is the model-implied distribution of the data, weighted with respect to the prior. Recently, iterated learning methods have been employed to uncover an implicit prior held by a group of participants. These methods can also be used to elicit information that is subsequently formalized as a prior distribution. (For a more in-depth discussion of hierarchical cognitive modeling, see Lee, 2008, discussed below.)

In sum, the paper gives an excellent overview of why and how one can specify prior distributions for cognitive models. Importantly, priors allow us to integrate domain-specific knowledge, and thus build stronger theories (Platt, 1964; Vanpaemel, 2010). For more information on specifying prior distributions for data-analytic statistical models rather than cognitive models see Rouder, Morey, Verhagen, Swagman, and Wagenmakers (in press) and Rouder, Engelhardt, McCabe, and Morey (2016).

8. Bayesian cognitive modeling

Source: Lee (2008) — Three case studies in the Bayesian analysis of cognitive models

Our final source (Lee, 2008) further discusses cognitive modeling, a more tailored approach within Bayesian methods. Often in psychology, a researcher will not only expect to observe a particular effect, but will also propose a verbal theory of the cognitive process underlying the expected effect. Cognitive models are used to formalize and test such verbal theories in a precise, quantitative way. For instance, in a cognitive model, psychological constructs, such as attention and bias, are expressed as model parameters. The proposed psychological process is expressed as dependencies among parameters and observed data (the "structure" of the model).

In peer-reviewed work, Bayesian cognitive models are often presented in visual form as a graphical model. Model parameters are designated by nodes, where the shape, shading, and style of border of each node reflect various parameter characteristics. Dependencies among parameters are depicted as arrows connecting the nodes. Lee gives an exceptionally clear and concise description of how to read graphical models in his discussion of multidimensional scaling (Lee, 2008, p. 2).

After a model is constructed, the observed data are used to update the priors and generate a set of posterior distributions. Because cognitive models are typically complex, posterior distributions are almost always obtained through sampling methods (i.e., MCMC; see van Ravenzwaaij et al., this issue), rather than through direct, often intractable, analytic calculations.

Lee demonstrates the construction and use of cognitive models through three case studies. Specifically, he shows how three popular process models may be implemented in a Bayesian framework. In each case, he begins by explaining the theoretical basis of each model, then demonstrates how the verbal theory may be translated into a full set of prior distributions and likelihoods. Finally, Lee discusses how results from each model may be interpreted and used for inference.

Each case example showcases a unique advantage of implementing cognitive models in a Bayesian framework (see also Bartlema, Voorspoels, Rutten, Tuerlinckx, & Vanpaemel, this issue). For example, in his discussion of signal detection theory, Lee highlights how Bayesian methods are able to account for individual differences easily (see also Rouder & Lu, 2005, in the *Further reading* appendix). Throughout, Lee emphasizes that Bayesian cognitive models are useful because they allow the researcher to reach new theoretical conclusions that would be difficult to obtain with non-Bayesian methods. Overall, this source not only provides an approachable introduction to Bayesian cognitive models, but also provides an excellent example of good reporting practices for research that employs Bayesian cognitive models.

Conclusion

By focusing on interpretation, rather than implementation, we have sought to provide a more accessible introduction to the core concepts and principles of Bayesian analysis than may be found in introductions with a more applied focus. Ideally, readers who have read through all eight of our highlighted sources, and perhaps some of the supplementary reading, may now feel comfortable with the fundamental ideas in Bayesian data analysis, from basic principles (Kruschke, 2015; Lindley, 1993) to prior distribution selection (Lee & Vanpaemel, this issue), and with the interpretation of a variety of analyses, including Bayesian analogs of classical statistical tests (e.g., t-tests; Rouder et al., 2009), estimation in a Bayesian framework (van de Schoot et al., 2014), Bayes factors and other methods for hypothesis testing (Dienes, 2011; Vandekerckhove et al., 2015), and Bayesian cognitive models (Lee, 2008).

Reviewers and editors unfamiliar with Bayesian methods may initially feel hesitant to evaluate empirical articles in which such methods are applied (Wagenmakers, Love, et al., this issue). Ideally, the present article should help ameliorate this apprehension by offering an accessible introduction to Bayesian methods that is focused on interpretation rather than application. Thus, we hope to help minimize the amount of reviewer reticence caused by authors' choice of statistical framework.

Our overview was not aimed at comparing the advantages and disadvantages of Bayesian and classical methods. However, some conceptual conveniences and analytic strategies that are only possible or valid in the Bayesian framework will have become evident. For example, Bayesian methods allow for the easy implementation of hierarchical models for complex data structures (Lee, 2008), they allow multiple comparisons and flexible sampling rules during data collection without correction of inferential statistics (Dienes, 2011; see also Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015, as listed in our *Further reading* appendix, and also Schönbrodt & Wagenmakers, this issue), and they allow inferences that many researchers in psychology are interested in but are not able to answer with classical statistics such as providing support for a null hypothesis (for a discussion, see Wagenmakers, 2007). Thus, the inclusion of more research that uses Bayesian methods in the psychological literature should be to the benefit of the entire field (Etz & Vandekerckhove, 2016). In this article, we have provided an overview of sources that should allow a novice to understand how Bayesian statistics allows for these benefits, even without prior knowledge of Bayesian methods.

Acknowledgments

The authors would like to thank Jeff Rouder, E.-J. Wagenmakers, and Joachim Vandekerckhove for their helpful comments. AE and BB were supported by grant #1534472 from NSF's Methods, Measurements, and Statistics panel. AE was further supported by the National Science Foundation Graduate Research Fellowship Program (#DGE1321846).

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of aic and bic. *Ecology*, *95*(3), 631–636. Retrieved from `http://tinyurl.com/aho2014` doi: dx.doi.org/10.1890/13-1452.1

Bartlema, A., Voorspoels, W., Rutten, F., Tuerlinckx, F., & Vanpaemel, W. (this issue). Sensitivity to the prototype in children with high-functioning autism spectrum disorder: An example of Bayesian cognitive psychometrics. *Psychonomic Bulletin and Review.*

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian analysis*, *1*(3), 385–402. Retrieved from `http://projecteuclid.org/euclid.ba/1340371035` doi: 10.1214/06-BA115

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*(2), 159–165. Retrieved from `http://www.jstor.org/stable/27855070`

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 317–335. Retrieved from `https://projecteuclid.org/euclid.ss/1177013238`

Cornfield, J. (1966). Sequential trials, sequential analysis, and the likelihood principle. *The American Statistician*, *20*, 18–23. Retrieved from http://www.jstor.org/stable/2682711

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, *25*(1), 7–29. Retrieved from http://pss.sagepub.com/content/25/1/7 doi: 10.1177/0956797613504966

DeGroot, M. H. (1982). Lindley's paradox: Comment. *Journal of the American Statistical Association*, 336–339. Retrieved from http://www.jstor.org/stable/2287246

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. Retrieved from http://tinyurl.com/dienes2011

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. Retrieved from http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00781/full

Dienes, Z., & McLatchie, N. (this issue). Four reasons to prefer Bayesian over orthodox statistical analyses. *Psychonomic Bulletin and Review*.

Dienes, Z., & Overgaard, M. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. *Behavioural methods in consciousness research*, 199–220. Retrieved from http://tinyurl.com/dienes2015

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychology research. *Psychological Review*, *70*(3), 193–242. Retrieved from http://tinyurl.com/edwards1963

Etz, A., & Vandekerckhove, J. (2016).

PLOS ONE, *11*, e0149794. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pone.0149794 doi: 10.1371/journal.pone.0149794

Etz, A., & Vandekerckhove, J. (this issue). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin and Review*.

Etz, A., & Wagenmakers, E.-J. (in press). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*.

Franke, M. (2016). Task types, link functions & probabilistic modeling in experimental pragmatics. In F. Salfner & U. Sauerland (Eds.), *Preproceedings of 'trends in experimental pragmatics'* (pp. 56–63).

Gallistel, C. (2009). The importance of proving the null. *Psychological review*, *116*(2), 439. Retrieved from `http://tinyurl.com/gallistel`

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (Vol. 3). Chapman & Hall/CRC.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460. Retrieved from `http://tinyurl.com/gelman2014`

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. Retrieved from `http://tinyurl.com/gelman2013` doi: 10.1111/j.2044-8317.2011.02037.x

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. Retrieved from `http://tinyurl.com/gigerenzer2004` doi: 10.1016/j.socec.2004.09.033

Goldstein, M., et al. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*(3), 403–420. Retrieved from `http://projecteuclid.org/euclid.ba/1340371036` doi: 10.1214/06-BA116

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *arXiv preprint arXiv:1703.05984*.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. Springer Science & Business Media.

Jaynes, E. T. (1986). Bayesian methods: General background. In J. H. Justice (Ed.), *Maximum entropy and bayesian methods in applied statistics* (pp. 1–25). Cambridge University Press. Retrieved from http://tinyurl.com/jaynes1986

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge university press.

Jeffreys, H. (1936). Xxviii. on some criticisms of the theory of probability. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *22*(146), 337–359. Retrieved from http://www.tandfonline.com/doi/pdf/10.1080/14786443608561691 doi: 10.1080/14786443608561691

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). Guilford New York, NY. Retrieved from http://tinyurl.com/kaplan2012

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. Retrieved from http://tinyurl.com/KassRaftery

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press. Retrieved from http://tinyurl.com/kruschke2015

Kruschke, J. K., & Liddell, T. (this issue). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review.*

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin and Review*, *15*(1), 1–15. Retrieved from http://tinyurl.com/lee2008cognitive

Lee, M. D., & Vanpaemel, W. (this issue). Determining priors for cognitive models. *Psychonomic Bulletin & Review.* Retrieved from https://webfiles.uci.edu/mdlee/LeeVanpaemel2016.pdf

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Lehmann, E. (1993). The fisher, neyman-pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, *88*(424), 1242–1249.

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. Retrieved from `http://tinyurl.com/lindley1993` doi: 10.1111/j.1467-9639.1993.tb00252.x

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, *49*(3), 293–337. Retrieved from `http://tinyurl.com/lindley2000`

Lindley, D. V. (2006). *Understanding uncertainty.* John Wiley & Sons.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... Wagenmakers, E.-J. (2015). JASP (version 0.7.1.12). *Computer Software.*

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. Retrieved from `http://tinyurl.com/zyvgp9y`

Matzke, D., Boehm, U., & Vandekerckhove, J. (this issue). Bayesian inference for psychology, Part III: Parameter estimation in nonstandard models. *Psychonomic Bulletin and Review.*

Mayer, J., Khairy, K., & Howard, J. (2010). Drawing an elephant with four complex parameters. *American Journal of Physics*, *78*(6), 648–649. Retrieved from `http://tinyurl.com/gtz9w3q`

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Vol. 122). CRC Press.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology.* Retrieved from `http://tinyurl.com/BFphilo`

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95. Retrieved from http://tinyurl.com/myung1997 doi: 10.3758/BF03210778

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, *5*(2), 241. Retrieved from http://tinyurl.com/nickerson2000 doi: 10.1037//1082-989X.S.2.241

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi: 10.1126/science.aac4716

Orwell, G. (1946). A nice cup of tea. *Evening Standard, January*.

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353.

Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, *81*(2), 216–232. Retrieved from http://www.jstor.org/stable/10.1086/675729

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. Retrieved from http://tinyurl.com/rouder2014 doi: 10.3758/s13423-014-0595-4

Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in anova. *Psychonomic Bulletin & Review*, *23*, 1779-1786.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. Retrieved from http://tinyurl.com/rouder2005

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903. Retrieved from http://tinyurl.com/rouder2012regression doi: 10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. Retrieved from http://tinyurl.com/rouder2012an

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520-547. Retrieved from http://tinyurl.com/jjubz9y

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (in press). Bayesian analysis of factorial designs. *Psychological Methods.* Retrieved from http://tinyurl.com/zh4bkt8

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*(1). Retrieved from http://www.collabra.org/article/10.1525/collabra.28/

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 and Review*, *16*(2), 225–237. Retrieved from http://tinyurl.com/rouder2009 doi: 10.3758/PBR.16.2.225

Rouder, J. N., & Vandekerckhove, J. (this issue). Bayesian inference for psychology, Part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin and Review.*

Royall, R. (1997). *Statistical evidence: A likelihood paradigm* (Vol. 77). CRC press.

Royall, R. (2004). The likelihood paradigm for statistical inference. In M. L. Taper & S. R. Lele (Eds.), *The nature of scientific evidence: Statistical, philosophical and empirical considerations* (pp. 119–152). The University of Chicago Press. Retrieved from http://tinyurl.com/royall2004

Schönbrodt, F. D., & Wagenmakers, E.-J. (this issue). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review.*

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods.* Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2604513 doi: 10.1037/met0000061

Senn, S. (2013). Invalid inversion. *Significance*, *10*(2), 40–42. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2013.00652.x/full

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Meth-*

ods for Psychology(3). Retrieved from http://www.tqmp.org/RegularArticles/vol12-3/p175/p175.pdf doi: 10.20982/tqmp.12.3.p175

Stone, J. V. (2013). *Bayes' rule: A tutorial introduction to Bayesian analysis.* Sebtel Press.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1-2. Retrieved from http://dx.doi.org/10.1080/01973533.2015.1012991

van Ravenzwaaij, D., Cassey, P., & Brown, S. (this issue). A simple introduction to Markov chain Monte-Carlo sampling. *Psychonomic Bulletin and Review.*

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford Handbook of Computational and Mathematical Psychology* (pp. 300–317). Oxford University Press. Retrieved from http://tinyurl.com/vandekerckhove2015

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842–860. Retrieved from http://tinyurl.com/vandeschoot

Van de Schoot, R., Winder, S., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (in press). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods.*

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498. Retrieved from http://tinyurl.com/vanpaemel2010 doi: doi:10.1016/j.jmp.2010.07.003

van Ravenzwaaij, D., Boekel, W., Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2014). Action video games do not improve the speed of information processing in simple perceptual tasks. *Journal of Experimental Psychology: General*, *143*(5), 1794–1805. Retrieved from http://tinyurl.com/vanRavenzwaaij doi: 10.1037/a0036923

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 14–57.

Retrieved from http://tinyurl.com/verhagen2014 doi: 10.1037/a0036731

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, *14*(5), 779–804. Retrieved from http://tinyurl.com/wagenmakers2007

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive psychology*, *60*(3), 158–189. Retrieved from http://tinyurl.com/wagenmakers2010 doi: 10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Morey, R. D. (this issue). Bayesian inference for psychology, Part II: Example applications with JASP. *Psychonomic Bulletin and Review.*

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. (this issue). Bayesian inference for psychology, Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review.*

Wagenmakers, E.-J., Morey, R. D., & Lee, M. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3). Retrieved from https://osf.io/3tdh9/

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2015). How to quantify the evidence for the absence of a correlation. *Behavior research methods*, 1–14.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: an empirical comparison using 855 t-tests. *Perspectives on Psychological Science*, *6*(3), 291–298. Retrieved from http://tinyurl.com/wetzels2011 doi: 10.1177/1745691611406923

Winkler, R. L. (2003). *An introduction to Bayesian inference and decision* (2nd ed.). Holt, Rinehart and Winston New York.

Appendix

Further reading

In this Appendix, we provide a concise overview of 32 additional articles and books that provide further discussion of various theoretical and applied topics in Bayesian inference. For example, the list includes articles that editors and reviewers might consult as a reference while reviewing manuscripts that apply advanced Bayesian methods such as structural equation models (Kaplan & Depaoli, 2012), hierarchical models (Rouder & Lu, 2005), linear mixed models (Sorensen, Hohenstein, & Vasishth, 2016), and design (i.e., power) analyses (Schönbrodt et al., 2015). The list also includes books that may serve as accessible introductory texts (e.g., Dienes, 2008) or as more advanced textbooks (e.g., Gelman et al., 2013). To aid in readers' selection of sources, we have summarized the associated focus and difficulty ratings for each source in Figure 1.A1.

Recommended articles

9. **Cornfield (1966)** — Sequential Trials, Sequential Analysis, and the Likelihood Principle. *Theoretical focus (3), moderate difficulty (5).*

   A short exposition of the difference between Bayesian and classical inference in sequential sampling problems.

10. **Lindley (2000)** — The Philosophy of Statistics. *Theoretical focus (1), moderate difficulty (5).*

    Dennis Lindley, a foundational Bayesian, outlines his philosophy of statistics, receives commentary, and responds. An illuminating paper with equally illuminating commentaries.

11. **Jaynes (1986)** — Bayesian Methods: General Background. *Theoretical focus (2), low difficulty (2).*

    A brief history of Bayesian inference. The reader can stop after finishing the section

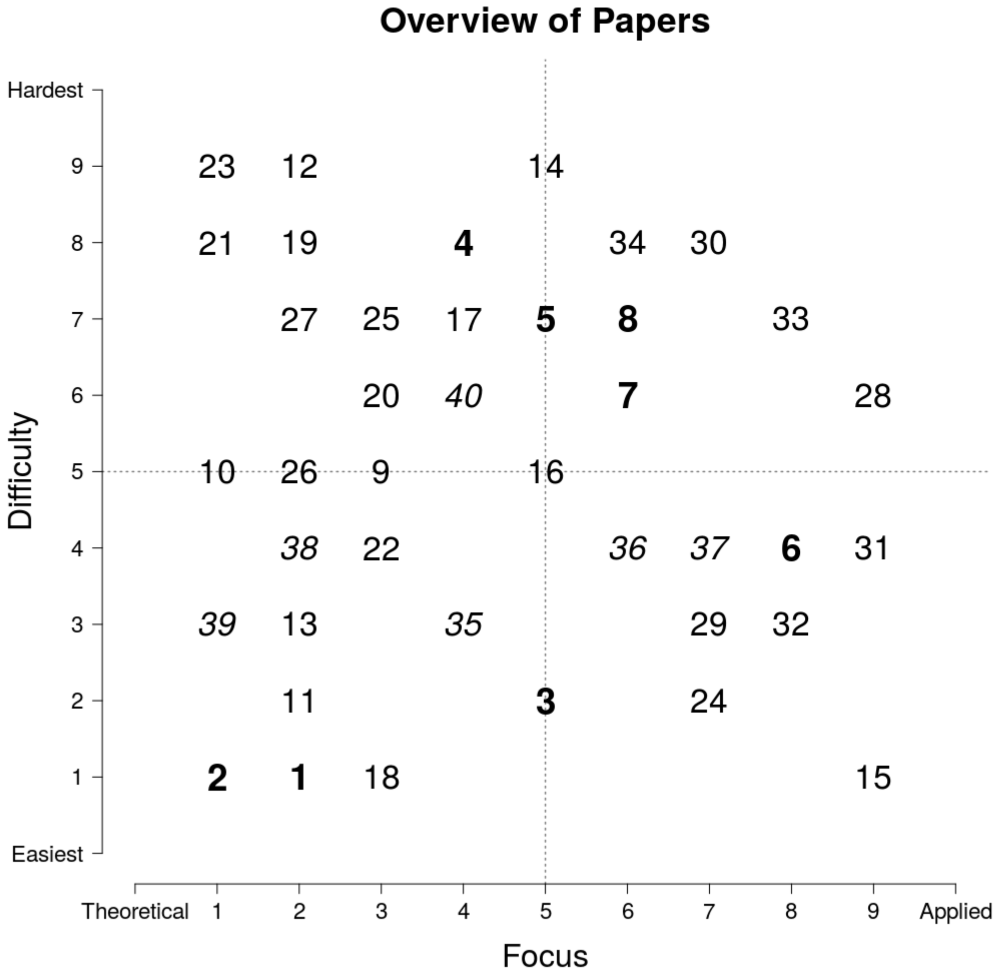


*Figure 1.A1. An overview of focus and difficulty ratings for all sources included in the present paper.*Sources discussed at length in the *Theoretical sources* and *Applied sources* sections are presented in bold text. Sources listed in the appended *Further reading* appendix are presented in light text. Source numbers representing books are italicized.

titled, "Is our logic open or closed," because the further sections are somewhat dated and not very relevant to psychologists.

12. **Edwards, Lindman, and Savage (1963)** — Bayesian Statistical Inference for Psychological Research. *Theoretical focus (2), high difficulty (9).*

The article that first introduced Bayesian inference to psychologists. A challenging but insightful and rewarding paper. Much of the more technical mathematical notation can be skipped with minimal loss of understanding.

13. **Rouder, Morey, and Wagenmakers (2016)** — The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Theoretical focus (2), low difficulty (3)*

    All forms of statistical analysis, both Bayesian and frequentist, require some subjective input (see also Berger & Berry, 1988). In this article, the authors emphasize that subjectivity is in fact desirable, and one of the benefits of the Bayesian approach is that the inclusion of subjective elements is transparent and therefore open to discussion.

14. **Myung and Pitt (1997)** — Applying Occam's Razor in Cognitive Modeling: A Bayesian Approach. *Balanced focus (5), high difficulty (9).*

    This paper brought Bayesian methods to greater prominence in modern psychology, discussing the allure of Bayesian model comparison for non-nested models and providing worked examples. As the authors provide a great discussion of the principle of parsimony, thus this paper serves as a good follow-up to our fifth highlighted source (Vandekerckhove et al., 2015).

15. **Wagenmakers, Morey, and Lee (2016)** — Bayesian Benefits for the Pragmatic Researcher. *Applied focus (9), low difficulty (1).*

    Provides pragmatic arguments for the use of Bayesian inference with two examples featuring fictional characters Eric Cartman and Adam Sandler. This paper is clear, witty, and persuasive.

16. **Rouder (2014)** — Optional Stopping: No Problem for Bayesians. *Balanced focus (5), moderate difficulty (5).*

    Provides a simple illustration of why Bayesian inference is valid in the case of optional stopping. A natural follow-up to our third highlighted source (Dienes, 2011).

17. **Verhagen and Wagenmakers (2014)** — Bayesian Tests to Quantify the Result of a Replication Attempt. *Balanced focus (4), high difficulty (7).*

Outlines so-called "replication Bayes factors," which use the original study's estimated posterior distribution as a prior distribution for the replication study's Bayes factor. Given the current discussion of how to estimate replicability (Open Science Collaboration, 2015), this work is more relevant than ever. (See also Wagenmakers, Verhagen, and Ly (2015) for a natural follow-up.)

18. **Gigerenzer (2004)** — Mindless Statistics. *Theoretical focus (3), low difficulty (1).*

    This paper constructs an enlightening and witty overview on the history and psychology of statistical thinking. It contextualizes the need for Bayesian inference.

19. **Ly et al. (2016)** — Harold Jeffreys's Default Bayes Factor Hypothesis Tests: Explanation, Extension, and Application in Psychology. *Theoretical focus (2), high difficulty (8).*

    A concise summary of the life, work, and thinking of Harold Jeffreys, inventor of the Bayes factor (see also Etz & Wagenmakers, in press). The second part of the paper explains the computations in detail for t-tests and correlations. The first part is essential in grasping the motivation behind the Bayes factor.

20. **Robert (2014)** — On the Jeffreys–Lindley Paradox. *Theoretical focus (3), moderate difficulty (6).*

    Robert discusses the implications of the Jeffreys–Lindley paradox, so-called because Bayesians and frequentist hypothesis tests can come to diametric conclusions from the same data—even with infinitely large samples. The paper further outlines the need for caution when using *improper priors*, and why they present difficulties for Bayesian hypothesis tests. (For more on this topic see DeGroot, 1982).

21. **Jeffreys (1936)** — On Some Criticisms of the Theory of Probability. *Theoretical focus (1), high difficulty (8).*

    An early defense of probability theory's role in scientific inference by one of the founders of Bayesian inference as we know it today. The paper's notation is somewhat outdated and makes for rather slow reading, but Jeffreys's writing is insightful nonetheless.

22. **Rouder, Morey, Verhagen, et al. (2016)** — Is There a Free Lunch in Inference? *Theoretical focus (3), moderate difficulty (4).*

    A treatise on why making detailed assumptions about alternatives to the null hypothesis is requisite for a satisfactory method of statistical inference. A good reference for why Bayesians cannot do hypothesis testing by simply checking if a null value lies inside or outside of a credible interval, and instead must calculate a Bayes factor to evaluate the plausibility of a null model.

23. **Berger and Delampady (1987)** — Testing Precise Hypotheses. *Theoretical focus (1), high difficulty (9).*

    Explores the different conclusions to be drawn from hypothesis tests in the classical versus Bayesian frameworks. This is a resource for readers with more advanced statistical training.

24. **Wetzels et al. (2011)** — Statistical Evidence in Experimental Psychology: An Empirical Comparison using 855 t-tests. *Applied focus (7), low difficulty (2).*

    Using 855 t-tests from the literature, the authors quantify how inference based on p values, effect sizes, and Bayes factors differ. An illuminating reference to understand the practical differences between various methods of inference.

25. **Vanpaemel (2010)** — Prior Sensitivity in Theory Testing: An Apologia for the Bayes Factor. *Theoretical focus (3), high difficulty (7).*

    The authors defend Bayes factors against the common criticism that the inference is sensitive to specification of the prior. They assert that this sensitivity is valuable and desirable.

26. **Royall (2004)** — The Likelihood Paradigm for Statistical Inference. *Theoretical focus (2), moderate difficulty (5).*

    An accessible introduction to the Likelihood principle, and its relevance to inference. Contrasts are made among different accounts of statistical evidence. A more complete account is given in Royall (1997).

27. **Gelman and Shalizi (2013)** — Philosophy and the Practice of Bayesian Statistics. *Theoretical focus (2), high difficulty (7).*

    This is the centerpiece of an excellent special issue on the philosophy of Bayesian inference. We recommend that discussion groups consider reading the entire special issue (*British Journal of Mathematical and Statistical Psychology*, February, 2013), as it promises intriguing and fundamental discussions about the nature of inference.

28. **Wagenmakers et al. (2010)** — Bayesian Hypothesis Testing for Psychologists: A Tutorial on the Savage-Dickey Ratio. *Applied focus (9), moderate difficulty (6).*

    Bayes factors are notoriously hard to calculate for many types of models. This article introduces a useful computational trick known as the "Savage-Dickey Density Ratio," an alternative conception of the Bayes factor that makes many computations more convenient. The Savage-Dickey ratio is a powerful visualization of the Bayes factor, and is the primary graphical output of the Bayesian statistics software JASP (Love et al., 2015).

29. **Gallistel (2009)** — The Importance of Proving the Null. *Applied focus (7), low difficulty (3).*

    The importance of null hypotheses is explored through three thoroughly worked examples. This paper provides valuable guidance for how one should approach a situation in which it is theoretically desirable to accumulate evidence for a null hypothesis.

30. **Rouder and Lu (2005)** — An Introduction to Bayesian Hierarchical Models with an Application in the Theory of Signal Detection. *Applied focus (7), high difficulty (8).*

    This is a good introduction to hierarchical Bayesian inference for the more mathematically inclined readers. It demonstrates the flexibility of hierarchical Bayesian inference applied to signal detection theory, while also introducing augmented Gibbs sampling.

31. **Sorensen et al. (2016)** — Bayesian Linear Mixed Models Using Stan: A Tutorial for Psychologists. *Applied focus (9), moderate difficulty (4).*

    Using the software Stan, the authors give an accessible and clear introduction to hierarchical linear modeling. Because both the paper and code are hosted on github, this article serves as a good example of open, reproducible research in a Bayesian framework.

32. **Schönbrodt et al. (2015)** — Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences. *Applied focus (8), low difficulty (3).*

    For Bayesians, power analysis is often an afterthought because sequential sampling is encouraged, flexible, and convenient. This paper provides Bayes factor simulations that give researchers an idea of how many participants they might need to collect to achieve moderate levels of evidence from their studies.

33. **Kaplan and Depaoli (2012)** — Bayesian Structural Equation Modeling. *Applied focus (8), high difficulty (7).*

    One of few available practical sources on Bayesian structural equation modeling. The article focuses on the Mplus software but also stands a general source.

34. **Rouder et al. (in press)** — Bayesian Analysis of Factorial Designs. *Balanced focus (6), high difficulty (8).*

    Includes examples of how to set up Bayesian ANOVA models, which are some of the more challenging Bayesian analyses to perform and report, as intuitive hierarchical models. In the appendix, how to use the BayesFactor R package and JASP software for ANOVA is demonstrated. The relatively high difficulty rating is due to the large amount of statistical notation.

Recommended books

35. **Winkler (2003)** — Introduction to Bayesian Inference and Decision. *Balanced focus (4), low difficulty (3).*

    As the title suggests, this is an accessible textbook that introduces the basic concepts and theory underlying the Bayesian framework for both inference and decision-making. The required math background is elementary algebra (i.e., no calculus is required).

36. **McElreath (2016)** — Statistical Rethinking: A Bayesian Course with Examples in R and Stan. *Balanced focus (6), moderate difficulty (4).*

    Not your traditional applied introductory statistics textbook. McElreath focuses on education through simulation, with handy R code embedded throughout the text to give readers a hands-on experience.

37. **Lee and Wagenmakers (2014)** — Bayesian Cognitive Modeling: A Practical Course. *Applied focus (7), moderate difficulty (4).*

    A textbook on Bayesian cognitive modeling methods that is in a similar vein to our eighth highlighted source (Lee, 2008). It includes friendly introductions to core principles of implementation and many case examples with accompanying MATLAB and R code.

38. **Lindley (2006)** — Understanding Uncertainty. *Theoretical focus (2), moderate difficulty (4).*

    An introduction to thinking about uncertainty and how it influences everyday life and science. Lindley proposes that all types of uncertainty can be represented by probabilities. A largely non-technical text, but a clear and concise introduction to the general Bayesian perspective on decision making under uncertainty.

39. **Dienes (2008)** — Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. *Theoretical focus (1), low difficulty (3).*

    A book that covers a mix of philosophy of science, psychology, and Bayesian inference. It is a very accessible introduction to Bayesian statistics, and it very clearly contrasts the different goals of Bayesian and classical inference.

40. **Stone (2013)** — Bayes' Rule: A Tutorial Introduction to Bayesian Analysis. *Balanced focus (4), moderate difficulty (6).*

    In this short and clear introductory text, Stone explains Bayesian inference using accessible examples and writes for readers with little mathematical background. Accompanying Python and MATLAB code is provided on the author's website.

CHAPTER 2: USING HIERARCHICAL MODELING TO QUANTIFY THE ROBUSTNESS OF EXPERIMENTALLY OBSERVED EFFECTS

This chapter was published in March 2018 as a peer-reviewed article in the *Proceedings of the National Academy of Sciences*.[1] My primary role in this work was to design and perform the research (i.e., developing the radical randomization approach to experimental design and coordinating data collection across all seven labs), in addition to contributing to the writing of the paper.

Metastudies for robust tests of theory

Beth Baribault[a], Chris Donkin[b] , Daniel R. Little[c],
Jennifer S. Trueblood[d], Zita Oravecz[e], Don van Ravenzwaaij[f],
Corey N. White[g], Paul De Boeck[h], & Joachim Vandekerckhove[a]

[a]University of California, Irvine

[b]University of New South Wales

[c]University of Melbourne

[d]Vanderbilt University

[e]Pennsylvania State University, State College

[f]University of Groningen

[g]Missouri Western State University

[h]Ohio State University

Abstract

We describe and demonstrate an empirical strategy useful for discovering and replicating empirical effects in psychological science. The method involves the

[1] https://www.pnas.org/content/pnas/115/11/2607.full.pdf

design of a *meta-study*, in which many independent experimental variables—that may be moderators of an empirical effect—are indiscriminately randomized. *Radical randomization* yields rich data sets that can be used to test the robustness of an empirical claim to some of the vagaries and idiosyncrasies of experimental protocols and enhances the generalizability of these claims. The strategy is made feasible by advances in hierarchical Bayesian modeling which allow for the pooling of information across unlike experiments and designs, and is proposed here as a gold standard for replication research and exploratory research. The practical feasibility of the strategy is demonstrated with a replication of a study on subliminal priming. All materials and data are freely available online via `https://osf.io/u2vwa/`.

Introduction

Imagine, if you will, an experiment in the psychological laboratory. In the experiment, a single participant provides data in each of two conditions. Further suppose an effect is observed in the form of a mean difference between the two conditions. Unless there are strong reasons to believe that all humans are largely interchangeable with respect to this particular effect, readers and reviewers will reasonably point out that this effect might be idiosyncratic to the participant and hence not generalize to the broader population.

One potential remedy is for the researcher to replicate the experiment with the same participant and one newly recruited participant – thereby enacting a systematic manipulation of the suspected moderating variable (i.e., participant identity). Such a design enables at least two related claims: possibly that there are individual differences in the magnitude of the effect, and possibly that the effect occurs in some participants but is absent in others.

This strategy is, however, clearly limited: it does not allow for population-level inference. Rather than merely observing that some individual differences *could* occur, we might instead be interested in whether the effect holds for *most* humans, or *on average* across humans, or perhaps for *all* humans. Such claims call for a hierarchical strategy in which not one or two but many participants are randomly sampled from the population towards which we wish to generalize. If the resultant sample is representative of the population, then the sample mean effect will be an unbiased estimate of the population mean effect and the sample variance in the effect will permit statements about the generality of its occurrence.

In the same way that psychological scientists typically want to generalize from one participant to all potential participants (within certain boundaries), so too will they often want to generalize from a small set of conditions to all conditions (within certain boundaries). For example, researchers who want to claim that stress impairs memory presumably believe that this effect is not specific to the particular aspects of one specific experiment. However, testing the myriad experimental facets, or moderators, involved (e.g., setting, stimuli, etc.) can be burdensome, time-consuming, and expensive. The strategy of *random selection* is a sound and viable one for potential moderators of an experimental effect, including potential moderators other than participant identity. In particular, we believe that extensive randomization can lead to scientific conclusions that are more general in scope, more robust to incidental variations in experimental setup, and more likely to replicate in future studies.

In what follows, we will we introduce the concept of a *meta-study*, in which we combine radical randomization of experimental features and systematic pooling of information with a Bayesian hierarchical model. We argue that sampling from a population of possible experiments in the same way one would sample from a population of possible participants is a practically feasible approach that can increase the robustness of empirical findings in psychology.

Causes of nonreplication and variations on replication

Replicability of empirical findings has been a central topic in recent psychological science. Following a series of dramatic revelations in which researchers have appeared unable to reliably replicate empirical effects once thought to be robust, there is now talk of a “crisis of confidence” (Pashler & Wagenmakers, 2012) in the field. While there are a number of possible explanations for the lack of replicability (Francis, 2012), one commonly indicated problem is the issue of *publication bias*: the preference to publish statistically significant results (i.e., results that lead to the rejection of a null hypotheses; Guan & Vandekerckhove, 2016; Rosenthal, 1979). This statistical significance filter (Vasishth & Gelman, 2017) biases the published record towards results that capitalize on measurement noise and fluke outcomes (Sterling, 1959).

Moreover, evidence from psychological studies—even if published without bias towards certain outcomes—is often weak due to traditions of insufficient sample sizes and noisy measurement tools, which leads to generally low ability to detect true effects and a concomitant increase in false positive results (Etz & Vandekerckhove, 2016; Gelman & Loken, 2014). The combination of publication bias and low standards of evidence would naturally cause frequent failures to replicate, since effects claimed in the published literature are likely to be false alarms. Given the uncertain nature of one-off effects found in the literature, replication of empirical results is a clear gold standard of convincing evidence: greater confidence is warranted in theories whose predictions repeatedly come true (Fisher, 1935) or whose predictions survive repeated falsification attempts (Popper, 1963).

At the same time, even when a published effect is true, it is possible for effects to fail to replicate strictly due to seemingly innocuous differences in the implementation of the experiment (i.e., due to "hidden moderators" that may occur in replication studies). Small variations in experiments are of course unavoidable: exact replication is strictly impossible. However, for the purposes of creating generalizable knowledge what matters most is recreating the necessary and sufficient conditions that will show the effect as predicted by some theory. By implication, small experiment variations that are not theoretically relevant should have only minimal impact on the size of a true effect. Indeed, theoretical statements made by researchers almost without fail imply some degree of robustness to irrelevant variables. It was recently proposed that authors make these claims explicit as part of every paper (Simons, Shoda, & Lindsay, in press; Kenett & Rubinstein, 2017).

Such robustness is, of course, a testable assertion. We could take any one of these suspected hidden moderators, systematically vary it as an independent variable in an experiment, and quantify any differences so obtained. Much theoretical knowledge grows exactly in this fashion.

A related distinction that is often made among replication attempts is that between *direct* and *conceptual* replications. A direct replication is one in which the replicating team attempts to follow the original protocol as closely as possible, allowing for no moderating variables that might distort the findings or obfuscate the effect seen in the original publication. In a direct replication, the *exact same theoretical prediction*—that is, the same *hypothesis*—is tested. A conceptual replication, on the other hand, is one in which the replicating team tests the same *theory*, but uses a different instantiation of theory to hypothesis, with entirely different values on some independent variables and possibly different dependent and independent variables as well. In such a replication, the issue at hand is the robustness of a reported effect to theoretically irrelevant design variations.

Both of these approaches have associated problems. A common concern about direct replications is that it is typically impossible to copy a protocol exactly: replications tend to take place at a different time and place from the original, with different subjects, and they are often by a different lab with slightly different ineffable and undocumented practices, and not all the relevant details are reported in the original publication. Conceptual replications, on the other hand, lack falsification power: a lack of effect may be due to one of the many differences between the original and the replication. While irrelevant within the adopted theoretical framework, an innocuous difference in design might in fact be a genuine moderating factor. As such, the masking of an otherwise replicable effect by a hidden moderator and a genuine failure to replicate are strictly unable to be teased apart with conventional techniques.

Radical randomization

Here we present an alternative take on replication that involves the *radical randomization* (RR) of many features of an experiment. As an example, imagine a study in which researchers are interested in some difference between two manners of stimulus presentation. A visual stimulus (e.g., the symbol `v`) is either presented to the participant normally for a short time (e.g., 30 ms), or it is presented with temporal masking – meaning that it is preceded and followed by visual masks (e.g., strings of symbols such as `&&&`). These masks are called *forward masks* and *backward masks*, respectively, and their addition sometimes suppresses the conscious perception of the temporally flanked stimulus. Such an experiment has a few immutable features that are necessary to address the question at hand (critically, some stimuli need to be masked while others are not). However, many of the features of this experiment are chosen largely arbitrarily: presumably there is nothing special about the symbol `v` and the same differences could be illustrated with the symbol `b` instead; and presumably `###` is as effective a forward or backward mask as `&&&`. If the effect exists, it should shine through—if perhaps diminished—for many different symbols and many different small variations on the experimental setup.

In a RR design, this presumption of robustness is put to a critical test. Rather than consistently using the symbol `v`, we instead randomly choose any symbol from a set, and then choose a new symbol whenever we can (without harming the validity of the study). Such a design could be considered *defensive* in the sense that it hardens our conclusions against minor infidelities in future replication attempts (i.e., replication attempts that are not strictly faithful and hence are not direct replications) – infidelities such as using a different symbol. That is, the RR design makes conclusions more robust because it mimics some of the potential variance between an experiment and future replication attempts that are—as all replications are—inexact.

In order to distinguish those immutable IVs that are needed to define the effect of interest from the innocuous design features (strictly speaking also IVs) that are randomized, it will be useful to introduce some new terminology. Borrowing from Generalizability Theory (Cronbach, Rajaratnam, & Gleser, 1963), we call these to-be-randomized IVs *facets*, and we call a study with many facets a *meta-study.* While a typical IV has a limited set of values that we normally call *conditions*, the values of a facet are drawn randomly from a potentially infinite population. We call the values of a facet that happened to be drawn for a particular meta-study its *levels*, and we call each cell in the multifaceted design a *micro-experiment.* The immutable IVs that occur in each micro-experiments will be called *elementary IVs.* Finally, it will sometimes be useful to think of the population of possible micro-experiments, which is defined by the space spanned by all the facets of a study. We call this the *method space.*

Facets can be simple design choices (e.g., the exact stimuli selected from a larger pool), natural constraints (e.g., the geographical location of the lab), or explicitly labeled nuisance variables that are randomized (e.g., individual differences between participants). The goal of introducing variability in a facet is to investigate the generality of an effect within a much broader subspace of the method space than is commonly the case. If an effect remains, despite variability in some design features, we establish robustness: invariance of the effect to reasonable variation in the facet. Alternatively, the effect may turn out be sensitive to such variability.

What constitutes "reasonable variation'—as formalized by the distribution from which levels of a facet are drawn—is up to the judgment of the researcher. The sampling distribution of a facet determines the "universe of intended generalization": the range within which we aim to establish the existence of the effect. In general, levels should be sampled so that they well represent the range of the facet across which one hopes to draw conclusions.

Facets may be of particular interest when they are predicted—by one theory or another—to moderate an empirical effect. In such cases, establishing the moderating influence or the invariance of the effect are both of theoretical interest. However, the purpose of a RR procedure is not to build or refine theories as much as it is to establish that an effect holds.

Researchers setting up a meta-study are therefore recommended to be liberal in which facets they select for randomization.

We are of course not the first to suggest randomization of experimental features. Indeed, in 1973 psycholinguist H. H. Clark (Clark, 1973) suggested it as a treatment for what he called the language-as-fixed-effect fallacy, and R. A. Fisher (Fisher, 1935) famously proposed it to avoid systematic effects of sampling locations in agricultural experiments. Our position might be characterized as an objection to a broader error of inappropriate use of fixed effects.

Finally, we should point out that randomization itself is not unique in its suitability toward the goal of obtaining a representative sample (Worrall, 2007). We merely propose it here as a convenient practical approach to exploring the space of possible micro-experiments.

Individually weak, jointly powerful

The RR approach that we propose involves the implicit construction of many micro-experiments and randomly sampling among them. A micro-experiment might consist of all the trials that share a level of one selected facet (hundreds or thousands of trials), but may be as small as all the trials in a single block by a participant (a few dozen trials). What constitutes a micro-experiment is less a design decision than a feature of the statistical analysis: it is a grouping of observations that is homogeneous in the facet(s) of interest (but has variability in the elementary IVs so that contrasts can be computed).

Individually, these micro-experiments do not deliver much evidence for or against the existence of an effect. However, a key component of the approach is the use of modern statistical techniques (e.g., Bayesian hierarchical modeling and meta-analysis; Gelman & Hill, 2006; Sutton & Abrams, 2001) to pool information across data sets efficiently.

Theory-testing

A meta-study serves to make a stronger statement about the existence of an empirical effect – namely, its persistence across variations on an experiment. To test an effect in such an hierarchical scenario, it is more beneficial to increase the number of independent variations than it is to increase the number of data points. Hence, by randomly sampling many locations in the method space and conducting a small independent experiment in each location, the multifaceted design allows robust and statistically powerful statements about the effect.

A theory, with an intended universe of generalizability, can be formalized as an effect size function over a *region* within a method space – rather than over a point, which would represent a more local hypothesis. The region of the method space within which an effect presents itself allows us to make empirically-backed statements about the constraints on generality—that is, the boundary conditions of the theory—that are usually only implicit in psychological theories.

While this strategy seems straightforward—perhaps even obvious—it is to the best of the authors' knowledge essentially unused in psychological or cognitive science. Over time, research groups with a concerted study program eventually develop a portfolio of experiments that vary in small ways, and in that sense these groups work to establish robustness (or observe the lack of it). However, the systematic execution of such a population of experiments—in what we here call a meta-study—does not occur, leading to the potential for bias and correlated error. We believe that the multifaceted design has great potential as a defensive design strategy that allows for more general statements and tests of theory, and is likely to yield conclusions that are more robust to small variations in design implementation.

Statistical analysis of multifaceted designs

The multifaceted design affords a number of different statistical approaches. In this section, we discuss three possibilities. In the case example, we will demonstrate all three.

In what follows, we will assume an experimental meta-study with some set of elementary independent variables that are theoretically interesting (i.e., whose effect on a dependent variable we are hoping to quantify) and some set of facets. Most facets are not relevant according to the theory we are testing, but might be relevant according to some unspecified rival theory or be relevant in ways that are simply not yet discovered.

Global tests

Many experimental studies are specifically designed to answer a particular question, often of the unary form “is A different from 0” or the binary form “is B greater than C”? Even though we often have multiple, randomly selected participants and we expect there to be person-level variability, the random effect of participant identity is often ignored on the (reasonable) assumption that with a sufficiently large sample, any interindividual differences will “wash out” so the sample is balanced and the sample mean effect is a good estimate of the population mean effect. With the same argument, we can—in a first pass—ignore the differences between the randomly sampled levels of the facets in an experiment. This way, we are able to test for the existence of an inequality *on average* over the range of possible values of the facet.

The formulation of the model is somewhat standard. Letting $y_{m(i)}$ stand for the dependent variable observed at trial i (which is nested in micro-experiment m) and letting $x_{km(i)}$ stand for for the corresponding value of the kth elementary IV X_k (where conventionally $x_{0m(i)} = 1$ to represent the intercept), the global test model has a set of regression weights β_k and a variance ς^2. Errors $\epsilon_{m(i)}$ are i.i.d. standard normal:

$$y_{m(i)} = \sum_{k=0}^{K} \beta_k x_{km(i)} + \varsigma \epsilon_{m(i)}.$$

This fairly common formulation subsumes as special cases the models associated with the t test (if $K = 1$ and X_1 is binary), linear regression (if X are continuous), or ANOVA (if $K > 1$ and all X_k are binary).

We emphasize, however, that such a global test is only valid if the results are relatively homogeneous between micro-experiments. In the same way that ignoring large individual differences may invalidate the results of a conventional experiment, if a facet causes true heterogeneity in the effect size, the global test can be a poor approximation, and it is important to evaluate whether the test is appropriate before drawing conclusions from it.

Level-2 heterogeneity and moderation

Experimental effect sizes are inherently unstable. Even in the absence of explicit moderators, any set of experiments will show variance *even in the true effect size* – that is, above and beyond measurement error. This instability—which occurs due to ephemeral differences even between superficially identical designs—is sometimes referred to as *level-2 heterogeneity.*

The global hypothesis test above makes no statement about the robustness of the finding to variations in the experimental setup. In order to evaluate robustness, we can apply an hierarchical model in which a facet is allowed to interact with any or all of the elementary IVs (including the intercept). We then inspect if and how the effect varies over the range of each of the individual facets. In the hierarchical model, the regression weights are decomposed to yield the following random-effects model equation:

$$y_{m(i)} = \sum_{k=0}^{K} \left(\beta_k + \sigma_k \gamma_{km}\right) x_{km(i)} + \varsigma \epsilon_{m(i)}.$$

Here, the new parameter γ_{km} indicates the unique contribution of the facet to the effect of the kth elementary IV. The parameter is i.i.d. standard normal. Of primary interest in this scenario is σ_k, the level-2 variance of the contribution of the facet to the effect size β_k, and potentially the pattern of change in γ_{km} across its levels m. The former quantifies the heterogeneity of the effect size: σ_k can be compared to the fixed effect size β_k for reference; the ratio $\rho_k = \sigma_k/\beta_k$ is sometimes called the *coefficient of variation.* The parameter ρ_k may be interpreted as a measure of robustness, with small values (say, less than 1/3 or 1/4) indicating robustness and large values indicating sensitivity to the facet k. The changes in γ_{km} over the facet allow us to visualize and study its influence.

While it is sometimes sufficient to visualize an effect or a pattern of effects across values of a moderator, we occasionally need to test whether an effect is nominally present or absent in a given condition. For this purpose, we can use a Bayes factor (or likelihood ratio), which expresses by how much the relative probability of a pair of hypotheses changes when the data are taken into account. That is, if $\mathcal{H}_a$ and $\mathcal{H}_b$ are the hypotheses under consideration and x is the data, the Bayes factor is given by

$$B_{ab} = \frac{P(\mathcal{H}_a|x)/P(\mathcal{H}_b|x)}{P(\mathcal{H}_a)/P(\mathcal{H}_b)}.$$

We will interpret $B_{ab} \geq 10$ as strong support for $\mathcal{H}_a$.

Planned meta-analysis

A meta-study will typically lead to somewhat larger data sets than are common in psychological science. In order to apply a high-dimensional statistical model to a large data set, we use one particularly useful approximation that changes our analysis from a standard hierarchical model into a *planned meta-analysis*. The approximation is based on the central limit theorem, which allows us to substitute n_m normally distributed data points $y_{m(i)}$ with variance ς^2 by their means $\bar{y}_m$ with standard deviation equal to the standard error of measurement s_m:

$$\bar{y}_m = \sum_{k=0}^{K} (\beta_k + \sigma_k \gamma_{km}) \, x_{km} + s_m \epsilon_m.$$

A conventional meta-analysis involves a set of studies, each of which can be represented as a point in the method space, with the exact location chosen by the experimenters. The meta-analyst then computes a weighted average of effect sizes across these studies. While conventional meta-analysis is often plagued by severe issues such as publication bias, this is not a concern for the meta-study. Similarly, the issue of hidden moderators is reduced here since at least some differences between micro-experiments are recorded: facets are explicitly identified and their levels are not arbitrarily chosen but—to the extent possible—fairly and independently sampled from a well-defined population distribution.

In the following section, we will apply these methods and analyses to an experimental study in cognitive science. For the purposes of exposition, we will omit some detail regarding the experiment (full detail is available via `https://osf.io/u2vwa/`).

The effect of masked cues on cognitive control

As a toy demonstration, we replicate a recently published experiment in cognitive psychology.[2] Reuss et al. (see esp. Fig. 1 Reuss, Kiesel, & Kunde, 2015) describe an experiment in which a cue that is presented for a subliminal amount of time (i.e., too briefly to be consciously detected) influences how participants balance speed and accuracy in a response time task. This design has obvious facets (e.g., the color of the cue) whose exact values are not expected to affect the finding of subliminal perception: If the effect is robust, it should appear at all values of the facet; If it is fickle, it should appear in some (contiguous) value ranges but not in others; If it is false, it should not consistently appear in any range of values.

The basic task

In the experiment, participants were shown a "bullseye" stimulus consisting of a dot surrounded by nine concentric circles. The stimulus appeared either in the right or the left half of the screen and participants were instructed to move the mouse pointer from the center of the screen to the center of the bullseye and then click the left mouse button. Shortly before the presentation of the stimulus, a single-letter cue was presented, instructing participants to either favor accuracy (measured in distance from the center) or favor speed. Additionally, the cue was either masked (by the rapid presentation of two three-symbol strings like ### and &&&) or not, giving rise to four experimental conditions. Of primary interest is the effect of the masked cue instruction on the speed and accuracy of the responses that Reuss et al. (Reuss et al., 2015) first reported.

[2] The experiment was approved by the institutional review boards of UC Irvine (#2015-1802), Syracuse University (#13-269), Vanderbilt University (#151563), the University of Groningen (#15122-NE), the University of New South Wales (#153-2387), and the Melbourne School of Psychology (#1544198.3). All participants provided informed consent at the beginning of the experiment and were informed that participation was voluntary.

Table 2.1
Heterogeneity over facets.

Facet	**Levels**	**Original**	$\hat{\rho}^a$
First forward mask duration	0 – 50 ms	40 ms	0.42
Second forward mask duration	0 – 50 ms	30 ms	0.52
Total forward mask duration	0 – 100 ms	70 ms	0.59
First backward mask duration	0 – 50 ms	40 ms	0.49
Second backward mask duration	0 – 50 ms	30 ms	0.52
Total backward mask duration	0 – 100 ms	70 ms	0.69
Masked cue duration	0 – 50 ms	30 ms	0.90
Blank interval duration	250 – 750 ms	500 ms	0.93
Intertrial interval duration	500 – 1500 ms	1000 ms	0.55
Mask and cue color	*(13 colors)* [b]	white	0.13
Mask and cue contrast	$0.5 \leq x \leq 1.0$	1.0	0.21
Target center color	*(13 colors)* [b]	red	0.10
Target center contrast	$0.5 \leq x \leq 1.0$	1.0	0.21
Target surround contrast	$0.5 \leq x \leq 1.0$	1.0	0.22
First mask symbol	@ , #, $, %, &, ?	#	0.06
Second mask symbol	@ , #, $, %, &, ?	%	0.06
Location	*(6 locations)* [c]		0.36

[a]: $\hat{\rho}$ indicates the observed heterogeneity that the facet introduces in the effect of masked cues on accuracy (lower values indicate greater robustness);
[b]: 12 hues were sampled between integer multiples of 30° angles in HSV color space; the 13th color was white;
[c]: The locations were the research labs of authors CD (Sydney, Australia), CNW (Syracuse, NY), DRL (Melbourne, Australia), DvR (Groningen, the Netherlands), JST (Nashville, TN), and JV (Irvine, CA).

Sampling the method space

During the development of the study, the experimenters collaboratively constructed a list of facets to include. In Table 2.1, we list facets related to timing, including the duration of the first and second forward mask, of the first and second backward mask, of the masked and unmasked cue; facets related to color, including the hue and luminance of masks and cues; and other miscellaneous facets, such as the symbols used in the mask and the testing location.

Each of these facets was assigned a distribution from which its values were to be randomly sampled at the beginning of each micro-experiment. In almost all cases, this involved a uniform distribution over a range of integer values (e.g., the variables relating to presentation time were naturally expressed as an integer number of frames). For one facet, variance was introduced not through random sampling but by a convenience sample: the experiment was conducted in 6 different geographical locations.

The experiment

Each participant's session of the experiment began with 16 practice trials whose facets were set to match the original study by Reuss et al. as closely as possible. After that, each block of trials consisted of (1) 40 "bullseye" trials whose facets were set to a random value sampled from the corresponding distribution; and (2) 40 "cue identification" trials whose facets were set to the same values used in the immediately preceding bullseye block. The first 8 trials of each type were considered practice trials as well. The goal of the cue identification trials was to confirm the true subliminal nature of the masked cue. Crucially, all facets' values were resampled at the start of each block of bullseye trials, making each block of trials a unique *micro-experiment*.

Practice trials were discarded. At each bullseye trial, we recorded two dependent variables: (1) the participant's response time and (2) the distance (in mm) between the center of the stimulus and the point where they clicked. In the cue identification trials, we recorded (1) the response time and (2) the (binary) accuracy. We discarded trials where the reaction time was too high (over 2500 ms) or too low (under 150 ms) and where the participant clicked without moving the pointer.

Each of the 6 participating labs decided how many blocks each participant would complete (all labs chose 14 blocks, which made for approximately one-hour sessions) and how many participants would be recruited; with no fixed stopping rule set. Labs recruited between 47 and 78 participants from their institutional human subjects pools, for a total of 346 participants and up to 4,844 micro-experiments, all with randomly drawn levels on each facet.

The dependent variable

Throughout the following analyses, the quantity of interest is the *magnitude of the conditional effect of the cue when it is masked* – that is, the difference between the masked-cue, accuracy-instruction condition and the masked-cue, speed-instruction condition. For the purposes of exposition, we will focus only on the dependent variable "accuracy" (negatively coded as the distance from the center of the bullseye), but similar results were found for the "reaction time" dependent variable.

Level-2 variability

In order to quantify the heterogeneity between the 4,844 micro-experiments, we applied an hierarchical Bayesian model (Stan Development Team, 2014) that included a unique effect size parameter for each micro-experiment (i.e., a random effect of micro-experiment). This results in a distribution of effect sizes with as many values as there were micro-experiments. Focusing on the effect of masked cues only, the mean of that effect size distribution was estimated at $\hat{\beta} \approx 3.36$ mm. However, its population standard deviation was $\hat{\sigma} \approx 6.46$ mm and the coefficient of variability was $\hat{\rho} \approx 2$, which indicates that the effect is sufficiently sensitive to the differences between micro-experiments that it will occasionally vanish.

A histogram of the distribution of effect sizes over micro-experiments (Fig. 2.1) shows the large variability. To construct these histograms, we computed Bayes factors[3] to express the statistical support for a non-zero effect in each micro-experiment. The sample effect sizes more consistent with a zero effect make up the inverted histogram. The figure shows that three-quarters of the individual micro-experiments in the masked condition appear more consistent with *no* effect than with a positive effect and a small number show an effect

[3] The Bayes factors express how much less likely the effect size of 0 mm is under its posterior distribution than under its prior distribution. The prior distribution of the effect size $\hat{\beta}$ is derived from the prior distributions of the condition means, which was in turn derived from the source paper (Reuss et al., 2015). Assuming a repeated measures correlation of no more than 0.5, the effect size prior worked out to a normal distribution with mean 0 mm and standard deviation 10 mm. This test is maximally sensitive to effect sizes that are slightly smaller than the global mean effect size in the original paper. None of our conclusions regarding Figure 2.1 are sensitive to reasonable variation in these assumptions.

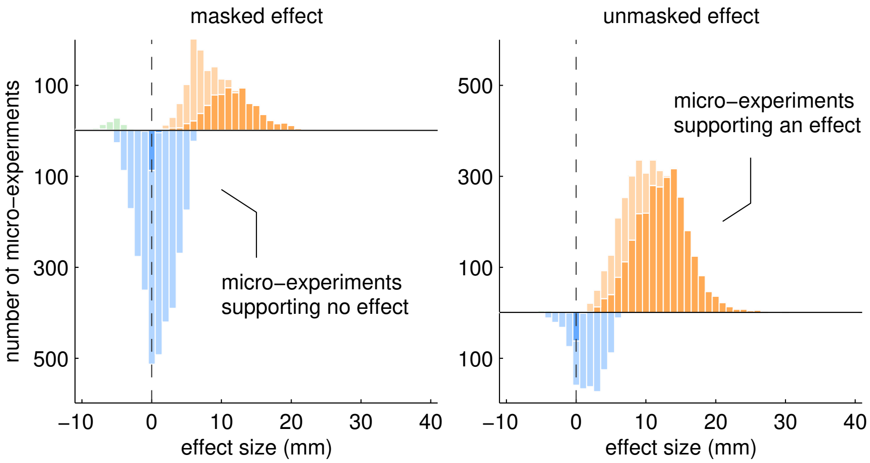


Figure 2.1. *Level-2 variability.* Histograms of estimated effect sizes across micro-experiments are split between masked (left) and unmasked (right) conditions and between micro-experiments that support an effect (regular bars) versus no effect (inverted bars). Darker bars indicate stronger support with a Bayes factor of at least 10. A majority of micro-experiments show support for the unmasked effect, but a similarly large number support no effect of the masked cue.

in the opposite direction. By contrast, in the unmasked condition, the large majority of micro-experiments are more consistent with a positive effect.

The large variability appears to suggest the existence of one or more moderating variables hidden in our design. We can quantify the heterogeneity of this effect by applying a sequence of hierarchical models. In each model, we will estimate the variability of the effect size across levels of one facet (i.e., a random effect of the facet). Each such analysis will yield an estimated coefficient of variability associated with that facet. These estimates are given in Table 2.1. The largest heterogeneity is seen in the various timing facets, and the effect is particularly unstable across levels of 'masked cue duration' amd 'blank interval duration', while it appears to be relatively robust to changes in colors and symbols.

Moderator analysis

The observed heterogeneity can be explored by the explicit introduction of potential moderators of the effect. One candidate moderator that is not included in Table 2.1 is the subliminality of the cue as presented. Recall that after each block of bullseye trials, participants completed a block of trials in which they were asked only to identify the cue. In

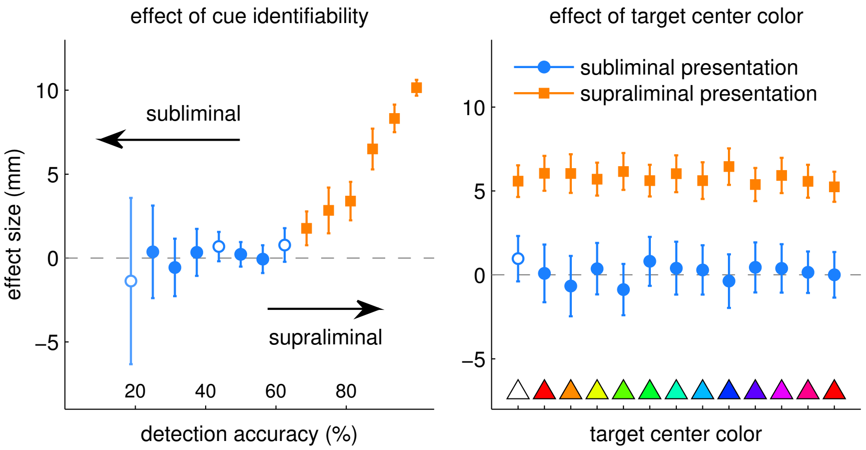


Figure 2.2. *The sensitivity and robustness of the effect to two moderators.* **Left:** Micro-experiments support an effect when participants are able to consciously identify the cue (square markers), but not otherwise (round markers). **Right:** The data are split by subliminality. The facet "target center color" was varied over 13 possible levels, but the facet does not appear to moderate the effect of interest. That is, the effect appears robust against this facet. **Both:** Error bars show 99% credibility intervals. Solid square markers indicate strong evidence (BF > 10) for a nonzero value. Solid round markers indicate strong evidence for a zero value. Empty markers indicate ambiguous evidence.

these cue-identification blocks, the cue was presented with the same settings (i.e., the same values on the relevant facets) as in the bullseye trials. We can quantify the subliminality of the cue under these conditions by the accuracy in the cue-identification trials.

Figure 2.2 (left panel) shows how the effect of the masked cue varies as a function of the subliminality of the cue presentation. Only in those micro-experiments where the cue identification accuracy is at least 68% does an effect of the masked cue appear. In the figure, square markers are filled if the data strongly support an effect (with a Bayes factor of at least 10), round markers are filled if an effect size of zero is strongly supported, and empty markers indicate ambiguity. Each facet can be explored in a similar way to evaluate whether it moderates the effect of interest.

The level-2 variability analysis hinted at the presence of a potential moderator, and Figure 2.2 identifies subliminality as one. We can construct similar figures to indicate the *lack* of a systematic effect of a facet. For example, a facet that is an unlikely moderator is the color of the target center. In Figure 2.2 (right panel), we graph the effect size as a function of this facet, splitting micro-experiments according to whether the cues were consciously visible. The effect appears to be robust to changes in this facet since it occurs across all levels of the facet for supraliminal trials (squares) and nowhere for subliminal trials (circles).

Conclusion

The effect of masked cues is strongly qualified by the moderator analysis. Masked cues seem to have an effect on participant behavior only in those settings where the cue is consciously visible. We find no evidence of an effect of subliminally presented cues. To the contrary: our data are more consistent with *no* effect when the cue presentation is truly subliminal.

Discussion

Robustness and generalizability of empirical results are critical considerations regarding the reproducibility crisis that has beset psychological science. The radical randomization approach to experimental design, in which features of an experimental design are strategically randomized, allows researchers to make statements that are less sensitive to unavoidable between-study variability. When a single experiment demonstrates the existence of some effect, there is the risk that the effect is isolated to a particular "sweet spot" in the method space. By contrast, the meta-study allows us to make statements about effects in regions in a method space: a well-defined and formalized universe of intended generalization.

In our view, meta-studies complement the standard approach to empirical research. The radical randomization approach speaks to the robustness of empirical effects, but such information is only useful to the extent that it informs the development of substantive theory. Experiments with tight control and fixed effects are an established means of generating theoretical explanations for data; we view meta-studies as an efficient way of testing such theories by complementing the fixed effect approach with random effects.

The strategy has some weaknesses to keep in mind. First, it is impractical in certain settings, such as when data is expensive to collect. However, it is particularly well suited for "many labs" style projects in which an ad-hoc consortium of research labs collaborates in data collection. Still, a meta-study could very reasonably be run within a single lab – from a logistical standpoint, the cost to each lab that contributed to the applied example was comparable to that of a typical experiment in cognitive science (arguably it was slightly lower since the study materials were produced entirely by the UC Irvine and UNSW labs). Second, in all but some cases it will be impossible for a research team to identify all facets that might moderate an effect. It serves to remember that claims of generality remain confined to the actually realized method space. However, the randomization of experimental features does provide for a built-in test of some robustness to small variations in experimental features, it can be used to spot weaknesses in an experimental design as well as in empirical claims, and it can be used to generate novel hypotheses when a facet unexpectedly turns out to be influential.

The major strength of radical randomization, and the reason why we recommend it, is that it allows for *defensive design*: a design strategy under which studies are optimized for generalizability, replicability, and robustness.

References

Clark, H. H. (1973). The language–as–fixed–effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*, e0149794. (Via `escholarship.org/uc/item/1nj1r7b1`)

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis – a "garden of forking paths" – explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460.

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*, *23*(1), 74–86. (Via `escholarship.org/uc/item/2682p4tr`)

Kenett, R. S., & Rubinstein, A. A. (2017). A generalization approach to reproducibility claims.
(Via `ssrn.com/abstract=3035070`)

Pashler, H., & Wagenmakers, E.-J. (2012). Editor's Introduction to the Special Section on Replicability in Psychological Science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.

Popper, K. (1963). *Conjectures and refutations: the growth of scientific knowledge.* Routledge & Kegan Paul.

Reuss, H., Kiesel, A., & Kunde, W. (2015). Adjustments of response speed and accuracy to unconscious cues. *Cognition*, *134*, 57–62.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (in press). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science.* (Via `osf.io/preprints/psyarxiv/w9e3r`)

Stan Development Team. (2014). *Stan: A c++ library for probability and sampling.* (Via mc-stan.org)

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*(4), 277–303.

Vasishth, S., & Gelman, A. (2017). The illusion of power: How the statistical significance filter leads to overconfident expectations of replicability. *arXiv preprint arXiv:1702.00556*. (Via arxiv.org/abs/1702.00556)

Worrall, J. (2007). Why there's no cause to randomize. *The British Journal for the Philosophy of Science*, *58*(3), 451–488.

CHAPTER 3: COGNITIVE LATENT VARIABLE MODELING: A TECHNIQUE FOR HIERARCHICALLY COMBINING COGNITIVE AND PSYCHOMETRIC MODELS

Abstract

Cognitive latent variable models are a powerful new class of hierarchical Bayesian models that combine a cognitive model of the data with a psychometric model of the latent structure of cognitive constructs. This fusion of techniques allows the researcher to infer a small set of unobserved, large-scale cognitive factors from the observed data, as seen through the lens of a cognitive model. In this tutorial, we outline a general method for implementing cognitive latent variable models using JAGS, a Bayesian sampling package. Specifically, we describe the development and use of a set of cognitive latent variable models designed to infer which theorized attention constructs are most likely to have produced participants' observed performance on a battery of attention tasks. We aim not only to demonstrate how a researcher may construct their own cognitive latent variable models, but also to showcase the unique benefits of this new approach to cognitive modeling.

I. Introduction

As cognitive scientists, we are generally trying to determine what unobserved cognitive abilities underlie observed performance on behavioral tasks. Through experimental methods, these abilities may be indirectly probed by manipulating one or more variables that are theoretically related to the target ability while holding all other variables constant. However, this process makes a plethora of assumptions, which may or may not be met. We must assume that the experimental manipulation is in fact manipulating the variable of interest, that the data systematically vary with the variable of interest, that participants and items are interchangeable, and more. By using computational models, we are able directly account for one or more of these ambiguous assumptions, and thus may generate more direct inferences about unobserved cognitive abilities.

There are two general approaches to modeling performance on behavioral tasks. The first, cognitive modeling, postulates that an unobserved but quantifiable psychological process has generated the observed data. The second, latent variable modeling, proposes that the combination of a small number of latent abilities, which vary across individuals and tasks, generates the observed data. In order to join these traditionally disparate modeling approaches, Batchelder (1998, 2010) developed a *cognitive psychometrics* technique in which both modeling approaches are sequentially applied in order to infer differences with respect to a particular aspect of cognition across items or between groups of individuals (e.g., Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002).

More recently, Vandekerckhove (2014) introduced *cognitive latent variable models*, a unified method for implementing both a cognitive model and a latent variable model simultaneously as components of a single hierarchical Bayesian model. In this paper, we demonstrate how cognitive latent variable models may be constructed from cognitive and latent variable model components. We begin with an overview of the theoretical background for this new class of models, then outline a general procedure for implementing cognitive latent variable models in JAGS (Plummer, 2003), a software package that is commonly used to estimate parameters in complex Bayesian models.

Cognitive models

Cognitive models, also known as cognitive process models, propose that the observed data are the result of a psychological process which the model is designed to express quantitatively. Because each parameter in the model is included to capture a different aspect of the underlying process, cognitive model parameters often have meaningful psychological interpretations. Distributions of the estimated parameters, which may or may not be specific to participants or items, are used as the basis for inferences about unobserved cognitive abilities and other constructs.

A primary benefit of this approach is that it allows the researcher to disentangle the influence of psychological constructs of interest from other sources of variability. Consider a two-alternative forced choice recognition memory task: In the absence of a cognitive model, the raw response time data might be used directly to conclude that participants who have a higher mean response time have a inferior memory ability. However, this assumes that response time is a measure of an individual's memory ability and their memory ability only. More realistically, it may be that certain factors irrelevant of the memory demand of the task, such as a cautious individual's desire to avoid responding carelessly in a laboratory setting, or a tired individual's slight general delay in pressing the response buttons, are contributing substantively and systematically to the observed response times. Diffusion models (e.g., Stone, 1960) are a type of cognitive model that account for the interplay among exactly these type of factors in response time data. A diffusion model can be used to infer separately each participant's ability, cautiousness, and overall delay in response by including each as a parameter in a decision accumulation process (as discussed in the *Cognitive components* section). In this way, cognitive models provide a way for researchers to make inferences only about the particular cognitive ability that they initially intended to study.

By directly incorporating domain knowledge about the psychological construct and the context in which it is being studied, cognitive models allow for more meaningful conclusions to be made based on the observed data. However, most cognitive models are designed to analyze data from a single experimental paradigm. In order to analyze data from multiple different types of tasks simultaneously, a different approach to data analysis is typically used.

Latent variable models

Latent variable models, also known as psychometric models, propose that each observed datum (e.g., for a given participant on a given task) is the result of a linear combination of underlying variables, plus error. In contrast to the cognitive modeling approach, in a latent variable model, the underlying variables need not have psychological interpretations. Instead, the goal of this approach is sometimes simply to compress the data by identifying a small number of latent variables that account for most of the variability in the observed data (see Bollen, 2002, for an overview of perspectives).

From a psychometric perspective, these models are seen as inferring the *true scores* by accounting for individual and task-specific differences in the observed scores. These true scores are determined by the small set of latent variables (e.g., factors in factor analysis, components in principal component analysis). Discrepancies between the observed and true scores are assumed to be due to measurement error. In the context of cognitive science research, these models are often used to attempt to uncover a small number of cognitive factors that together predict performance across a diverse set of tasks, or to uncover a few underlying psychological constructs that are each measured by a large number of task items. An emphasis is often placed on the latent variable model's utility in disentangling unique properties of participants or items, or in determining whether theorized constructs are truly separable, depending on the research context. The primary benefit of this approach is that it allows the researcher to provide a comprehensive, yet simply expressed, account of the nature of large-scale, unobservable influences while accounting for individual and task differences.

Cognitive latent variable modeling

Often, researchers who wish to make inferences about latent cognitive ability from observed data using traditional models will operate in one of these two data analysis paradigms. However, this effectively forces a choice between speaking about large-scale latent constructs (such as attention and working memory) or psychological processes (such as signal detection or categorization rules). In order to most directly infer what unobserved cognitive abilities underlie observed data across many cognitive tasks, it would be advantageous to combine the two approaches.

This is especially desirable when the experimental design involves different types of data being recorded in different tasks. In order to analyze data from all tasks simultaneously, a psychometric model may be used to infer abilities that predict performance across tasks. However, this assumes that the data are an excellent proxy for cognitive ability, which, as discussed earlier, may not be so. In order to most directly assess the contribution of cognitive ability to the observed data, a cognitive model appropriate for each data type may be used to infer ability separately from other psychological factors (e.g., bias, risk-taking) based on the observed data in each task. To analyze data from all tasks in a way that assesses both individual differences and psychological processes, a researcher may wish to use both modeling approaches to analyze their data. A simple way to accomplish this would be to apply each type of model in sequence, as in Batchelder's cognitive psychometrics technique (e.g., Batchelder, 2010).

In a sequential procedure, the data for all participants from each task would be submitted to the cognitive model separately. The resulting estimates of a parameter selected by the researcher (e.g., that which most closely reflects cognitive ability) for each participant on each task would then be submitted to the psychometric model. This second model would in turn infer a small set of latent variables that predict the cognitive ability parameter values across tasks. However, this procedure requires that the output of the cognitive model be reduced from posterior distributions to point estimates. A general benefit of Bayesian methods is that uncertainty is naturally quantified through the spread of the posterior distribution. By discarding this information, a researcher risks introducing bias, especially if the posterior distributions are notably skewed or extremely wide (Pagan, 1984). At a minimum, the sequential approach leads to a loss of power, by collapsing many data points (i.e., scores on many trials) to a single summary statistic (i.e., a cognitive parameter estimate).

An alternative approach to merging the cognitive and psychometric approaches is to implement both simultaneously by constructing a *cognitive latent variable model* (Vandekerckhove, 2014). This is accomplished by treating one or more models of each type as components of a single hierarchical Bayesian model. Specifically, the model components are linked by allowing the large-scale latent ability parameters of the psychometric model to combine to determine a selected parameter of the cognitive model, which, in conjunction with the other cognitive model parameters, describes the distribution of the observed data. This composite model is therefore a single-step, coherent inference procedure.

From a statistical standpoint, this approach is preferable as it allows for uncertainty to be propagated from the cognitive model component to the psychometric component through the simultaneous estimation of all model parameters. However, the true appeal of the cognitive latent variable modeling technique is that it allows for more powerful statements to be made about the latent structure of cognitive abilities. From a psychological perspective, the benefit of this approach is the unique statement made by the composite model, that the observed data are the result of unobserved psychological processes, which are themselves heavily influenced by a few unobserved psychological constructs.

Goals of this tutorial

In this tutorial paper, we demonstrate how a cognitive latent variable model (CLVM) may be constructed from cognitive model and psychometric model components using freely available software packages. We first explain the how each model component may be implemented as a Bayesian model, beginning with defining statistical statements, and building up to model specification code. We then demonstrate how these components may be fused to create a CLVM, again including full model specification code. By including CLVMs that incorporate different cognitive components, we are able to showcase the flexibility of the framework and how the general CLVM approach may be tailored to different research contexts. We also demonstrate how different hypotheses about the latent structure of psychological constructs may be expressed in the psychometric component in order to show how CLVMs may be used for theory testing. Finally, we apply the set of CLVMs under consideration to a large, heterogeneous attention dataset and discuss practical details such as our use of selected software features to facilitate our chosen method of model comparison. In our discussion of the implications of the model comparison results for attention research, we review the unique benefits of this comprehensive new approach to modeling cognition.

II. Case example: A comparative evaluation of attention theories

Attention is one of a suite of higher-order cognitive functions, along with metacognition, decision-making, and working memory abilities. However, as attention is a latent construct, it cannot be directly measured. A multitude of tasks have been designed to assess attention indirectly (for an overview, see Lezak, Howieson, & Loring, 2004), each of which taps one or more aspects of the attention construct. However, it is unclear whether these aspects should be viewed as independent component abilities or as merely superficially different manifestations of a unitary latent ability.

Posner's theory of attention is a predominant theory of the former type. It posits that attention comprises three functionally and anatomically separable abilities including *alerting*, the ability to sustain attention over time, *orienting*, the ability to restrict the focus of attention to a small number of stimuli, and *executive control*, the ability to shift attention allocation to meet task demands (Posner & Petersen, 1990). This theory has heavily influenced research across psychology for nearly three decades (Raz & Buhle, 2006; Petersen & Posner, 2012).

The independence of these abilities is supported by behavioral, clinical, and imaging research (Fernandez-Duque & Posner, 2001; Petersen & Posner, 2012). While some cortical areas are believed to be common substrates for all three attention abilities, there is evidence of distinct neural substrates each for alerting, orienting, and control (Posner & Fan, 2004; Yin et al., 2012), which suggests anatomical independence for each network. Their separability is further supported by clinical research that observes selective deficits in just one of the three attentional networks (e.g., Wang et al., 2005). Their functional independence is supported by behavioral data collected primarily through the use of the Attention Network Test (Fan, McCandliss, Sommer, Raz, & Posner, 2002; Fan, McCandliss, Fosella, Flombaum, & Posner, 2005), which was designed to measure alerting, orienting, and control abilities separately through differences in mean response time.

Other research has suggested that these abilities many not be separable. There is now a growing body of evidence that some or all of the attention networks proposed by Posner and Petersen interact in replicable ways (Callejas, Lupianez, Funes, & Tudela, 2005; Fuentes & Campoy, 2008; Fan et al., 2009; Weinbach & Henik, 2012; McConnell & Shore, 2011). However, the interpretation of these findings is debatable. It may be that the alerting, orienting, and control are not independent abilities, or it may be that, although the networks interact during cognitive tasks, the distinction between the three abilities is significant and useful (Fan et al., 2009).

Finally, other theoretical accounts of the latent structure of attention and related executive cognitive abilities exist that propose alternative divisions (e.g., Mirsky, Anthony, Duncan, Ahearn, & Kellam, 1991; Miyake et al., 2000). Thus, exactly which component cognitive abilities give rise to the psychological construct of attention is still an open question.

In the case example described here, we sought to resolve this issue as directly as possible through the use of multiple, competing CLVMs. Specifically, we explored which of seven theoretical accounts of latent attention ability is best able to describe participants' performance on a small battery of attention tasks. In the following sections, we describe this research with a heavy emphasis on the development, testing, and use of CLVMs.

Attention dataset

Participants. In our preregistration (`https://osf.io/qxk2s/`), we declared our intention to collect data from 50–70 participants based on previous similar work (Pe, Vandekerckhove, & Kuppens, 2013; Vandekerckhove, 2014). 60 undergraduates (54 female, 6 male) participated in the experiment. Mean age of these participants was 22.22 (range: 18–64).

Participants reported ethnicities including Asian (66.7%), Latinx (21.7%), White (5.0%), Black (3.3%), and Pacific Islander (3.3%).

Only partial data was collected from three participants due to computer failures; data from these participants was subsequently excluded. An additional seven participants were excluded due to lack of engagement[1] during one or more tasks. Ultimately, data from 50 participants was included for analysis.

Materials & Procedure. All participants completed a small battery of computerized attention tasks, presented in a random order using PsychoPy (Peirce, 2007). This 1.5-hour battery included five commonly used response-time tests and two self-report scales:

- *Attention Network Test (ANT):* This combined cueing and flanker test was explicitly designed to assess all three of Posner's theorized components of attention (e.g., Posner & Petersen, 1990), including alerting, orienting, and conflict (Fan et al., 2002, 2005). Each trial began with one of three cue types. If a cue was presented, it indicated that the target was about to appear and, in some trial types, also indicated the upcoming target's location. The target arrow, which appeared randomly on the top or bottom of the screen, was flanked by arrows pointing in either a congruent or incongruent direction.[2] Participants' goal in each trial was to judge the direction of the center arrow with a keypress (see Figure 3.1). The full combination of cue and flanker conditions allowed for six trial types in total.

- *Continuous Performance Test, X version (CPT-X):* In this intentionally repetitive and monotonous test of sustained attention ability, participants judge whether each item

[1] A lack of engagement was operationalized as (1) taking more than 2 hours to complete the battery and/or (2) accuracy at chance for one or more of the response time tests. This latter criterion was assessed using a Bayes factor (BF) comparing the null hypothesis that a participant was responding with chance accuracy, $\mathcal{H}_0 : p = 0.5$, to the alternative hypothesis that a participant was responding with some level of accuracy above chance, $\mathcal{H}_1 : p \sim \text{Uniform}(.5, 1)$. For this analysis, a lack of response was considered an incorrect response. If the Bayes factor for any of the five response time tests provided any level of evidence for the null hypothesis over the alternative ($BF_{01} > 1$), the participant was excluded. This assessment was part of our automated data preprocessing (using preregistered code available at `https://osf.io/qhhvg/`).

[2] We also included trials with neutral flankers (dashes) in order to more closely duplicate the original design of the ANT. However, because these trials were not used to assess latent attention abilities in the original research (Fan et al., 2005), we likewise discarded this data before data processing and data analysis.

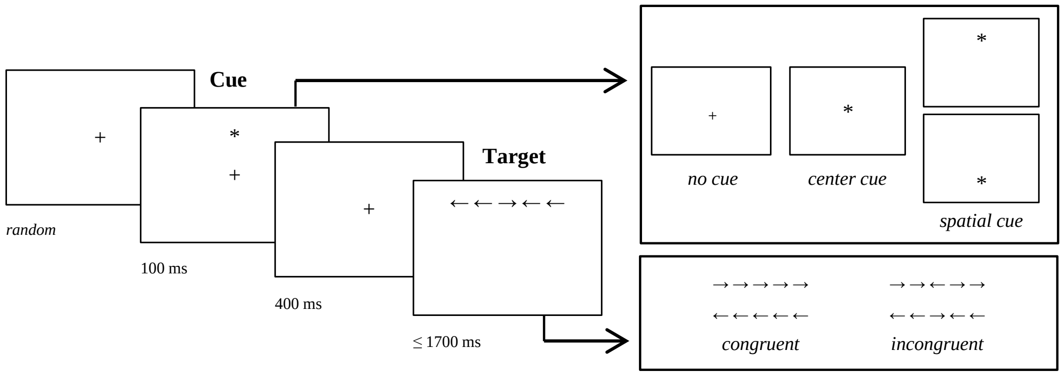


Figure 3.1. *Attention Network Test.* At left, a cascade shows the sequence of presentation and timing for a single 3500 ms trial. At right, cue types and flanker types are shown.

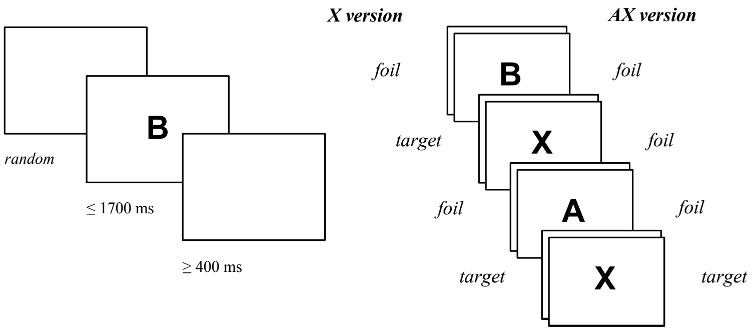


Figure 3.2. *Continuous Performance Tests.* On the left, a cascade shows the timing for a single ≤ 2500 ms trial. On the right, how the trial type (target vs. foil) and, consequently, the correct response to a given stimulus, depends on the version of the test is shown for a sequence of four trials.

in a long sequence does or does not follow a rule (Rosvold, Mirsky, Sarason, Bransome, & Beck, 1956; Conners & Staff, 2000). We adapted two versions of this test (see Figure 3.2) such that a response should be given on all trials.[3] In the *X* version, the rule is simple: If the letter presented is X, the stimulus is a target; if the letter presented is not X, it is a foil.

- *Continuous Performance Test, AX version (CPT-AX):* In the *AX* version, the rule is

[3] In traditional continuous performance tests, a response is given for foil trials, but withheld for target trials (Conners & Staff, 2000), or vice versa (Rosvold et al., 1956). In either method, it is intentional that for some trials no data is collected. In order to model performance on the entire CPT, we instructed participants to respond with one keypress for foils (i.e., the F key) and a different keypress for targets (the J key). This mild change was instituted so that response time and accuracy data were generated for all CPT trial types. Because the expected effects of trial type on performance were still observed (see Appendix A), we conclude that the dual-response adaptation was successful.

more complex: If the letter presented is X, and the previous letter was A, the stimulus is a target; if the letter presented is not X and/or the previous letter was not A, the stimulus is a foil. Because targets are infrequent, the participant must remain vigilant in order to quickly and effectively switch response modes when a rare target trial does occur.

- *Number-Letter (NL):* This response time test was designed to measure the effect of switching the focus of attention (Rogers & Monsell, 1995). In each trial, a stimulus consisting of a number and a letter as a pair (e.g., 7A) was presented. If the stimulus appeared on the top half of the screen, the task was to judge whether the number is odd or even; if the stimulus appeared on the bottom half of the screen, the task was to judge whether the letter is a consonant or a vowel (see Figure 3.3). Trials where the previous trial required a different judgment are classified as switch trials. Trials where the previous trial required the same judgment were designated as no-switch trials.

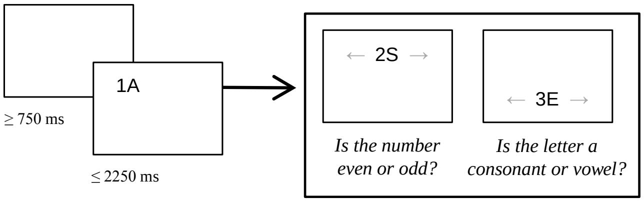

1A
≥ 750 ms
≤ 2250 ms
← 2S →
← 3E →
Is the number even or odd?
Is the letter a consonant or vowel?

Figure 3.3. Number-Letter. A cascade (left) shows the timing for a single ≤ 3000 ms trial. The vertical position of the stimulus determines the required judgment (right). The trial type (switch vs. no switch) depends on the judgment required on the previous trial (not pictured).

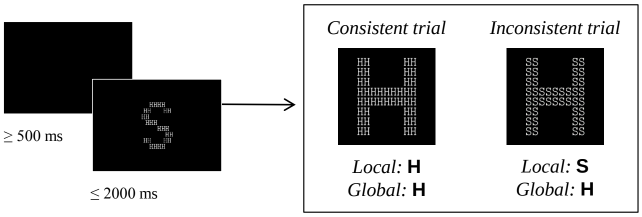


Figure 3.4. Local-Global. A cascade (left) shows the timing for a single ≤ 2500 ms trial. The correct response depends on whether the instruction for the current block is to judge the local or the global feature (right).

- *Local-Global (LG):* This response time test was included to probe selective attention and inhibition. In each trial, a Navon figure (i.e., a letter composed of smaller letters Navon, 1977) was presented. Navon stimuli were either consistent (e.g., H composed of smaller Hs) or inconsistent (e.g., S composed of smaller Hs). In each block, the task was to identify either the global feature, meaning the large-scale letter, or the local feature, meaning the small-scale letter, with the corresponding keypress (see Figure 3.4). Participants were expected to exhibit a global precedence effect such that global features are automatically more readily perceived, and the reporting of local features would require suppression of the global percept. The full combination of feature and consistency conditions allowed for four trial types in total.

- *Attentional Function Index (AFI):* This scale, developed for use in clinical research, measures the self-reported ability to accomplish everyday tasks that rely on attention ability in a general sense (see Figure 3.5A; Cimprich, Visovatti, & Ronis, 2011)[4].

- *Attention-Related Cognitive Errors Scale (ARCES):* This scale, developed for use in cognitive research, measures the self-perceived frequency of everyday behavioral errors caused by lapses in sustained attention (see Figure 3.5B; Cheyne, Carriere, & Smilek, 2006).

(A) AFI

At this time, how well do you feel you are functioning in each of the areas below?

1. Getting started on activities (tasks, jobs) you intend to do.

Not at all □1 □2 ... □10 Extremely well

(B) ARCES

Use the scale to indicate how frequently you have the following experiences.

6. I begin one task and get distracted into doing something else.

Never □1 □2 □3 □4 □5 Very often

Figure 3.5. *Example items from each survey task.*

[4] The original pencil-and-paper version of the AFI allowed participants to mark their response to each item on a 100mm line. To adapt this scale for presentation on a computer, we replaced the response line with a 10-point Likert scale.

Demographic information was collected separately on paper using three free-response items at the conclusion of the experiment.

Because participants' performance on each trial type of a given response time test and on each survey might rely on different components of attention, we treated each trial type and survey as a separate task for the purposes of our analyses. Across the resultant 18 tasks, data types included response times, choice accuracy, and Likert scale ratings. During automated data preprocessing, data from trials where the participant did not respond were excluded (1.8% of the data). For the response time tasks, response times that were unrealistically fast, as determined by an exponentially weighted moving average procedure[5], were also excluded (2.9% of the response time data; 2.8% of the data). Planned manipulation checks were performed to ensure that the expected effects for each response time test were observed. Descriptive statistics and manipulation check results are presented in Appendix A.

Model design

Cognitive components. Data from 16 of the 18 tasks included response time data. As discussed earlier, although raw response time is often treated as a proxy of cognitive ability, it is a complex and noisy measure. The two models we selected as options for the cognitive component of the CLVM were designed to estimate task-specific ability separately from other sources of variability in observed response times. Both are evidence accumulation process models.

This general type of model proposes that when a stimulus is presented, an individual samples information from the stimulus sequentially in time. Each sample provides some amount of information, or *evidence*, for or against a choice. Samples continue to be collected over time until the amount of evidence accumulated passes a threshold, at which point a decision is made. This idea successfully accounts for latencies in response time in simple re-

[5] This algorithm determined the lowest response time at which a participant responds above chance accuracy, then censored response times below that threshold (for further details, see Vandekerckhove & Tuerlinckx, 2007). Code for this function is available at `osf.io/x99sz/`.

action time tasks beyond that which is accounted for by conductance lag and motor response execution (Carpenter, 1981).

Two features of this process are captured by most evidence accumulation models. The *speed of evidence accumulation* reflects an individual's ability to complete the task at hand. An individual who is skilled at the task will gain more information with each sample from the stimulus, and therefore will accumulate enough evidence to make a decision more quickly than an individual who is less skilled at the task and gains less information with each sample from the stimulus. The *evidence threshold* reflects an individual's personal level of cautiousness in making a decision. An individual who is cautious may set a high threshold of evidence for making a decision, and thus require more samples to be accumulated before making a decision, leading to longer response times overall. More complex models include parameters that capture additional features of this proposed process (e.g., Stone, 1960; Ratcliff & Rouder, 1998).

A benefit of these models is that they naturally account for speed-accuracy tradeoffs (i.e., the finding that slow decisions tend to be accurate, while fast decisions tend to be error-prone; Wickelgren, 1977). Given a high speed of evidence accumulation, a low threshold will lead to fast response times, however, the decisions made will be less accurate as less samples are required to reach the decision threshold. In contrast, a high threshold will lead to slower response times, however, the decisions that are ultimately made will be more accurate, as more samples from the stimulus were collected before reaching the decision threshold.

The first option for the cognitive component is the LATER model (Linear Approach to Threshold with Ergodic Rate; Carpenter, 1981). In this model, the speed of evidence accumulation is captured by the *mean decision rate* parameter, ν, and the evidence threshold is captured by the *threshold* parameter, θ. Thus, ν reflects task-specific ability, and θ reflects an individual's cautiousness. The accumulation process is deterministic and is assumed to be linear (see Figure 3.6A). The observed response time, x, for an individual participant, p, on each trial, i, is determined by the following relationship:

$$\frac{1}{x_{pi}} = \frac{\gamma_{pi}}{\theta_p}$$

where the trial-specific decision rate, γ, is a draw from a hierarchical distribution of rates:

$$\gamma_{pi} \sim \text{Normal}(\nu_p, \omega_p^2)$$

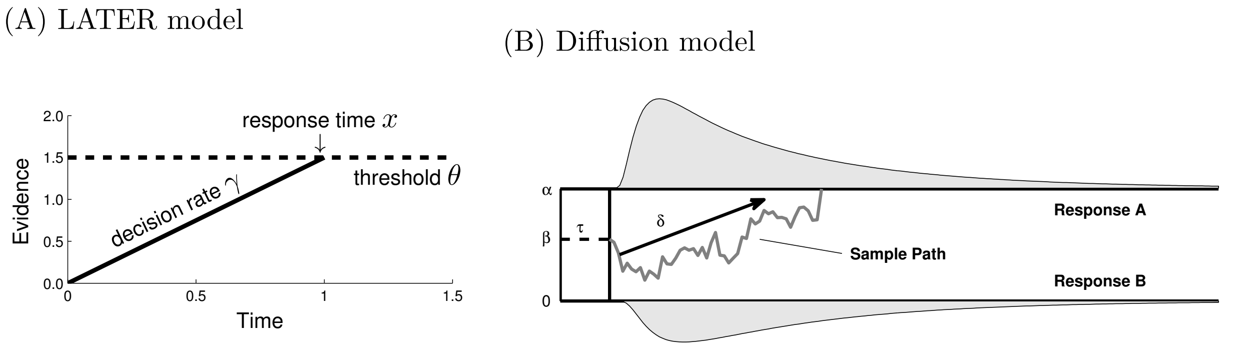
If ω^2 is fixed to 1 for all participants, combining the two equations above gives the defining equation of the model:

$$\frac{1}{x_{pi}} \sim \text{Normal}(\frac{\nu_p}{\theta_p}, \frac{1}{\theta_p^2}) \tag{1}$$

To implement this as a Bayesian model, we must also set priors on the parameters that will be estimated:

$$\begin{aligned} \nu_p &\sim \text{Normal}(0, 1) \\ \theta_p &\sim \text{Uniform}(0, 10) \end{aligned} \tag{2}$$

In this way, the LATER model is able to quantitatively describe a possible cognitive process that might underlie observed response times. However, the LATER model is not able to account for response accuracy, which is known in our attention dataset. In order to incorporate both sources of information, we include another evidence accumulator model, the diffusion model, as a second option for the cognitive component.



(A) LATER model

(B) Diffusion model

Figure 3.6. *Schematic representation of the cognitive components.*

Various versions of the diffusion model have been widely and successfully used in cognitive science to address the underlying cognitive processes that produce response times in binary choice contexts (Wagenmakers, 2009). In the version of the diffusion model that we elected to use (Stone, 1960; see Figure 3.6B), task-specific cognitive ability is captured by the *drift rate* parameter, δ, which expresses the average rate of the stochastic evidence accumulation process. Because the diffusion model describes a choice between two distinct response options, there is now a threshold for each possible response. Thus, the threshold parameter is reconceptualized as the distance between two thresholds, or the *boundary separation*, α. Because a wider boundary separation implies that more evidence must be accumulated to reach either decision threshold, α is interpreted as an individual's cautiousness. As a second result of this reconceptualization, the model also includes a *bias* parameter, β, to reflect where the evidence accumulation process begins relative to the two thresholds. If β is high, then less evidence is required to reach the top threshold than is required to reach the bottom threshold; if β is low, then more evidence is required to reach the top threshold than is required to reach the bottom threshold. As such, β is interpreted as a bias toward either response before the stimulus is seen (or no bias, if $\beta = \frac{1}{2}$). The final parameter of the diffusion model is the *non-decision time* parameter, τ. This accounts for time that must be reserved for stimulus encoding and motor response execution, and thus is not reflective of the cognitive evidence accumulation process.

The distribution of choice response times for a given participant on each trial is described as a Wiener diffusion process:

$$x_{pi} \sim \text{Wiener}(\alpha_p, \beta_p, \delta_p, \tau_p) \tag{3}$$

where x represents signed response times[6]. In our use of the diffusion model, we defined the two choice options (*Response A* and *Response B* in 3.6B) as the correct versus incorrect

[6] Such that $x = ax_{raw}$, where x_{raw} is raw response time and a is choice accuracy, defined as 1 for correct responses and -1 for incorrect responses.

choice on a given trial. Because of this, the bias parameter, β, may be set to 0.5, as we expect no bias toward the correct or incorrect response before the stimulus is seen. To implement this as a Bayesian model, we also set priors on each parameter to be estimated:

$$\begin{gathered}\alpha_p \sim \text{Uniform}(0.01, 4)\\ \beta_p = 0.5\\ \delta_p \sim \text{Normal}(0, 1)\\ \tau_p \sim \text{Uniform}(0.01, 1)\end{gathered} \tag{4}$$

We will return to the equations for the LATER model and the diffusion model in a later section when we establish the model specification for each component in JAGS.

Psychometric component. Factor analysis is used to reduce a number of observed variables to a smaller set of latent variables. When the observed variables are data from cognitive tasks, the latent variables may be interpreted as unobserved cognitive constructs. For this analysis, we elected to use *confirmatory factor analysis* as it as it allows us to express different theoretical accounts of which latent variables might comprise the attention construct a priori.

The general factor analysis model can be written as a single mathematical statement:

$$\mathrm{X} = \iota + \Lambda\Phi + \varepsilon$$

In this equation, X is a matrix of observed data for N tasks from P participants. This data is primarily determined by a product of the factor loadings matrix, Λ, and the factor scores matrix, Φ. In the factor loadings matrix, Λ, each row represents a task, and each column represents one of the F proposed factors. As such, each entry, λ_{nf}, expresses the strength and direction of the relationship between observed scores on the nth task (e.g., the spatial cue/congruent flanker condition of the ANT) to an unobserved fth factor (e.g., alerting ability). If an entry λ_{nf} of the loadings matrix is 0, it would indicate no relationship between the nth task and the fth factor; if an entry λ_{nf} of the loadings matrix is -.8, it would indicate an inverse relationship. In the factor scores matrix, Φ, each of the F rows represents a factor, and each of the P columns represents a participant. This is interpreted similarly: Each entry, ϕ_{fp} expresses how an individual participant p would score on the unobserved fth factor (e.g., orienting ability), if that could be measured. Lastly, ι is an N-length vector of intercepts, or baselines scores for each task.

The product of the factor loadings matrix and the factor scores matrix, $\Lambda\Phi$, plus the corresponding intercepts, ι, is an $N \times P$ matrix of the *true scores* for each participant on each task, which we denote M. Some error of measurement, ε, is added to these true scores to reflect the discrepancy between the true scores, M, and the observed scores, X.

In order to implement factor analysis as a Bayesian model, we rework the general factor analysis above by first defining the true scores in a deterministic statement:

$$\mu_{np} = \iota_n + \vec{\Lambda}_n \vec{\Phi}_p \tag{5}$$

then defining the distribution of the observed data stochastically, using a variance parameter, σ^2, in place of explicit error, in a second statement:

$$\mathrm{x}_{npi} \sim \text{Normal}(\mu_{np}, \sigma_n^2) \tag{6}$$

and finally setting priors for all free model parameters:

$$\iota_n \sim \text{Normal}(0, 1)$$

$$\lambda_{nf} \sim \text{Normal}(0, 1)$$

$$\phi_{fp} \sim \text{Normal}(0, 1)$$

$$\sigma_n \sim \text{Uniform}(0, 10)$$

However, as we will specifically use *confirmatory* factor analysis, not all entries in the factor loadings matrix will be freely estimated[7]; some entries will be fixed[8] to 0 or 1 in order to express our hypotheses about relationships between tasks and latent attention abilities.

To allow for clarity and brevity in our in-text model descriptions, we include a loadings matrix with fewer tasks ($N = 5$) than in our attention dataset and a small number of factors ($F = 2$). We refer to this example loadings matrix as Λ_0 to distinguish it from the seven loadings matrices used in our final data analysis (Λ_1–Λ_7). In the example loadings matrix, printed below, we use a common method which ensures model identification. Specifically, we fix the entries on the diagonal ($\lambda_{nf}, n = f$) to 1, fix entries above the diagonal ($\lambda_{nf}, n < f$) to 0, and allow all other entries to be estimated.

$$\Lambda_0 = \left[\begin{array}{cc} 1 & 0 \\ \lambda_{2,1} & 1 \\ \lambda_{3,1} & \lambda_{3,2} \\ \hline \lambda_{4,1} & \lambda_{4,2} \\ \lambda_{5,1} & \lambda_{5,2} \end{array}\right]$$

In this matrix, we will assume that the first three rows reflect data from tasks where response time and accuracy data were collected, and that the final two rows reflect data from tasks where survey data were collected.

To incorporate this example loadings matrix in the confirmatory factor analysis model, we redefine the priors above and set an assignment or prior for each loading in a separate

[7] We also do not allow all entries in the factor loadings matrix to be freely estimated to ensure model identification; including both a completely free factor loadings matrix and a completely free factor scores matrix would lead to an unidentified model.

[8] In a classical implementation of factor analysis, individual loadings are normalized ($-1 \leq \lambda_{nf} \leq 1, \forall n \in \{1, 2, ...N\}, \forall f \in \{1, 2, ...F\}$), and so fixing an entry to 1 implies that the corresponding task is a perfect measure of a given factor. This is not so in the Bayesian implementation as loadings are not normalized ($\lambda_{nf} \in \mathbb{R}, \forall n, f$). Fixing a single entry in a column of the loadings matrix to 1 simply allows that loading to serve as a scaling factor for the remaining free loadings on that factor. Thus, fixing a single entry implies only that the corresponding task has a *nonzero* relationship to the factor. Fixing multiple entries in a column to 1 implies that the corresponding tasks have the same strength relationship to that factor, and scales the remaining nonzero loadings.

statement:

$$
\begin{aligned}
\iota_n &\sim \text{Normal}(0,1) \\
\lambda_{1,1} &= 1 \\
\lambda_{1,2} &= 0 \\
\lambda_{2,1} &\sim \text{Normal}(0,1) \\
\lambda_{2,2} &= 1 \\
\lambda_{3,1} &\sim \text{Normal}(0,1) \\
\lambda_{3,2} &\sim \text{Normal}(0,1) \\
\lambda_{4,1} &\sim \text{Normal}(0,1) \\
\lambda_{4,2} &\sim \text{Normal}(0,1) \\
\lambda_{5,1} &\sim \text{Normal}(0,1) \\
\lambda_{5,2} &\sim \text{Normal}(0,1) \\
\phi_{fp} &\sim \text{Normal}(0,1) \\
\sigma_n &\sim \text{Uniform}(0,10)
\end{aligned}
\tag{7}
$$

The numbered equations above will determine our JAGS model specification for the psychometric component in the next section.

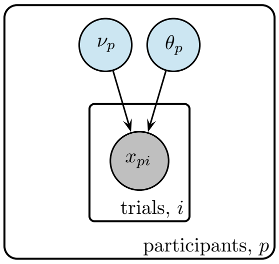
Cognitive latent variable models. To create each of our CLVMs, cognitive model and psychometric model components were simultaneously implemented in a hierarchical Bayesian model. In this demonstration, all model specifications are written for JAGS (Plummer, 2003), an automated Gibbs sampling package. We facilitated calls to JAGS from MATLAB by using the Trinity interface (Vandekerckhove, 2015). Readers may wish to consult Matzke, Boehm, and Vandekerckhove (2018) to establish familiarity with how these three utilities work in concert. (We also relied on the the `jags-wiener` module (Wabersich & Vandekerckhove, 2014) to implement the diffusion model in JAGS.)

The entire implementation process may be distilled to three steps:

Step 1: Write a separate model specification for the cognitive model component and for the *psychometric model component.*

We began by writing a JAGS model specification for each component as if we intended to use them separately. This is accomplished by taking the distributional statements that define each model, as outlined in the previous two sections, and and translating them, line by line, into JAGS code.

(A)

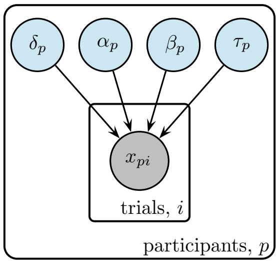


(B)

```
#data (eqn. 1)
for(p in 1:P){
   for(i in 1:I){
      x[p,i] ~ dnorm(nu[p]/theta[p],theta2[p])
   }
}

#priors (eqn. 2)
for(p in 1:P){
   nu[p] ~ dnorm(0,1)
   theta2[p] <- pow(theta[p],2)
   theta[p] ~ dunif(0,10)
   }
}
```

(C)



(D)

```
#data (eqn. 3)
for(p in 1:P){
   for(i in 1:I){
      x[p,i] ~ dwiener(alpha[p],tau[p],0.5,  delta[p])
   }
}

#priors (eqn. 4)
for(p in 1:P){
   alpha[p] ~ dunif(.01,4)
   tau[p] ~ dunif(.01,1)
   delta[p] ~ dnorm(0,1)
}
```

Figure 3.7. Cognitive components. Graphical model (A) and model specification (B) for the LATER model; graphical model (C) and model specification (D) for the diffusion model.

The JAGS model specification for the LATER model, one of the two options for the cognitive component, is presented in Figure 3.7B. This is a direct translation of the likelihood distribution in Equation 1 and the prior distributions in Equation 2. A model specification for the diffusion model (see Figure 3.7D) can be created in a similar fashion from the likelihood distribution in Equation 3 and the prior distributions in Equation 4. Graphical models for each cognitive component are presented in parts A and C of Figure 3.7.

The JAGS model specification for confirmatory factor analysis, the latent variable component, is presented in Figure 3.8B. This is a direct translation of Equations 5, 6, and 7. The graphical model for this component is presented in part A of Figure 3.8.

Translation of these equations does require the addition of a few lines of code to reflect the use of precision, rather than variance or standard deviation, to parameterize the normal distribution in JAGS. For example, in the model specification for factor analysis, the second argument to `dnorm` is not the standard deviation, σ, on which we have set a prior, but rather the precision, $\frac{1}{\sigma^2}$. We accommodate this alternative parameterization by using an intermediate variable, `invsigma2`, to indicate the inverted and squared version of `sigma` (see Line 26 in Figure 3.8B). Similarly, in the model specification for the LATER model, the desired variance, $\frac{1}{\theta^2}$ is already inverted, so we need only to square the threshold, `theta`, to generate the correct precision, `theta2` (see Line 11 in Figure 3.7B).

If there are data from any task that are not applicable as input to the cognitive model, it is during this step that JAGS code for the distribution of data from these tasks and for any necessary priors should be written. In our case example, data from the cognitive tests, x, is modeled as an evidence accumulation process, but data from the self-report scales, y, is not. Instead, we assume that the self-report scale data, when standardized, is normally distributed:

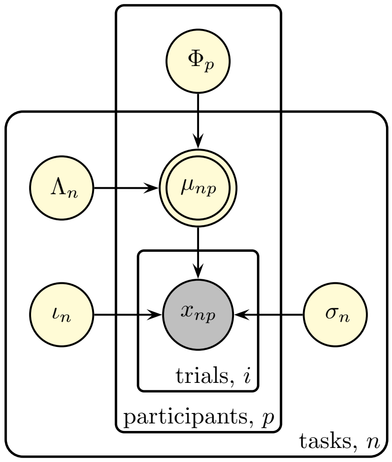
$$\mathrm{y}_{pi} \sim \mathrm{Normal}(\kappa_p, \eta^2) \tag{8}$$

with associated priors:

$$\begin{aligned} \kappa_p &\sim \mathrm{Normal}(0, 1) \\ \eta &\sim \mathrm{Uniform}(0, 5) \end{aligned} \tag{9}$$

Based on the above equations, we write a few lines of JAGS code for the distribution of the scale data and its associated priors. Again, this code is written as if the data from each self-report scale were the only data to be modeled. This model specification and a representative graphical model are presented in Figure 3.9. As in the specification for the confirmatory factor analysis model, we add a line to reflect the transformation of the standard deviation, η, to a precision $\frac{1}{\eta^2}$ (see Line 11 in Figure 3.9B).

(A)



(B)

```
#data (eqn. 6)
for(n in 1:N){
   for(p in 1:P){
      for(i in 1:I){
         x[n,p,i] ~ dnorm(mu[n,p],invsigma2[n])
      }
   }
}

#factor analysis, true scores (eqn. 5)
for(n in 1:N){
  for(p in 1:P){
    mu[n,p] <- iota[n] +                     inprod(lambda[
         n,],phi[,p])
  }
}

#priors (eqn. 7)
lambda[1,1] <- 1
lambda[1,2] <- 0
lambda[2,1] <- 1
lambda[2,2] ~ dnorm(0,1)
lambda[3,1] ~ dnorm(0,1)
lambda[3,2] ~ dnorm(0,1)
lambda[4,1] ~ dnorm(0,1)
lambda[4,2] ~ dnorm(0,1)
lambda[5,1] ~ dnorm(0,1)
lambda[5,2] ~ dnorm(0,1)
for(n in 1:N){
   iota[n] ~ dnorm(0,1)
   invsigma2[n] <- pow(sigma[n],-2)
   sigma[n] ~ dunif(0,10)
}
for(f in 1:F){
   for(p in 1:P){
      phi[f,p] ~ dnorm(0,1)
   }
}
```

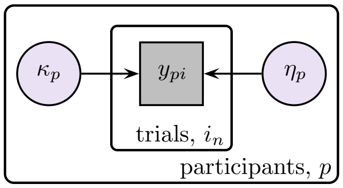
Figure 3.8. Psychometric component. Graphical model (A) and model specification (B) for confirmatory factor analysis.

Step 2: Identify which parameter in the cognitive model is most closely related to the latent constructs captured by the psychometric model.

In the confirmatory factor analysis model, the matrix of true scores, $M = \Lambda\Phi$, captures all information about the latent cognitive constructs. This is due to the reliance of the true scores, M, on the proposed factors through the factor loadings matrix, Λ, which captures relationships between latent constructs and tasks, and the factor scores matrix, Φ, which captures relationships between latent constructs and individual participants.

To create the CLVM, we will allow the unobserved cognitive constructs inferred by the psychometric model, as expressed in the matrix of true scores, to fully determine a parameter

(A)



(B)

```
#data (eqn. 8)
for(p in 1:P){
   for(i in 1:I){
      y[p,i] ~ dnorm(kappa[p],inveta2)
   }
}

#priors (eqn. 9)
for(p in 1:P){
   kappa[p] ~ dnorm(0,1)
}
inveta2 <- pow(eta,-2)
eta ~ dunif(0,5)
```

Figure 3.9. Covariate data model. Graphical model (A) and model specification (B) for the normally-distributed survey task data.

of the data distribution in the cognitive model and of the normal distribution for the scale data. In this step, we must therefore select which parameter in each distribution is most suitable for this assignment based on domain knowledge. Specifically, we will select the parameter in each data distribution that we expect to most closely relate to latent attention ability.

As mentioned in the *Cognitive components* section above, task-specific ability is captured by the mean decision rate, ν, in the LATER model, and by the drift rate, δ, in the diffusion model. In the normal distribution of the survey data, the mean score, κ, should reflect (self-perceived) attention ability. Thus, we select these three variables to be determined by the psychometric component in the next step.

Because we will specifically link these variables to the true scores, μ, special care should be taken to ensure that the dimensions for all selected parameters can reasonably be made to match the dimensions of μ in the next step. In the psychometric component, μ is indexed by task and participant. In order to link this to the data distributions for the response time data and the scale data in the next step, all of the selected parameters must be made to vary across tasks in addition to varying across participants.

It is sensible to expect that the task-specific ability parameter values in the cognitive components, ν and δ, will differ across the response time tasks. Thus, before moving on to the next step, we will edit each cognitive model specification such that the response time data, `x`, and the selected parameters for each model, `nu` and `delta`, are now indexed by task in addition to their respective existing indices (i.e., `x[`***n,*** `p,i]`, `nu[`***n,*** `p]`, `delta[`***n,*** `p]`). To accommodate these new indices, we also add loops over the response time tasks (n = 1–3 in our example loadings matrix, Λ_0) where necessary.

For the survey data, it is sensible to expect that the mean score parameter values, κ, will differ across the two survey tasks. Therefore, we will edit the model specification for the survey data similarly such that the data, `y`, and the mean score, `kappa`, are now indexed by task (i.e, `y[`***n,*** `p,i]`, `kappa[`***n,*** `p]`). Finally, loops over the survey tasks (n = 4–5 in our example loadings matrix, Λ_0) are added where necessary.

Step 3: Link the two component models hierarchically by replacing the selected cognitive model parameter(s) with the selected parameter(s) of the psychometric model.

To compile the model specification for each version of the the CLVM, we begin by pooling all relevant model specifications from the previous steps in a single text document. For the LATER model version, we collect the code for the LATER model (as in Figure 3.7B), for the survey data (as in Figure 3.9B), and for confirmatory factor analysis (as in Figure 3.8B). For the diffusion model version, we collect the code for the diffusion model (as in Figure 3.7D), for the survey data (as in Figure 3.9B), and for confirmatory factor analysis (as in Figure 3.8B). To shape each assortment of code into a viable model specification, we will make some slight changes and deletions.

The first action will make the statistical statement that ultimately distinguishes the CLVM approach by linking the cognitive component to the psychometric component directly. Now, we replace the parameters selected in *Step 2* with the true scores, μ), in each data distribution statement. In each cognitive component, we replace the task-specific ability parameters, `nu[n,p]` and `delta[n,p]`, each with the true scores, `mu[n,p]`. In the survey data model specification, we will replace the mean score, `kappa[n,p]` with the true scores, `mu[n,p]`. As a consequence of this substitution, the selected parameters, ν, δ, and κ, are no longer used in the model. Therefore, we may delete the corresponding prior for each of these parameters (Line 10 in Figure 3.7B, Line 12 in Figure 3.7D, and Line 10 in Figure 3.9B). We should also now delete redundant statements. Specifically, we delete the now-unused likelihood distribution and its surrounding loops from the psychometric model component (Lines 1–8 in Figure 3.8B) in both versions. As a consequence of this deletion, the parameter σ is no longer used, and so the prior for `sigma` and its reparameterization as `invsigma2` (Lines 26 and 27 in Figure 3.8B) should be removed also.

Final model specifications for the LATER model version of the CLVM and the diffusion model version of the CLVM are presented in Appendix B in Figure B1 and B2 respectively, and represented as graphical models in Figure 3.10. In the model specifications presented in Appendix B, we rely on the example loadings matrix, Λ_0, for clarity. However, in the CLVMs we used to analyze our attention dataset, each model specification expressed one of seven loadings matrices.

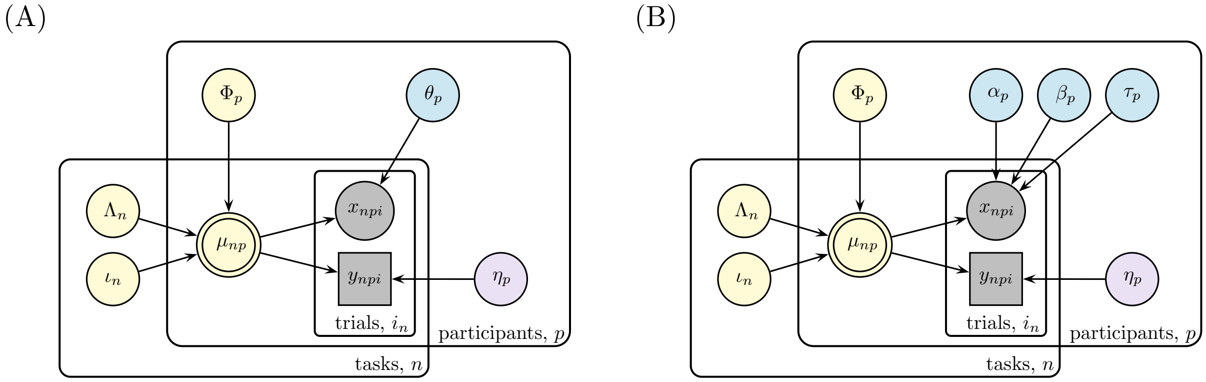


Figure 3.10. Graphical models for each version of the CLVM. Response time data (x) is LATER model-distributed in A and diffusion model-distributed in B. The survey data (y) is normally distributed in both A and B.

Loadings matrices. Each loadings matrix used in this analysis was designed to capture a different theoretical account of the latent structure of the psychological construct of attention. By incorporating each of the seven loadings matrices in each of seven otherwise identical CLVMs, we were able to compare the theoretical accounts of attention ability by performing a model comparison across the CLVMs. Below, we briefly describe the conceptual basis for each matrix. A summary is presented in Table 3.1 and a more detailed account of how we developed each loadings matrix, including their exact specification, is included in Appendix C.

Our first loadings matrix, Λ_1: *Posner*, was designed to express Posner and Petersen's (1990) theory of attention, as described at the beginning of this section. To capture their hypothesized three components of attention, we fixed a subset of entries in the loadings matrix to 0 or 1 based on the knowledge of what each test and each survey was designed to measure. In this way, we were able to define one factor each to reflect alerting, orienting, and control abilities.

Our second loadings matrix, Λ_2: *Posner + Working Memory*, extends Posner and Petersen's theory of attention through the addition of a fourth factor. Because two trial types of the CPT require that participants recall the stimulus from the last trial to respond correctly, we consider the possibility that disentangling working memory from attention abilities will allow for a better description of the observed data. Therefore, in this loadings matrix we include factors to reflect alerting, orienting, control, and working memory.

Λ_3: *Mirsky* reflects an alternative division of attention into component constructs proposed by Mirsky and colleagues (1991). While the components, sustain, focus-execute, shifting, and encode, are similar to those expressed in Λ_2, alerting, orienting, and control, and working memory, respectively, Mirsky's description of the some factors differs substantively in a number of ways. For example, Mirsky's shifting component only refers to breaking an attentional set, while Posner's control is much more broadly defined.

The next two loadings matrices, Λ_4 and Λ_5, blend ideas from Mirsky (Mirsky et al., 1991) and Miyake (Miyake et al., 2000) and so are referred to as *Composite 1* and *Composite 2* respectively. One of three constructs uncovered by Miyake and colleagues in their research on executive functions is *inhibition*, the specific ability to suppress a response. This is in contrast to shifting, present in both Miyake and Mirsky's theories, which emphasizes breaking and reforming an attentional set or filter. Both Λ_4 and Λ_5 offer two plausible theoretical blends of these theories of attention and executive function.

Our final two loadings matrices are unlike the previous matrices in that they do not divide attention into component abilities. Instead, we consider the possibility that all proposed divisions are superficial, and therefore attention is best described a single ability. In Λ_6: *Unitary*, the sole factor included is a general attention factor which relates to scores on all tasks. In Λ_7: *Unitary + Working Memory*, we propose that the only useful distinction is that between attention and memory. Thus, this final loadings matrix includes two factors: a general attention factor, and a memory factor which only relates to the two tasks with a known memory load.

Table 3.1
Summary of theories under consideration

	Sustained	Selective	Shifting	Inhibition	Working memory
Λ_1: Posner	✓	✓	*control*		
Λ_2: Posner + Working Memory	✓	✓	*control*		✓
Λ_3: Mirsky	✓	✓	✓		✓
Λ_4: Composite 1	✓	✓	✓	✓	
Λ_5: Composite 2	✓	✓	✓	✓	✓
Λ_6: Unitary	*attention*				
Λ_7: Unitary + Working Memory	*attention*				✓

In the table above, each row describes a theory of attention ability, as quantified in a loadings matrix (see Appendix C). Each column represents a latent attention ability, or factor. A checkmark (or italicized factor name) indicates that a factor is included the corresponding loadings matrix.

For our analysis of the attention dataset, we incorporated each loadings matrix in turn in the psychometric component of the LATER model version of the CLVM and the diffusion model version of the CLVM. This procedure generated two groups of seven models each.

Model comparison. To observe which of these theoretical accounts of latent attention abilities provided the best description of our attention dataset, we compared fit across all CLVMs that incorporated the LATER model and across all CLVMs that incorporated the diffusion model. We opted to use the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002, 2014) as our method of model comparison as it was designed to compare hierarchical Bayesian models specifically and is easily calculated from

MCMC samples. Similar to other model comparison metrics used in Bayesian analysis, the DIC balances goodness of fit against model complexity, as the DIC penalizes models with a high effective number of model parameters. Although the raw DIC value is not easily interpreted, comparatively low DIC values (i.e., at least 3–7 units lower; Spiegelhalter et al., 2002) indicate notably better fit. Other options for model comparison metrics include Bayes factors, AIC, and BIC (for a brief review, see Vandekerckhove, Matzke, & Wagenmakers, 2015).

In order to calculate the DIC for each model, we included `'dic'` as a module in the Trinity call for each model (see Matzke et al., 2018, for examples). This inclusion triggered JAGS to monitor the deviance in addition to the monitored model parameters. Chains for the deviance variable were submitted to `getdic`, a built-in Trinity function (Vandekerckhove, 2015). This function calculates DIC across the entire range of kept samples, across the first and last 50% of samples, and across each quarter of the samples. The calculation across subdivisions of samples facilitates observance of DIC stability, which is important to track as DIC convergence is notoriously more difficult to establish than chain convergence.

As such, we required that a high level of convergence was attained across all parameters for each model. We operationalized "a high level of convergence" as (1) a Gelman-Rubin diagnostic statistic ($\hat{R}$; Gelman & Rubin, 1992) under 1.01 for all model parameters, and (2) a DIC value that appears stable across quartiles of samples.

Results

The model comparison results are presented in Table 3.2. For ease of interpretation, rather than presenting raw DIC values, we have presented differences in DIC from the best-fitting model within each set. Thus, ΔDIC was calculated by subtracting the lowest DIC from the DIC value for each model.

To assess the comparative merit of the seven theoretical accounts of attention ability, we can observe which loadings matrix enabled the CLVMs to best to fit the observed data.

Table 3.2
Model comparison results

Cognitive component	*Psychometric component*	Δ *DIC*
LATER model	Λ_1: Posner	2.69×10^2
	Λ_2: Posner + WM	2.06×10^2
	Λ_3: Mirsky	1.99×10^2
	Λ_4: Composite 1	3.47×10^2
	Λ_5: Composite 2	2.90×10^2
	Λ_6: Unitary	0.87×10^2
	Λ_7: Unitary + WM	**0**
Diffusion model	Λ_1: Posner	3.44×10^3
	Λ_2: Posner + WM	2.93×10^3
	Λ_3: Mirsky	2.38×10^3
	Λ_4: Composite 1	4.28×10^3
	Λ_5: Composite 2	4.52×10^3
	Λ_6: Unitary	0.56×10^3
	Λ_7: Unitary + WM	**0**

Overall, the model with the lowest DIC value, and therefore the theoretical account that provided the best description of the observed data, was Λ_7, the *Unitary + Working Memory* account. It is reassuring that, qualitatively, we can draw the similar general conclusions from the set of CLVMs incorporating the LATER model and the set of CLVMs incorporating the diffusion model. Specifically, we observe a similar same rank order of models by DIC value when only the response time data is used (in the LATER model versions) and when both the response time data and accuracy data are used (in the more nuanced diffusion model versions). This suggests our conclusions about the relative suitability of the seven theories of attention are robust to the likelihood.

One could question whether the superior fit of the *Unitary + Working Memory* model might be accounted for by the addition of the working memory component. It is clear that the *Unitary + Working Memory* model fit better than the *Unitary* model, due to the inclusion of the working memory factor in the former account. However, although we believe that this is a contributing factor, the beneficial addition of a memory factor cannot be the sole reason for Λ_7's superior performance. Multiple other accounts (Λ_2, Λ_3, and Λ_5) also include a working memory factor, but the models incorporating these loadings matrices do not consistently have lower DIC values than those models that do not include a working memory factor (Λ_1, Λ_4, and Λ_6).

Our results suggest that the lack of division of attention into component abilities is the main reason that the *Unitary + Working Memory* model outperformed the other models in each set. The two models that proposed a single factor to capture unobserved attention abilities (Λ_6, the *Unitary account*, and Λ_7, the *Unitary + Working Memory* account) bested all models that divided attention into component abilities by a comfortable margin. Of those models that propose separable components of attention ability, the models that proposed the highest number of divisions of attention, Λ_4 and Λ_5, the two *Composite* accounts, consistently yielded the worst fit. Therefore, we believe that the division of attention into component abilities is what accounts for the most of the comparatively poor fit of all models as compared to the *Unitary + Working Memory* model.

It certainly still may be that there are independent components of attention ability; our results simply indicate that a unified theory of attention is the best description of this particular dataset out of the particular set of models we considered. To validate these results, we would suggest this study be replicated with a sample of participants that would be expected to have a more diverse range of attention abilites than our sample of college students. However, this conclusion is supported by previous research, mentioned at the beginning of this section, that reports significant and replicable interactions among the attention subsystems proposed by Posner and Petersen (1990). In light of our results, it appears possible that these previous findings were observed because attention ability is not composed of separable abilities. Our findings suggest that attention, similar to intelligence (g), is best described as a single latent ability that affects performance across superficially diverse tasks. We contend that attention ability is truly distinct only from other executive functions, such as working memory. Therefore, viewing attention as one of a small collection of executive functions may be the most viable framework to guide future research on attention.

III. Conclusion

Through the use of CLVMs, we were able to take a more direct approach to comparing theories of attention than would be possible with either a cognitive model alone or a psychometric model alone. Only a CLVM approach allows the researcher to infer latent abilities from data from all tasks while also accounting for the complexity of the psychological processes that produced the data. In general, the CLVM technique expands the scope of research questions that may be asked by making the unique psychological statement that observed task performance is due to unobserved task-specific ability, which in turn is a result of underlying general cognitive abilities. This is because CLVMs takes a more statistically nuanced approach by removing sources of variability that are not cognitive in nature before inferring large-scale latent abilities. As such, this approach also makes the unique statistical statement that the latent variables uncovered by the psychometric model determine a parameter of the cognitive model, which in turn generates the observed data.

Because the CLVM approach is a Bayesian hierarchical modeling technique, it not only inherits the benefits of Bayesian modeling generally, such as the propagation of uncertainty through all levels of a model, but also gains the subtlety of hierarchical modeling, allowing one model to account for the structure of a parameter of multiple other models. In this way, the CLVM approach implements the core idea of Batchelder's cognitive psychometrics technique in a more statistically powerful way.

In this tutorial, we sought to introduce a general implementation procedure for CLVMs in JAGS that should be accessible to any researcher who has becomes familiar with Bayesian sampling software. As demonstrated in our case example, switching out the cognitive component, from the simpler LATER model to the more complex diffusion model, was a straightforward process after the initial CLVM was constructed. To construct a new CLVM, both cognitive model components and psychometric components may be flexibly interchanged. Other cognitive models that might be used in place of the options discussed here include models such as the linear ballistic accumulator model (Brown & Heathcote, 2008) which may be used to describe multi-alternative choice response times, or multinomial processing tree models (Batchelder & Riefer, 1999) which capture elaborate decision-making processes underlying categorical choice data. Other latent variable models that might be used in place of factor analysis include regression, structural equation models, and nonlinear combinations of latent variables. A few studies including CLVMs with different components already exist. For example, Guan, Lee, and Vandekerckhove (2015) combined a cognitive model of optimal stopping behavior with a regression model to infer latent risk-taking and bias, and Nunez, Srinivasan, and Vandekerckhove (2015) combined a diffusion model with a regression model that incorporates neural data.

In this article, we hope to have demonstrated the unique capabilities of CLVMs through our case example of a comparative evaluation of theories of attention. Through the inclusion of one or more cognitive components, CLVMs can be tailored to the specific research context and account for the latent generating processes of a wide variety of data types. Through the inclusion of the psychometric component, CLVMs are able to assess common influences across a heterogeneous dataset in a single model. To create a new CLVM, model components may be easily switched out by changing just a few lines in JAGS. It is our hope that this tutorial will enable researchers to build their own cognitive latent variable models, and use this flexible and powerful new approach to data analysis to answer a host of novel questions about the structure of cognitive constructs underlying diverse data.

References

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*(4), 331.

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 71–93). Washington, DC: American Psychological Association.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*(1), 57–86.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual review of psychology*, *53*(1), 605–634.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *3*(57), 153–178.

Callejas, A., Lupianez, J., Funes, M. J., & Tudela, P. (2005). Modulations among the alerting, orienting and executive control networks. *Experimental Brain Research*, *167*(1), 27–37.

Carpenter, R. H. S. (1981). Oculomotor procrastination. In D. F. Fisher, R. A. Monty, & S. J. W (Eds.), *Eye movements: Cognition and visual perception* (pp. 237–246). Erlbaum.

Cheyne, J. A., Carriere, J. S., & Smilek, D. (2006). Absent-mindedness: Lapses of conscious awareness and everyday cognitive failures. *Consciousness and Cognition*, *3*(15), 578–592.

Cimprich, B., Visovatti, M., & Ronis, D. L. (2011). The Attentional Function Index – a self-report cognitive measure. *Psycho-Oncology*, *2*(20), 194–202.

Conners, C., & Staff, M. (2000). Conners' continuous performance test ii. *Multi-Health Systems Inc., North Tonawanda, NY*.

Fan, J., Gu, X., Guise, K. G., Liu, X., Fossella, J., Wang, H., & Posner, M. I. (2009).

Testing the behavioral interaction and integration of attentional networks. *Brain and cognition*, *70*(2), 209–220.

Fan, J., McCandliss, B. D., Fosella, J., Flombaum, J., & Posner, M. I. (2005). The activation of attentional networks. *Neuroimage*, *2*(26), 471–479.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *3*(14), 340–347.

Fernandez-Duque, D., & Posner, M. I. (2001). Brain imaging of attentional networks in normal and pathological states. *Journal of Clinical and Experimental Neuropsychology*, *23*(1), 74–93.

Fuentes, L. J., & Campoy, G. (2008). The time course of alerting effect over orienting in the attention network test. *Experimental Brain Research*, *185*(4), 667–672.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *4*(7), 457–472.

Guan, M., Lee, M. D., & Vandekerckhove, J. (2015). A hierarchical cognitive threshold model of human decision making on different length optimal stopping problems. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Lezak, M., Howieson, D., & Loring, D. (2004). *Neuropsychological assessment* (4th ed.). Oxford University Press.

Matzke, D., Boehm, U., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part III: Parameter estimation in nonstandard models. *Psychonomic bulletin & review*, *25*(1), 77–101.

McConnell, M. M., & Shore, D. I. (2011). Mixing measures: testing an assumption of the attention network test. *Attention, Perception, & Psychophysics*, *73*(4), 1096–1107.

Mirsky, A. F., Anthony, B. J., Duncan, C. C., Ahearn, M. B., & Kellam, S. G. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology* *review*, *2*(2), 109–145.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex frontal lobe tasks: A latent variable analysis. *Cognitive Psychology*, *1*(41), 49-100.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383.

Nunez, M. D., Srinivasan, R., & Vandekerckhove, J. (2015). Individual differences in attention influence perceptual decision making. *Frontiers in Psychology*(8), 1–13.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 221–247.

Pe, M., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion*, *13*(4), 739.

Peirce, J. W. (2007). PsychoPy — psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual review of neuroscience*, *35*, 73.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125).

Posner, M. I., & Fan, J. (2004). Attention as an organ system. In J. R. Pomerantz (Ed.), *Topics in Integrative Neuroscience: From Cells to Cognition* (p. 266-290). Oxford: Oxford University Press.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *1*(13), 25–42.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Raz, A., & Buhle, J. (2006). Typologies of attentional networks. *Nature Reviews Neuroscience*, *7*(5), 367–379.

Riccio, C. A., Reynolds, C. R., Lowe, P., & Moore, J. J. (2002). The continuous performance test: A window on the neural substrates for attention? *Archives of clinical neuropsychology*, *17*(3), 235–272.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*(2), 184.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207.

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology*, *5*(20), 343–350.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *4*(64), 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 485–493.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *3*(25), 251–260.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

Vandekerckhove, J. (2015). *Trinity: A matlab interface for bayesian analysis.* Retrieved from http://tinyurl.com/matlab-trinity.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In (p. 300). Oxford University Press, USA.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the ratcliff diffusion model to experimental data. *Psychonomic bulletin & review*, *14*(6), 1011–1026.

Wabersich, D., & Vandekerckhove, J. (2014). Extending jags: A tutorial on adding custom distributions to jags (with a diffusion model example). *Behavior research methods*, *46*(1), 15–28.

Wagenmakers, E. J. (2009). Methodological and empirical developments for the ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*(5), 641–671.

Wang, K., Fan, J., Dong, Y., Wang, C.-q., Lee, T. M., & Posner, M. I. (2005). Selective impairment of attentional networks of orienting and executive control in schizophrenia. *Schizophrenia research*, *78*(2), 235–241.

Weinbach, N., & Henik, A. (2012). The relationship between alertness and executive control. *Journal of experimental psychology: human perception and performance*, *38*(6), 1530.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85.

Yin, X., Zhao, L., Xu, J., Evans, A. C., Fan, L., Ge, H., … Liu, S. (2012). Anatomical substrates of the alerting, orienting and executive control components of attention: focus on the posterior parietal lobe. *PLoS One*, *7*(11), e50590.

Appendix A

Manipulation check results

Eight secondary confirmatory analyses were performed to check expected effects (based on the corresponding source material) for each of the five response time tasks.

For each comparison below, the first condition listed (e.g, the no cue condition) was expected to be more difficult than the second condition listed (e.g, the center cue condition). As such, we expected to observe that response times would be higher in the first condition than in the second condition. We report standardized effect sizes ($\delta = \frac{\mu_1 - \mu_2}{\sigma}$), which were expected to be positive, and Bayes factors (BF_{10}), which were expected to be > 1. Bayes factors indicate the strength of evidence for (or against) the hypothesis that a difference in means between the two conditions is observed ($\mathcal{H}_1 : \delta \neq 0$ vs. $\mathcal{H}_0 : \delta = 0$ for all comparisons). Analysis code for the Bayes factors and a note on our method of calculation is available at `osf.io/u7h27/`.

Table A1

Descriptive statistics and Bayes factors

		Accuracy	*Response time*		
	Conditions	Means	Means	δ	BF_{10}
ANT, Alerting effect	No cue	.973 (.002)	669.80 (2.67)	.24	3.84×10^4 ✓
	Center cue	.968 (.003)	627.91 (2.44)		
ANT, Orienting effect	No cue	.973 (.002)	669.80 (2.67)	.60	3.91×10^4 ✓
	Spatial cue	.986 (.002)	561.46 (2.66)		
ANT, Control effect	Congruent flanker	.992 (.001)	658.82 (2.18)	.44	3.20×10^4 ✓
	Incongruent flanker	.959 (.002)	580.97 (2.08)		
CPT-X, Vigilance effect	Target	.723 (.009)	443.03 (2.75)	.55	4.04×10^4 ✓
	Foil	.995 (.001)	378.30 (0.89)		
CPT-AX, Vigilance effect	Target	.880 (.008)	381.96 (2.37)	.32	2.78×10^4 ✓
	Foil	.985 (.001)	343.82 (0.94)		
NL, Switching effect	Switch	.846 (.008)	1070.10 (6.44)	.73	5.21×10^4 ✓
	No switch	.927 (.006)	849.90 (6.30)		
LG, Global precedence effect	Local	.927 (.004)	587.02 (2.45)	.59	1.92×10^4 ✓
	Global	.948 (.004)	493.04 (2.69)		
LG, Consistency effect	Conflicting	.904 (.005)	559.18 (2.86)	.22	3.67×10^4 ✓
	Consistent	.971 (.003)	523.13 (2.44)		

Mean accuracy is reported as a proportion; mean response time is reported in ms. For each condition mean, standard error of the mean is in parentheses. $BF_{10} \geq 10$ provide support for the alternative hypothesis; $\frac{1}{10} < BF_{10} < 10$ are inconclusive; $BF_{10} \leq \frac{1}{10}$ are provide support for the null hypothesis. Results that are consistent with the expected results (i.e., $\delta > 0$ and $BF_{10} \geq 10$) are noted with a ✓.

Appendix B

Model specifications

```
#data: response time tasks
for(n in 1:3){
   for(p in 1:P){
      for(i in 1:I){
         x[n,p,i] ~ dnorm(mu[n,p]/theta[p],theta2[p])
      }
   }
}

#priors: response time tasks
for(p in 1:P){
   theta2[p] <- pow[theta[p],2]
   theta[p] ~ dgamma(2,2)
}

#data: survey tasks
for(n in 1:2){
   for(p in 1:P){
      for(i in 1:I){
         y[n,p,i] ~ dnorm(mu[n,p],inveta2[n])
      }
   }
}

#priors: survey tasks
for(p in 1:P){
   inveta2[p] <- pow(eta[p],-2)
   eta[p] ~ dgamma(1,2)
}

#factor analysis, true scores
for(n in 1:N) {
  for(p in 1:P) {
    mu[n,p] <- iota[n] + inprod(lambda[n,],phi[,p])
  }
}


#priors: factor analysis
for(n in 1:N){
   iota[n] ~ dnorm(0,1)
}
lambda[1,1] <- 1
lambda[1,2] <- 0
lambda[2,1] <- 1
lambda[2,2] ~ dnorm(0,1)
lambda[3,1] ~ dnorm(0,1)
lambda[3,2] ~ dnorm(0,1)
lambda[4,1] ~ dnorm(0,1)
lambda[4,2] ~ dnorm(0,1)
lambda[5,1] ~ dnorm(0,1)
lambda[5,2] ~ dnorm(0,1)
for(f in 1:F){
   for(p in 1:P){
      phi[f,p] ~ dnorm(0,1)
   }
}
```

Figure B1. *Model specification for the LATER model version of the CLVM.*

```
#RT data: diffusion model
for(n in 1:3){
   for(p in 1:P){
      for(i in 1:I){
         x[n,p,i] ~ dwiener(alpha[p],tau[p],0.5,mu[n,p])
      }
   }
}

#priors: diffusion model
for(p in 1:P){
   alpha[p] ~ dunif(.01,4)
   ter[p] ~ dunif(.01,1)
}

#survey data: normal distribution
for(n in 1:2){
   for(p in 1:P){
      for(i in 1:I){
         y[n,p,i] ~ dnorm(mu[n,p],inveta2[p])
      }
   }
}

#priors: normal distribution
for(n in 1:2) {
  inveta2[n] <- pow(eta[n],-2)
  eta[n] ~ dunif(0,10)
}

#factor analysis, true scores
for(n in 1:N) {
  for(p in 1:P) {
    mu[n,p] <- iota[n] + inprod(lambda[n,],phi[,p])
  }
}
```

```

#priors: factor analysis
for(n in 1:N){
   iota[n] ~ dnorm(0,1)
}
lambda[1,1] <- 1
lambda[1,2] <- 0
lambda[2,1] <- 1
lambda[2,2] ~ dnorm(0,1)
lambda[3,1] ~ dnorm(0,1)
lambda[3,2] ~ dnorm(0,1)
lambda[4,1] ~ dnorm(0,1)
lambda[4,2] ~ dnorm(0,1)
lambda[5,1] ~ dnorm(0,1)
lambda[5,2] ~ dnorm(0,1)
for(f in 1:F){
   for(p in 1:P){
      phi[f,p] ~ dnorm(0,1)
   }
}
```

Figure B2. *Model specification for the diffusion model version of the CLVM.*

Appendix C

Loadings matrices

Each loadings matrix descried here represents a theoretical account of the latent structure of attention. In each matrix (Λ_i), rows correspond to observed variables (i.e., tasks), while columns correspond to the proposed factors (i.e., latent constructs). A verbal theory of attention is quantified in each matrix by fixing a subset of entries. This allows us to express *a priori* beliefs about the relationships between tasks and factors and effectively lends psychological interpretation to the factors.

Entries fixed to 0 imply that there is no relationship between the corresponding task and latent construct. As was mentioned in an earlier footnote, entries fixed to 1 do *not* imply that the corresponding task is a perfect measure of a given latent construct, as would be so in a classical implementation of confirmatory factor analysis; in the Bayesian implementation of confirmatory factor analysis, fixing an entry to 1 makes the weaker statement that there is a positive relationship of unknown relative strength between the corresponding task and factor.

Each loadings matrix is preceded by a brief description of the proposed latent cognitive abilities as defined by these fixed entries. For visual clarity, subscripts for all distinct freely estimated loadings parameters have been omitted (e.g., $\lambda_{3,5}$ and $\lambda_{3,6}$ are both represented as λ). In addition, factor names are printed above the corresponding column and task names are printed to the right of the corresponding row for every proposed loadings matrix. Finally, horizontal lines are overlaid on each matrix to make the division between tasks representing trial types in a single test more readily apparent.

Λ_1: Posner

Our first loadings matrix is inspired by inspired by Posner and Petersen's highly influential theory of attention (Posner & Petersen, 1990; Petersen & Posner, 2012). As mentioned earlier, they propose that the psychological construct of "attention" comprises three anatomically and functionally separable abilities: *altering*, the ability to sustain attention over an extended period, *orienting*, the ability to select and attend only to a small number of stimuli, and *control*, the ability to shift and change attention allocation to meet task demands. In order to express these abilities in a loadings matrix (Λ_1, printed below), we fixed a subset of matrix entries to either 0 or 1 based on prior knowledge of the abilities that each task was designed to measure.

The ANT was explicitly designed to assess altering, orienting, and control abilities separately (Fan et al., 2002, 2005). In this test, differences in mean response times between the "no cue" condition and the "center cue" condition are interpreted as a measure of alerting ability (Fan et al., 2005). When no cue is presented, participants must maintain a high level of attentional vigilance in order to be respond to the target within the time limit, and thus alerting ability is required on these trials. When a cue of any type is presented, participants know that the target will appear soon, thus circumventing the need for alerting ability on those trials. Therefore, for tasks where no cue is presented, the loading on the alerting factor is permitted to be nonzero, however, for tasks where a cue of any type is presented, the loading on the alerting factor is set to 0.

Differences in mean response times between the "center cue" condition and the "spatial cue" condition are interpreted as a measure of orienting ability (Fan et al., 2005). When a central cue or no cue is presented, participants must be able to quickly find and focus on the target in order to respond within the time limit, and thus orienting ability is required on those trials. When a spatial cue is presented, participants already know what position to attend to, thus circumventing the need for orienting ability. Therefore, for tasks where a central cue or no cue is presented, the loading on the orienting factor is permitted to be nonzero, however, for tasks where a spatial cue is presented, the loading on the orienting factor is set to 0.

Differences in mean response times between the “incongruent flanker” condition and the “congruent flanker” condition are interpreted as a measure of control ability (Fan et al., 2005). When incongruent flankers appear, participants must be able to effectively inhibit a response to the four flankers, which point in the opposite direction to the single target arrow, in order to respond accurately to the target within the time limit. Thus, control ability is required on the incongruent flanker trials. When a congruent flanker is presented, the information conveyed by the flankers need not be inhibited as it is consistent with the target, and so control ability is not required. Therefore, for tasks where incongruent flankers appear, the loading on the control factor is permitted to be nonzero, however, for tasks where congruent flankers appear, the loading on the control factor is set to 0.

Various versions of the CPT are widely used as measures of sustained attention ability (see Riccio, Reynolds, Lowe, & Moore, 2002, for a review). As sustained attention is very similar to Posner and Petersen’s description of alerting, all tasks derived from each version of the CPT (CPT-X and CPT-AX) should load on the alerting factor. We do not expect any condition of this task to require orienting ability, so all loadings on the orienting factor were set to 0. However, the target trials in each version of the CPT may rely on control abilities. Because target trials are rare, participants must to inhibit their prepotent response (i.e., the response to the foil stimuli) in order to respond accurately to a target within the time limit. Therefore, the tasks including targets trials in each version of the CPT are permitted to load on the control factor. For model identification purposes, we have set the loading for the foil tasks in the CPT-AX on the alerting factor to 1 ($\lambda_{10,3} = 1$).

The NL task was developed to assess the cost of switching attention (Rogers & Monsell, 1995). On each trial, participants must selectively attend to either the number portion or the letter portion of the stimulus based on the position of the stimulus. Therefore, regardless of condition, this task should relate to orienting ability. Switching cost is assessed using the difference in performance on “switch” trials, where participants must attend to a different portion of the stimulus than they did on previous last trial, and “no-switch” trials, where participants must attend to the same portion of the stimulus as on the previous trial. Because an effortful switch in the focus of attention is required for successful performance on switch trials, this task condition is permitted to load on on the control factor. For model identification purposes, we have set the loading for the no-switch task on the orienting factor to 1 ($\lambda_{11,2} = 1$).

The LG task is used to demonstrate the global precedence effect (Navon, 1977). In the “local” condition, participants must not allow the automatic global percept to interfere with their response. Therefore, local task conditions are permitted to load on the control factor to reflect this effortful reallocation of attention to the local feature over the global feature. In the “conflicting” conditions, participants must ensure their attention is oriented to the correct feature (either local or global) in order to respond correctly. Therefore, conflicting conditions are permitted to load on the orienting factor. For model identification purposes, we have set the loading for the local judgment and consistent stimulus task on the control factor to 1 ($\lambda_{14,3} = 1$).

Finally, because the AFI conceptualizes attention as a general ability, this task is permitted to load on all three latent attention abilities. In contrast, because the ARCES was designed to measure the frequency of lapses in sustained attention specifically, this task is only permitted to load on the alerting factor.

$$\Lambda_1 = \begin{array}{cccl}
\text{alerting} & \text{orienting} & \text{control} & \\
\lambda & \lambda & 0 & \textit{ANT: no cue, congruent} \\
0 & \lambda & 0 & \textit{ANT: center cue, congruent} \\
0 & 0 & 0 & \textit{ANT: spatial cue, congruent} \\
\lambda & \lambda & \lambda & \textit{ANT: no cue, incongruent} \\
0 & \lambda & \lambda & \textit{ANT: center cue, incongruent} \\
0 & 0 & \lambda & \textit{ANT: spatial cue, incongruent} \\
\cline{1-3}
\lambda & 0 & \lambda & \textit{CPT-X: targets} \\
\lambda & 0 & 0 & \textit{CPT-X: foils} \\
\cline{1-3}
\lambda & 0 & \lambda & \textit{CPT-AX: targets} \\
1 & 0 & 0 & \textit{CPT-AX: foils} \\
\cline{1-3}
0 & 1 & 0 & \textit{NL: no switch} \\
0 & \lambda & \lambda & \textit{NL: switch} \\
\cline{1-3}
0 & 0 & 0 & \textit{LG: global, consistent} \\
0 & 0 & 1 & \textit{LG: local, consistent} \\
0 & \lambda & 0 & \textit{LG: global, conflicting} \\
0 & \lambda & \lambda & \textit{LG: local, conflicting} \\
\cline{1-3}
\lambda & \lambda & \lambda & \textit{AFI} \\
\cline{1-3}
\lambda & 0 & 0 & \textit{ARCES}
\end{array}$$

Λ_2: Posner + Working Memory

This loadings matrix extends the *Posner* loadings matrix (Λ_1) through the addition of a fourth factor. We conjecture that attention may be divided into the same three latent abilites as proposed in the Posner structure, however, we add a fourth factor to reflect the known working memory load in two of the 16 tasks.

In the CPT-X and CPT-AX, participants' task is to judge whether the current stimulus is a target or a foil based on whether a rule is obeyed. In the CPT-AX, the rule not only relies on the current stimulus' identity, but also relies on the identity of the previous stimulus. Thus, in order to make the correct response on both trial types, participants must accurately recall the stimulus from the last trial. In the simpler CPT-X — and in all other tasks for that matter — there is no such memory demand. As such, the working memory factor is defined by nonzero loadings for the two memory-dependent CPT-AX tasks, and by loadings fixed to 0 for all other tasks. For model identification purposes, the first nonzero loading is fixed to 1 ($\lambda_{9,4} = 1$).

In the loading matrix printed below (Λ_2), the first three columns, corresponding to alerting, orienting, and control abilities, respectively, are identical to the three columns in the Λ_1 loadings structure. The fourth column of Λ_2 represents the added working memory factor.

$$\Lambda_2 = \begin{array}{|cccc|l}
\text{alerting} & \text{orienting} & \text{control} & \text{working memory} & \\
\lambda & \lambda & 0 & 0 & \textit{ANT: no cue, congruent} \\
0 & \lambda & 0 & 0 & \textit{ANT: center cue, congruent} \\
0 & 0 & 0 & 0 & \textit{ANT: spatial cue, congruent} \\
\lambda & \lambda & \lambda & 0 & \textit{ANT: no cue, incongruent} \\
0 & \lambda & \lambda & 0 & \textit{ANT: center cue, incongruent} \\
0 & 0 & \lambda & 0 & \textit{ANT: spatial cue, incongruent} \\
\hline
\lambda & 0 & \lambda & 0 & \textit{CPT-X: target} \\
\lambda & 0 & 0 & 0 & \textit{CPT-X: foil} \\
\hline
\lambda & 0 & \lambda & 1 & \textit{CPT-AX: target} \\
1 & 0 & 0 & \lambda & \textit{CPT-AX: foil} \\
\hline
0 & 1 & 0 & 0 & \textit{NL: no switch} \\
0 & \lambda & \lambda & 0 & \textit{NL: switch} \\
\hline
0 & 0 & 0 & 0 & \textit{LG: global, consistent} \\
0 & 0 & 1 & 0 & \textit{LG: local, consistent} \\
0 & \lambda & 0 & 0 & \textit{LG: global, conflicting} \\
0 & \lambda & \lambda & 0 & \textit{LG: local, conflicting} \\
\hline
\lambda & \lambda & \lambda & 0 & \textit{AFI} \\
\hline
\lambda & 0 & 0 & 0 & \textit{ARCES}
\end{array}$$

Λ_3: Mirsky

Our third loadings matrix is inspired by inspired an alternative theory of latent attention ability proposed by Mirsky and colleagues (1991). They contend that attention is composed of four independent abilities, including: *sustain*, the ability to maintain an attentional focus over time, *focus-execute*, the ability to select a specific task or stimulus to attend to, *shift*, the ability to shift or change the focus of attention flexibly, and *encode*, the ability to register, retain, and possibly manipulate information.

The former two abilities, sustain and focus-execute, map fairly closely to Posner and Petersen's (1990) theorized alerting and orienting abilities, respectively. As such, we allow for the first two columns in Λ_3 below, the sustain and focus-execute factors, to be defined by duplicating the first and second columns of the *Posner* loadings matrix (Λ_1).

The third ability, shift, is a new conception of the executive aspect of attention which focuses exclusively on switching attention, in contrast to Posner and Petersen's (1990) control ability, which is more broadly defined. Conditions of the CPT-X, CPT-AX, and NL tasks that previously loaded on the control factor are now permitted to load on the shifting factor. For target trials in each version of the CPT, participants must switch from one mode of responding (correct response behavior on foil trials) to another (correct response behavior on target trials). Therefore, these two tasks may load on the shift factor. The "switch" condition of the NL task is also permitted to load on the shift factor, as successful performance on this trial type requires a shift in one's mental set. For model identification purposes, this loading is fixed to 1 ($\lambda_{12,3} = 1$). Finally, the AFI may capture the shift ability, as it measures the ability to complete everyday tasks that draw on attention ability in a general sense. As such, this survey is allowed a nonzero loading. All other tasks do not involve a mental shift or purport to measure this ability, and therefore are not permitted to load on the shifting factor.

The fourth and final component of attention proposed by Mirsky is encode. The definition of this ability is strikingly similar to descriptions of working memory. As such, we allow the both conditions of the CPT-AX, the only response time tasks that entail a memory load, to have nonzero loadings for the encode factor. We also permit our general attention measure, the AFI, to load on this factor. All other tasks do not involve a memory load or do not measure activities affected by attention-related memory and therefore are not permitted to load on the shifting factor.

$$\Lambda_3 = \begin{array}{ccccl}
\text{sustain} & \text{focus-execute} & \text{shifting} & \text{encode} & \\
\lambda & \lambda & 0 & 0 & \textit{ANT: no cue, congruent} \\
0 & \lambda & 0 & 0 & \textit{ANT: center cue, congruent} \\
0 & 0 & 0 & 0 & \textit{ANT: spatial cue, congruent} \\
\lambda & \lambda & 0 & 0 & \textit{ANT: no cue, incongruent} \\
0 & \lambda & 0 & 0 & \textit{ANT: center cue, incongruent} \\
0 & 0 & 0 & 0 & \textit{ANT: spatial cue, incongruent} \\
\hline
\lambda & 0 & \lambda & 0 & \textit{CPT-X: target} \\
\lambda & 0 & 0 & 0 & \textit{CPT-X: foil} \\
\hline
\lambda & 0 & \lambda & 1 & \textit{CPT-AX: target} \\
1 & 0 & 0 & \lambda & \textit{CPT-AX: foil} \\
\hline
0 & 1 & 0 & 0 & \textit{NL: no switch} \\
0 & \lambda & 1 & 0 & \textit{NL: switch} \\
\hline
0 & 0 & 0 & 0 & \textit{LG: global, consistent} \\
0 & 0 & 0 & 0 & \textit{LG: local, consistent} \\
0 & \lambda & 0 & 0 & \textit{LG: global, conflicting} \\
0 & \lambda & 0 & 0 & \textit{LG: local, conflicting} \\
\hline
\lambda & \lambda & \lambda & \lambda & \textit{AFI} \\
\hline
\lambda & 0 & 0 & 0 & \textit{ARCES}
\end{array}$$

Λ_4: Composite 1

As mentioned in the description for the previous loadings matrix, there are some similarities between Posner and Petersen's (1990) and Mirsky and colleagues' (1991) theories of the structure of attention ability. Related research on the latent structure of executive functions by Miyake and colleagues' (2000) further overlaps with these theories. In our fourth and fifth loadings matrices, we propose novel theories of attention that draw from all three of these sources to create plausible theories of underlying attention abilities.

A major area of overlap in the previously discussed theories is seen in the repeated descriptions of *sustained* attention, referred to as "alerting" by Posner and Petersen and as "sustain" by Mirsky and colleagues, and *selective* attention, referred to as "orienting" by Posner and Petersen and as "focus-execute" by Mirsky and colleagues. In this loadings matrix, we again include these abilities. The first factor, sustained, duplicates previous definitions of alerting (the first column in Λ_1) and sustain (the first column in Λ_3). The second factor, *selective*, duplicates previous definitions of orienting (the second column in Λ_1) and focus-execute (the second column in Λ_3).

These theories diverge in their descriptions of executive attention abilities. Posner and Petersen's third attention component, control, is thought to comprise all executive attention functions, including switching the focus of attention, inhibiting prepotent responses, and other supervisory functions. In contrast, Mirsky and colleagues provide a more constrained view of executive attention in their description of the shift ability. Finally, Miyake and colleagues explicitly capture different elements of executive attention as separable components, specifically shifting and inhibition abilities. We include *shifting*, described by both Mirsky and Miyake, as the third factor in the present loadings matrix. We have previously described how shifting may be defined in our discussion of Mirsky and colleagues' theory of attention. Therefore, to incorporate this ability in the present loadings matrix, we duplicate the third column of the Mirsky matrix (Λ_3).

The final component of attention ability included in this matrix is *inhibition.* As in

Miyake and colleagues' description, we view inhibition as the ability to suppress a prepotent response. For ANT trial types with incongruent flankers, participants must ignore the conflicting direction information given by the flanking arrows in order correctly respond to the center target arrow. Thus, successful performance on these trial types requires the inhibition of a response to the more numerous flankers. In the CPT-X and CPT-AX, a successful response on target trials, which are rare, requires that the prepotent response to foil trials, which are common, is inhibited (in addition to requiring a successful mental set update). In local conditions of the LG task, participants must inhibit a tendency to respond in a way that is influenced by the global precedence effect. For model identification purposes, we have set the loading for the local judgment and consistent stimulus task on the inhibition factor to 1 ($\lambda_{14,4} = 1$). The NL task is exclusively a switching attention task, and therefore is not permitted to load on the inhibition factor.

Because the AFI was designed to assess everyday functioning relying on attention ability in a general sense, it is permitted to load on the inhibition factor. Because the ARCES was designed to assess lapses in sustained attention ability only, it assumed to have no relationship to inhibition abilities ($\lambda_{18,4} = 0$).

One may also see this loadings matrix as a reconceptualization of Posner and Petersen's theory, where the control ability (column 3 in Λ_1) has been "split" to create two independent abilities, shifting and inhibition.

$$\Lambda_4 = \begin{array}{cccc|l}
\text{sustained} & \text{selective} & \text{shifting} & \text{inhibition} & \\
\lambda & \lambda & 0 & 0 & \textit{ANT: no cue, congruent} \\
0 & \lambda & 0 & 0 & \textit{ANT: center cue, congruent} \\
0 & 0 & 0 & 0 & \textit{ANT: spatial cue, congruent} \\
\lambda & \lambda & 0 & \lambda & \textit{ANT: no cue, incongruent} \\
0 & \lambda & 0 & \lambda & \textit{ANT: center cue, incongruent} \\
0 & 0 & 0 & \lambda & \textit{ANT: spatial cue, incongruent} \\
\hline
\lambda & 0 & \lambda & \lambda & \textit{CPT-X: target} \\
\lambda & 0 & 0 & 0 & \textit{CPT-X: foil} \\
\hline
\lambda & 0 & \lambda & \lambda & \textit{CPT-AX: target} \\
1 & 0 & 0 & 0 & \textit{CPT-AX: foil} \\
\hline
0 & 1 & 0 & 0 & \textit{NL: no switch} \\
0 & \lambda & 1 & 0 & \textit{NL: switch} \\
\hline
0 & 0 & 0 & 0 & \textit{LG: global, consistent} \\
0 & 0 & 0 & 1 & \textit{LG: local, consistent} \\
0 & \lambda & 0 & 0 & \textit{LG: global, conflicting} \\
0 & \lambda & 0 & \lambda & \textit{LG: local, conflicting} \\
\hline
\lambda & \lambda & \lambda & \lambda & \textit{AFI} \\
\hline
\lambda & 0 & 0 & 0 & \textit{ARCES}
\end{array}$$

Λ_5: Composite 2

This loadings matrix extends the *Composite 1* loadings matrix (Λ_4) through the addition of a fifth factor. Both Mirsky (1991) and Miyake (2000) discuss the ability to hold and manipulate information as a possible component of attention and executive function abilities. In addition to the four factors proposed in the Composite 1 structure, we include a factor to express working memory as an aspect of attention, which is defined by duplicating the *encode* factor from the *Mirsky* loadings matrix(Λ_3).

$$\Lambda_5 = \begin{array}{ccccc|l}
\text{sustained} & \text{selective} & \text{shifting} & \text{inhibition} & \text{encode} & \\
\cline{1-5}
\lambda & \lambda & 0 & 0 & 0 & \textit{ANT: no cue, congruent} \\
0 & \lambda & 0 & 0 & 0 & \textit{ANT: center cue, congruent} \\
0 & 0 & 0 & 0 & 0 & \textit{ANT: spatial cue, congruent} \\
\lambda & \lambda & 0 & \lambda & 0 & \textit{ANT: no cue, incongruent} \\
0 & \lambda & 0 & \lambda & 0 & \textit{ANT: center cue, incongruent} \\
0 & 0 & 0 & \lambda & 0 & \textit{ANT: spatial cue, incongruent} \\
\cline{1-5}
\lambda & 0 & \lambda & \lambda & 0 & \textit{CPT-X: target} \\
\lambda & 0 & 0 & 0 & 0 & \textit{CPT-X: foil} \\
\cline{1-5}
\lambda & 0 & \lambda & \lambda & 1 & \textit{CPT-AX: target} \\
1 & 0 & 0 & 0 & \lambda & \textit{CPT-AX: foil} \\
\cline{1-5}
0 & 1 & 0 & 0 & 0 & \textit{NL: no switch} \\
0 & \lambda & 1 & 0 & 0 & \textit{NL: switch} \\
\cline{1-5}
0 & 0 & 0 & 0 & 0 & \textit{LG: global, consistent} \\
0 & 0 & 0 & 1 & 0 & \textit{LG: local, consistent} \\
0 & \lambda & 0 & 0 & 0 & \textit{LG: global, conflicting} \\
0 & \lambda & 0 & \lambda & 0 & \textit{LG: local, conflicting} \\
\cline{1-5}
\lambda & \lambda & \lambda & \lambda & \lambda & \textit{AFI} \\
\cline{1-5}
\lambda & 0 & 0 & 0 & 0 & \textit{ARCES} \\
\cline{1-5}
\end{array}$$

Λ_6: Unitary

Unlike the previous loadings matrices, our penultimate proposed loadings matrix does not divide attention into component abilities. Instead, we propose that attention is best described as a single latent ability. Because all 18 tasks included in the battery may rely on attention, the column of the loadings matrix corresponding to the attention factor includes all nonzero loadings. Only one entry is fixed for model identification purposes ($\lambda_{1,1} = 1$).

$$\Lambda_6 = \begin{array}{c} \text{attention} \\ \left[\begin{array}{c} 1 \\ \lambda \\ \lambda \\ \lambda \\ \lambda \\ \lambda \\ \hline \lambda \\ \lambda \\ \hline \lambda \\ \lambda \\ \hline \lambda \\ \lambda \\ \hline \lambda \\ \lambda \\ \lambda \\ \lambda \\ \hline \lambda \\ \hline \lambda \end{array}\right] \end{array} \begin{array}{l} \\ \textit{ANT: no cue, congruent} \\ \textit{ANT: center cue, congruent} \\ \textit{ANT: spatial cue, congruent} \\ \textit{ANT: no cue, incongruent} \\ \textit{ANT: center cue, incongruent} \\ \textit{ANT: spatial cue, incongruent} \\ \textit{CPT-X: target} \\ \textit{CPT-X: foil} \\ \textit{CPT-AX: target} \\ \textit{CPT-AX: foil} \\ \textit{NL: no switch} \\ \textit{NL: switch} \\ \textit{LG: global, consistent} \\ \textit{LG: local, consistent} \\ \textit{LG: global, conflicting} \\ \textit{LG: local, conflicting} \\ \textit{AFI} \\ \textit{ARCES} \end{array}$$

Λ_7: Unitary + Working Memory

Our final loadings matrix is similar to the last in that we again contend that attention ability is best described as a unitary ability. However, in this theoretical account, we additionally consider the contribution of working memory. The first factor in the matrix below represents the unified attention factor, and is a duplicate of the the single column in the previous loadings matrix (Λ_6). The second factor represents the working memory factor, and is a duplicate of the working memory factor included in a previous loadings matrix (column 4 in Λ_2). Overall, this matrix quantifies the idea that task performance is best described by a small set of large-scale cognitive abilities.

$$\Lambda_7 = \begin{array}{cc|l}
\text{attention} & \text{working memory} & \\
\hline
1 & 0 & \textit{ANT: no cue, congruent} \\
\lambda & 0 & \textit{ANT: center cue, congruent} \\
\lambda & 0 & \textit{ANT: spatial cue, congruent} \\
\lambda & 0 & \textit{ANT: no cue, incongruent} \\
\lambda & 0 & \textit{ANT: center cue, incongruent} \\
\lambda & 0 & \textit{ANT: spatial cue, incongruent} \\
\hline
\lambda & 0 & \textit{CPT-X: target} \\
\lambda & 0 & \textit{CPT-X: foil} \\
\hline
\lambda & 1 & \textit{CPT-AX: target} \\
\lambda & \lambda & \textit{CPT-AX: foil} \\
\hline
\lambda & 0 & \textit{NL: no switch} \\
\lambda & 0 & \textit{NL: switch} \\
\hline
\lambda & 0 & \textit{LG: global, consistent} \\
\lambda & 0 & \textit{LG: local, consistent} \\
\lambda & 0 & \textit{LG: global, conflicting} \\
\lambda & 0 & \textit{LG: local, conflicting} \\
\hline
\lambda & 0 & \textit{AFI} \\
\hline
\lambda & 0 & \textit{ARCES}
\end{array}$$

CHAPTER 4: A HIERARCHICAL BAYESIAN APPROACH TO JOINT MODELING OF NEURAL AND BEHAVIORAL DATA

Abstract

The noisiness and complexity of behavioral data can make the creation of effective joint models of neural and behavioral data challenging. Cognitive models, especially those developed in mathematical psychology, can be incorporated to infer specific psychological variables that shape the observed behavioral data. However, previous work on joint models of neural and behavioral data that incorporates cognitive models has been primarily correlational in nature, and so is not suitable for theory testing. In this paper, we develop a new joint modeling framework called *neurocognitive process modeling*, in which a cognitive model of the behavioral data and a probabilistic model of the neural data are combined in a single hierarchical Bayesian model. We describe how this approach may be used to construct models that describe comprehensive latent processes, and how the models may be used for the comparative evaluation of complex hypotheses and theories about unobserved neurocognitive processes. In two case studies, we demonstrate how this approach extends the scope of questions that may be asked about the origin of behavior in neural signals.

Introduction

Different fields have developed a wide variety of approaches to inferring the neural processes and cortical systems that underlie behavior. In both computational and cognitive neuroscience, models may be used to express a unified quantitative account of behavioral and neural data, and therefore aid in generating new conclusions and insights about the underlying neurocognitive process. However, the significant noise and complexity inherent in behavioral data can make developing effective models challenging. In psychology, computational cognitive models are frequently used to quantify a theoretical underlying cognitive process. Cognitive models excel in accounting for how individual differences in psychological factors and features of the experimental context might systematically affect observed behavior. In recent decades, researchers have begun to link cognitive model accounts of behavioral data to patterns observed in neural data. To do so, there have been two primary approaches used: two-stage analyses and joint modeling.

Two-stage analysis

In a two-stage analysis, the goal is to assess the nature of the relationship between the neural data and the behavioral data, as seen through the lens of a cognitive model. First, the behavioral data is submitted to the cognitive model to derive estimates of various features of the proposed underlying cognitive process. Then, a summary statistic of the neural data, such as the mean activation across fMRI voxels in a given cortical area, are correlated to each of the cognitive model parameter point estimates. This process may be repeated in order to observe how changes in cognitive model parameters due to systematic alterations of elements of the behavioral task relate to fluctuations in the neural data.

However, this approach suffers from two shortcomings. First, it provides limited evidence about the nature of neurocognitive relationships. From this analysis, conclusions might be drawn about the relationship between neural activity and inferred features of a cognitive process. However, often an assumption is made that a significantly correlated cortical area is *the* neural substrate for the cognitive model parameter. This is not only faulty because correlations do not imply causation, but also because this method assumes that a summary statistic of the neural data is sufficient to capture the neural activity. However, just as using raw behavioral data ignores the complexity of the cognitive process, using raw neural data ignores complexity of the underlying neural process. Yet research in this vein (e.g., Forstmann et al., 2008; Van Veen, Krug, & Carter, 2008; Forstmann, Brown, Dutilh, Neumann, & Wagenmakers, 2010; van Maanen et al., 2011) often overstates the strength and nature of the evidence by making causal conclusions.

Second, even if researchers' conclusions are tempered, the method is statistically undesirable. In order to reach the second stage of the procedure, in which cognitive model parameters are correlated to the neural data, the uncertainty associated with the point estimates for each parameter must be discarded. By discarding this information, a researcher risks introducing bias, especially if the posterior distributions are notably skewed or extremely wide (Pagan, 1984). At a minimum, a two-stage approach leads to a loss of power, as many data points (i.e., response times on many trials) are collapsed to a single summary statistic (i.e., a cognitive parameter estimate).

Joint models

A joint modeling approach (e.g., in the style of Turner et al., 2013) resolves these issues by implementing both a neural model and a behavioral model in a single step as a hierarchical Bayesian model. A joint model is constructed by specifying a model of the neural data with a set of associated parameters:

$$x \sim \text{Neural}(\Delta)$$

a model of the behavioral data, such as a diffusion model (as in Turner, Van Maanen, & Forstmann, 2015) or a linear ballistic accumulator model (as in Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016), with a set of associated parameters:

$$y \sim \text{Behavioral}(\Theta)$$

and a stochastic linking function which may act as a joint distribution of the parameters of both model components:

$$(\Delta, \Theta) \sim \text{Joint}(\Omega)$$

This joint distribution may be any function that would allow for the parameters of the component models to be hierarchically linked, although in practice, only distributions that allow for the strength and the direction of the relationship between the each neural and behavioral model parameter have been used. For example, the joint distribution may be a multivariate normal distribution (as in Turner et al., 2013, 2016, 2015) or factor analysis (Turner, Wang, & Merkle, 2017). The approach is statistically sound, as it allows the uncertainty in the parameter estimates for each component to be propagated through the model.

Two analyses are typically performed with this style of model. First, the model may be used to make predictions about either the neural data or the behavioral data or both. Second, the parameters capturing the relationships between neural model parameters and behavioral model parameters (e.g., correlations, ρ) are used to make inferences about the structure of cognition. Both goals are accomplished by "treat[ing] the two sources of information as separate measurements of the same cognitive construct" (Turner et al., 2013, , p. 193).

This perspective suggests that the joint modeling approach is not intended to describe a latent process, but rather is designed to infer abstract relationships. As such, a joint modeling approach might be considered an approach to generating theories about latent processes. The traditional choices for the joint distribution makes it clear that the goal is ultimately to observe which, if any, features of the behavioral process are related to the neural data.

However, researchers may have strong hypotheses about the particular way in which the neural data determines behavior, and so may wish to more directly model and test the viability of a theorized latent process. As such, a correlational method may be restrictive, as it is agnostic about the source of the data. Furthermore, in practice, joint models do not account for the complexity of the neural data. Rather than specifying a true neural *model*, a normal distribution is almost exclusively used (Turner et al., 2015, 2016, 2017).

An alternative framework

In the novel approach proposed here, we do not intend to measure any singular abstract construct. Rather, we seek to establish a viable description of the underlying *process*. To do this, we include both a cognitive model of behavioral data capable of capturing individual differences in psychological factors as well as a nuanced model of the neural data that captures both domain knowledge about the area of cortex under study and theoretical ideas about how populations of neurons work together encode information. Similar to the previous joint modeling approach, the neural model and a cognitive model are both implemented in a hierarchical Bayesian framework. However, unlike the joint modeling approach, we employ a directed link between the model components in order to express a more theoretically-committed statement about the underlying neurocognitive process. The ultimate goal of this new approach, which we call *neurocognitive process modeling*, is to establish a comprehensive framework for describing comprehensive latent processes where models are easily altered and components are easily replaced, similar to the cognitive latent variable modeling framework established by Vandekerckhove (2014).

In this article, we describe how a pair of customized neurocognitive process models were constructed and used to analyze a published dataset (Li, Gerfen, & Svoboda, 2014), which includes electrophysiological recordings and binary behavioral responses. We place special emphasis on how we used two competing versions of the model to test a complex hypothesis about the underlying neural process. Specifically, the goals of the case study described here were to: (1) test which hypothesis about the neural process better describes the joint dataset, and (2) generate new insights about the underlying neurocognitive process through analysis of the posterior distributions of selected model parameters. Accomplishing the first objective will demonstrate that neurocognitive process model (NCPMs) are useful for hypothesis testing. Accomplishing the second objective will demonstrate how the hierarchical nature of the technique naturally leads to new conclusions about the nature of the latent neurocognitive process underlying the observed data.

Dataset

We analyzed a published, open dataset (Li et al., 2014) that included electrophysiological recordings from motor cortex and behavioral data from 99 sessions across 19 mice. In each trial, a trained mouse performed a simple spatial discrimination task while electrophysiology data was simultaneously recorded (see Figure 4.1). The goal of the task was to indicate with a lick whether a pole was presented in an anterior or posterior position relative to the right whisker pad. Each trial began when the pole descended, signaling the start of the sample period. The pole remained down for 1.3 seconds, during which time the mouse used its whiskers to detect the pole's position. When the pole was removed, the sample period ended and the 1.3-second delay period began. During the delay, the mouse attempted to retain the perceived position of the pole in memory until a brief auditory cue signaled the start of the response period. At any time after the cue, the mouse was permitted to lick one of two ports to indicate its response. Licking the left port indicated the pole was thought to have been in the anterior position, while licking the right port indicated the pole was thought to have been in the posterior position. In our analysis, the behavioral data, y, is the observed lick direction.

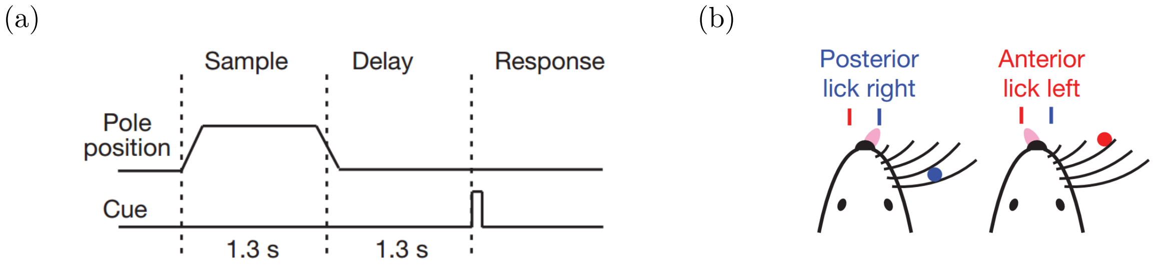


Figure 4.1. Spatial discrimination task. Mice were trained to discriminate between two bar positions. (a) The time course of each trial. (b) Mapping between stimulus bar position and correct response. Throughout this article, red will indicate a left response, and blue will indicate a right response. Reproduced from Li, Chen, Guo, Gerfen, & Svoboda, 2015.

While mice performed the behavioral task, extracellular recordings were taken from left anterior lateral motor cortex (ALM) using a 32-channel electrode. In each session, between

2 and 29 pyramidal units[1] (mean = 12.6) were identified. In our analysis, the neural data, x, are the binned spike counts (bin width = 100ms) for each 1.3-second trial epoch. Only units classified as pyramidal were included. Data from trials where the mouse licked before the cue, the mouse did not lick, or optogenetics were used to alter the neuronal response were excluded; no other exclusions of mice, sessions, units, or trials were made.

Models

In their 2015 *Nature* paper, the authors of this dataset describe their analysis of this data and their subsequent conclusions concerning the function and purpose of this section of motor cortex (Li et al., 2015). Specifically, Li and colleagues assert that a majority of pyramidal neurons in ALM encode the planned motor response in their average firing rate. As such, they contend that the ALM region in mice may be seen as a homologue of human premotor cortex.

We designed the neurocognitive process models described in this section not only to test and hopefully validate these conclusions, but furthermore to add to our understanding of the neurocognitive process by which ALM supports planned decision behaviors. The models were designed to be as simple as possible while still capturing key features of the experimental context and qualitative theories about the underlying neurocognitive process. These key features included: (1) Li and colleagues' (2015) finding that ALM neurons encode the planned lick direction in their firing rate, (2) the principle of population coding (i.e., pooling across neurons with different preferences), (3) the binary nature of the behavioral response, and (4) individual differences across mice. By accounting for these ideas directly in the models, each NCPM is able to make a comprehensive, strong theoretical statement

[1] When describing neural data collected through extracellular recordings, the term *unit* is generally preferred over *neuron*. This is because neurons' membrane potentials are not directly measured, but rather voltage fluctuations at various points in the extracellular space are recorded and subsequently analyzed to derive the timing and source of action potentials. This post hoc analysis, known as *spike sorting*, is not error-proof; sometimes signals from one neuron might be assigned to two separate sources, or signals from two neurons might be assigned to a single source. Because the source of action potentials uncovered in the voltage traces is inferred, the term *unit* is used to emphasize the uncertainty.

about the origin of motor behavior from neural signals.

The two versions of the model are distinguished by how they assume the planned lick direction is encoded in the spike train data. Our first model assumes, consistent with Li and colleagues' analysis, that the planned lick direction is encoded in the *average firing rate*. In other words, our first model assumes ALM neurons encode planned decision behaviors using a *rate code*. Our second model considers the alternative theory that the planned behavior is encoded in the firing rate in time-dependent manner. This *time-dependent rate code* model accounts for the possibility that ALM neurons rely on a difference in firing rate during a only short time interval to encode the planned behavioral response. By comparing the adequacy of these models in simultaneously describing both types of data, we intend to compare the viability of these theories about the latent neural process, while still accounting for the complexity of the behavioral data and potential individual differences across mice.

Rate code model

We begin our description of the neurocognitive process expressed in our first model (as depicted in Figure 4.2) with the distribution of the neural data, which is submitted to the model as binned spike counts. For each trial, t, the number of spikes, x, observed in each bin, i, for each unit, u, is assumed to be generated via a Poisson process:

$$\mathrm{x}_{mstui} \sim \text{Poisson}(\lambda_{mstu})$$

with some underlying rate, λ:

$$\lambda_{mstu} \sim \text{Gamma}(\alpha_m, \theta_m) \tag{1}$$

that is constrained to be positive.

Li and colleagues (2015) observed that a majority of units were selective for the planned behavioral response. This selectivity is captured by the preference parameter, π, for each

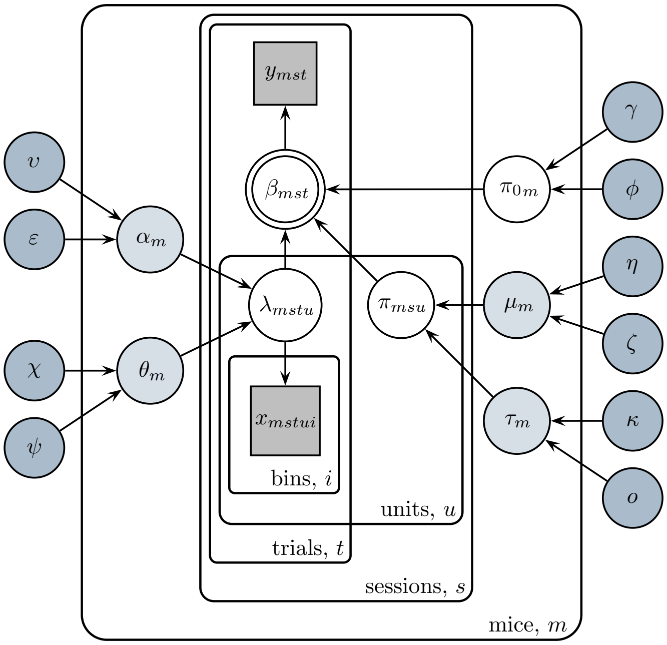


Figure 4.2. Graphical model representation of the rate code model. White and gray nodes indicate model parameters and data, respectively. Connections among the white and gray nodes express the core neurocognitive process captured by the model. Light blue nodes indicate mouse-level hyperparameters, and dark blue nodes indicate experiment-level hyperparameters.

unit:

$$\pi_{msu} \sim \text{Normal}(\mu_m, \tau_m)$$

where positive values for π indicate a selectivity for left responses, negative values π indicate a selectivity for right responses, and values near zero indicate that the unit is not selective.

The preference parameters are used to pool the rates across the population of neural units in order to produce a consensus regarding the planned response on each trial. Specifically, the rate parameters are summed using the preferences as weights. In this way, the responses of units with strong preferences ($|\pi| \gg 0$) are upweighted, while the responses of units with no or minimal preference ($|\pi| \approx 0$) are largely discounted. However, the neural population response may not be the only factor that determines the observed behavioral response. To account for the possibility that individual mice have a bias to lick in a particular direction regardless of the bar position presented, we include a baseline bias parameter, π_0:

$$\pi_{0_m} \sim \text{Normal}(\gamma, \phi)$$

which is interpreted in the same way as the preference parameter. Combining the mouse-specific baseline behavioral bias with the trial-specific consensus neural response provides the overall bias, β, toward a left or right response on a given trial:

$$\beta_{mst} = \pi_{0_m} + \sum_{u=1}^{U_{ms}} \lambda_{mstu} \pi_{msu} \tag{2}$$

By subjecting the bias parameter to a logistic function, it is constrained to the $[0, 1]$ interval. The transformed bias parameter may now be used as the success probability parameter in a Bernoulli process to generate the observed response:

$$\text{y}_{mst} \sim \text{Bernoulli}(\texttt{ilogit}(\beta_{mst})) \tag{3}$$

Values of β near 1 will generate a left response ($y = 1$) with a high probability, while values near 0 will generate a right response ($y = 0$) with a high probability. Values of β near 0.5 indicate that there is no tendency toward either a left or right response based on the mouse's behavioral tendencies or the neural response, and a random response would be generated.

The equations above serve to quantify the theoretical description of the underlying process occurring on each trial. Of special note is the deterministic statement in Equation (2), which links the behavioral model (i.e., the Bernoulli process) to the neural model (i.e., the Poisson process) by allowing the consensus in the neural population to mathematically determine the behavioral response probability. That this link between the model components is a deterministic statement, describing a unidirectional effect, rather than a multivariate stochastic distribution, describing bidirectional latent relationships among parameters, distinguishes this model as a neurocognitive process model.

In order to account for neural and behavioral data from all sessions and all mice simultaneously, random effects of session and mouse are included. To accomplish this hierarchical extension, prior distributions for mouse-level hyperparameters are specified:

$$\begin{aligned}
\mu_m &\sim \text{Normal}(\eta, \zeta) \\
\tau_m &\sim \text{Gamma}(\kappa, o) \\
\alpha_m &\sim \text{Gamma}(\upsilon, \varepsilon) \\
\theta_m &\sim \text{Gamma}(\chi, \psi)
\end{aligned}$$

Two of these mouse-level hyperparameters have straightforward interpretations: μ_m, the mean unit preference for an individual mouse, and τ_m, the standard deviation of unit preferences for an individual mouse. The proper interpretation of the other two mouse-level hyperparameters, α_m and θ_m, is less intuitive. However, these parameters need not have clear interpretations in and of themselves in order to be of meaningful theoretical consequence. These parameters are included to effect the assumption that the rate parameter inferred on each trial is drawn from some hierarchical distribution of rates that is unique to each mouse (see Equation (1)).

All four mouse-level parameters are themselves assumed to be drawn from experiment-wide distributions. To complete the model specification, a prior is specified for each parameter used to define an experiment-wide distribution:

$$\begin{aligned}
\eta &\sim \text{Normal}(0, 10) \\
\zeta &\sim \text{Gamma}(.1, .1) \\
\kappa &\sim \text{Gamma}(.1, .1) \\
o &\sim \text{Gamma}(.1, .1) \\
\upsilon &\sim \text{Gamma}(.1, .1) \\
\varepsilon &\sim \text{Gamma}(.1, .1) \\
\chi &\sim \text{Gamma}(.1, .1) \\
\psi &\sim \text{Gamma}(.1, .1)
\end{aligned}$$

With the exception of η, which captures the mean preference across all units in the entire experiment, assigning a semantic interpretation to many of these experiment-level hyperparameters again proves difficult. Yet each is necessary to include in order to provide a comprehensive account of the entire dataset.

Time-dependent rate code model

In the rate code model, the behavioral data is in part determined by the consensus in the neural response, as captured by the neural component. This neural component was built from the assumption that the behavioral response is encoded in the *average* firing rate specifically. This assumption about the method of encoding used by ALM neurons was based on Li and colleagues' (2015) use of differences in the total number of spikes observed on each trial type to assess selectivity. While this strategy is able to easily detect units that exhibit some degree of selectivity across an entire trial epoch, it would make it difficult to detect units that are reliably selective on shorter time scales. If ALM neurons instead rely on *time-dependent* firing rates to encode the planned behavior, then our first model might be underestimating the number of selective units.

In order to capture this possibility in our second model, rather than inferring a rate parameter directly from the binned spike count data, we will first assess the of the selective value of each time bin for a given unit, then use those values to produce a weighted rate. If this *time-dependent rate code* model provides a better account of the latent neurocognitive process, we expect that it might not only allow for an improved assessment of selectivity and enable the model to better capture the observed data.

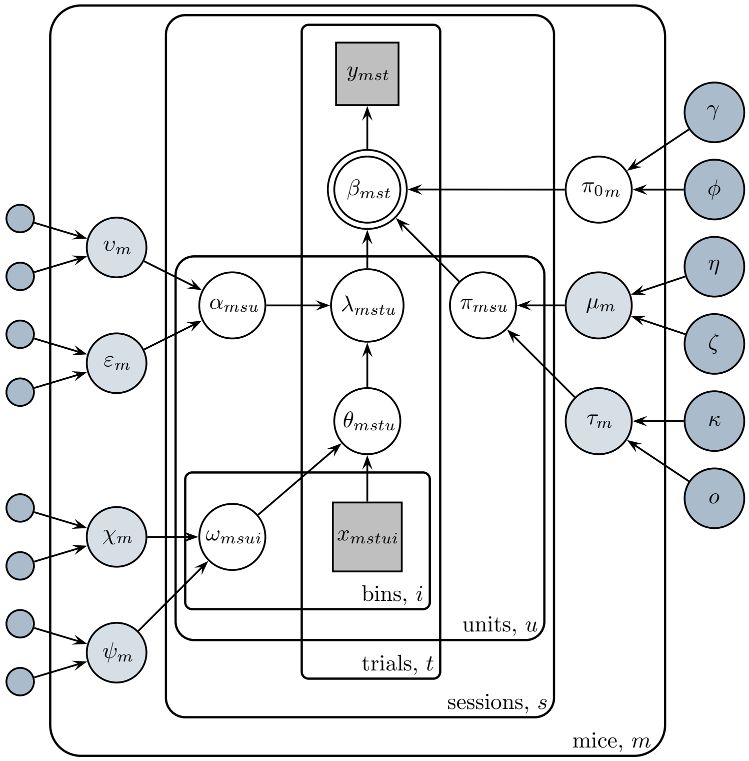


Figure 4.3. *Graphical model representation of the time-dependent rate code model.*

We implemented the time-dependent rate code model by extending the neural component of the rate code model, specifically by incorporating a gamma regression layer in the neural model component. In the rate code model, all I bins are viewed as interchangeable by the model. Whether a spike count was from the first bin or the last bin on a given trial was irrelevant; every bin observed on a given trial was assumed to have equal predictive value. Specifically, the number of spikes, x, observed on each trial, t, was assumed to be

Poisson-distributed with some underlying rate, λ:

$$\lambda_{mstu} \sim \text{Gamma}(\alpha_m, \theta_m)$$
$$x_{mstui} \sim \text{Poisson}(\lambda_{mstu})$$

To create our second model, these equations will removed from the model and replaced.

In our second model's alternative specification of the neural component, bins are no longer considered interchangeable; the identity of the bins is important, as now each is permitted to have a different predictive value. The rate observed on each trial is again assumed to be gamma-distributed:

$$\lambda_{mstu} \sim \text{Gamma}\left(\alpha_m, \frac{\alpha_m}{\theta_{mstu}}\right)$$

however, we now include a linear predictor of the mean of the gamma distribution:

$$\theta_{mstu} = .001 + \sum_{i=1}^{I} x_{mstui}\omega_{msui}$$

The two new equations above state that the mean rate observed on each trial, θ, is a linear combination of the binned spike count, x, and the corresponding importance of that bin in distinguishing between left and right responses, ω. (A small value, .001, is also included in the definition of the linear predictor to ensure that, in the event that no spikes are observed on a given trial ($\forall\, i : x_{mstui} = 0$), we will not produce an undefined value for the second parameter of the prior distribution for λ.) It is important to note that this bin importance parameter is not trial-specific, as it reflects the predictive value of a given time bin across trials for a given unit.

By setting a truncated normal prior on bin importance, it is constrained to positive values:

$$\omega_{msui} \sim \text{Normal}(\chi_m, \psi_m)_{\mathcal{I}(0,\infty)}$$

High values of ω indicate that the number of spikes observed in a given bin are informative with respect to the observed behavioral response. Values of ω near 0 indicate that a bin is not informative. The incorporation of bin weighting in the distribution for λ affects its interpretation. It is no longer a true rate parameter, but now is rather more of an assessment of the neural activity across a given trial, based on how that activity generally informs the eventual behavioral response.

The prior for ω is hierarchically extended to address the neural data across all sessions and all mice. For clarity and brevity, prior specifications for the mouse-level parameters, χ and ψ, and for the pursuant experiment-level hyperparameters are omitted here. We also omit the remainder of the model specification as it applies to the shape parameter, α, the unit preferences, π, the baseline behavioral biases, π_0, the combined trial-specific bias, β, and the behavioral response, y, as this will all be the same as in the rate code model.

Results

A comparative test of two complex hypotheses

We begin our analysis of Li and colleagues' dataset (Li et al., 2014) by performing a model comparison to test whether it is more likely that ALM neurons encode the planned behavioral response using a rate code or a time-dependent rate code. This not only tests an important research question, but also serves to demonstrate the how neurocognitive process models are naturally suited to theory testing.

We elected to use deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002, 2014) as our model comparison metric as it was designed to compare hierarchical Bayesian models specifically and is easily calculated from MCMC samples.[2] Similar to other model comparison metrics used in Bayesian analysis, the DIC balances goodness of fit against model complexity, as the DIC assesses deviance while heavily penalizing models

[2] DIC is highly sensitive to imperfect convergence of the MCMC chains. As such, we will require (1) a Gelman-Rubin diagnostic statistic ($\hat{R}$; Gelman & Rubin, 1992) under 1.01 for all model parameters, and (2) a DIC value that appears stable across quartiles of samples.

with a high effective number of model parameters. Although the raw DIC value is not easily interpreted, comparatively low DIC values (i.e., at least 3–7 units lower; Spiegelhalter et al., 2002) indicate notably better fit. Thus, we will observe the difference in DIC values for the two neurocognitive process models.

We found that the rate code model provided a better fit to the joint dataset than the time-dependent rate code model ($\Delta\text{DIC} = 75.12 \times 10^2$). This result provides quantitative evidence that it is more likely that ALM neurons encode the planned response in their average firing rate, than the alternative that ALM neurons are reliably selective on shorter time scales. Now that we have inferred which theoretical neurocognitive process is more likely, we may look more closely at the behavior of the superior model to learn more about the latent process.

New conclusions about the underlying neurocognitive process

Because we implemented the rate code model in a hierarchical Bayesian framework, we can use the posterior distributions of model parameters which have meaningful semantic interpretations as the basis for new inferences about the nature of the underlying neurocognitive process. We began by examining the selectivity of individual units as captured by the preference parameters, π. Specifically, we sought to evaluate Li and colleagues' (2015) conclusion that a *majority* of ALM neurons are selective for the behavioral response.

Visual comparison of unit preference parameter point estimates, $\hat{\pi}$, to the raw data (see Figure 4.4) suggests that the model produces sensible inferences about the selectivity of individual units. As mentioned in our earlier description of the rate code model's specification, preference parameters are interpreted such that positive values indicate a selectivity for left responses, negative values indicate a selectivity for right response, and values near 0 indicate no selectivity. Therefore units for which the 95% credible interval[3] of the preference

[3] Credible intervals are similar in concept to confidence intervals, but are more directly interpretable (see Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016, for an accessible overview). A 95% credible interval is a range of possible parameter values that contains the true value of the parameter with 95% probability. In our calculations, the lower and upper bounds are the 2.5th and 97.5th percentile of the posterior samples,

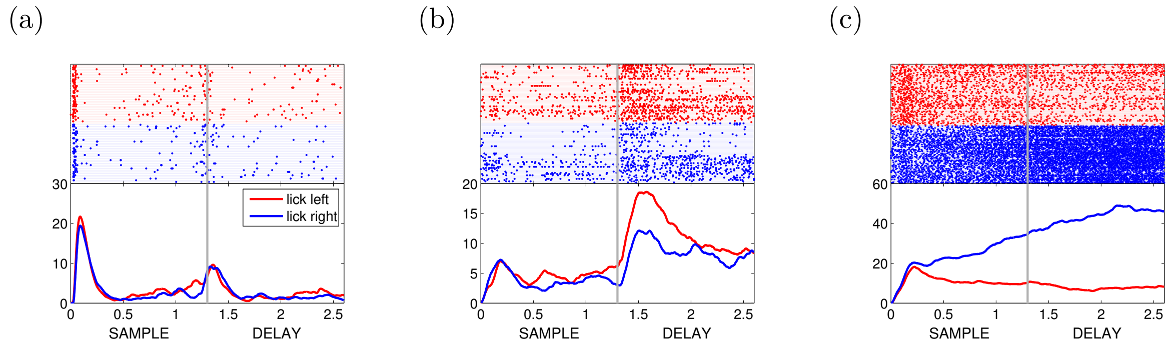


Figure 4.4. Raw data for units with different inferred preferences. Raster plots (top) and peristimulus time histograms (PSTHs; bottom) of units for which the model generated point estimates indicating: (a) no preference ($\hat{\pi} = 0.07$), (b) a moderate preference for a left response ($\hat{\pi} = 1.56$), and (c) a strong preference for right responses ($\hat{\pi} = -3.45$).

parameter posterior distribution did not include 0 were classified as selective. If the 95% credible interval included 0, the unit was classified as not selective.

These classifications were used to infer the proportion of units across the entire experiment that were selective for the upcoming response during the preparatory interval (i.e., the sample and delay periods, before a response might be executed). Only a minority of units (∼20%) were found to be selective for the upcoming motor response during the preparatory interval. As this general pattern was consistent across all 99 sessions (see Figure 4.5), it suggests that only a subset of ALM neurons encode the planned response. This conclusion stands in stark contrast to Li and colleagues' finding that ∼50% of units exhibit selectivity for the behavioral response before it is executed.

To explore this conflict further, we used the rate code model a second time to analyze the behavioral responses and the neural data from the response interval only. The number of units that were selective during this trial epoch was inferred in the same fashion. This second application of the model enabled us to assess separately the proportion of units that were selective for the planned response, the proportion of units that were selective for the executed response, and the proportion of units that were not selective at any point at all during the trial period. This also permitted us to compare our results more directly against

respectively.

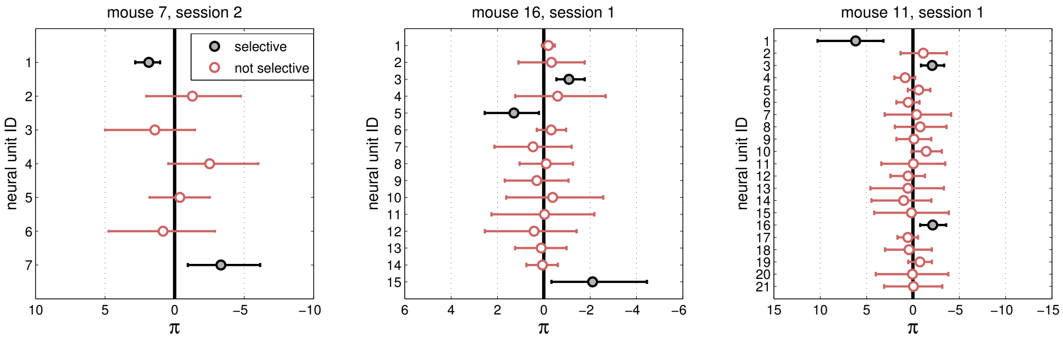


Figure 4.5. Posterior distributions of preference parameters for all units in each of three sessions. Circles represent point estimates, $\hat{\pi}$, and bars represent 95% credible intervals. Units classified as selective are marked in black. Units classified as not selective are marked in red. Regardless of the number of units observed in a given session, the same pattern emerges: Only a minority of units are selective for the upcoming response.

Li and colleagues' (2015) results. Our findings are presented in Figure 4.6.

We again observe a striking difference in conclusion. Li and colleagues (2015) assert that, across the entire trial period, 73% of units are selective for the motor response, while the neurocognitive process model suggests that just 43% are selective.

That the inferred number of selective units is lower is not entirely unexpected: Given that this is a hierarchical model, the number of units classified as selective would likely be lower due to shrinkage, meaning the drawing of all parameter estimates toward the group mean (in this case, the population mean of unit preferences, $\hat{\eta} \approx 0$). However, given that the proportion of units classified as selective during the response interval in our analysis is almost the same as in Li and colleagues' (2015) analysis, the effect of shrinkage may be minimal. We contend that the reason for the large discrepancy in the proportion of units classified as selective is that Li and colleagues' method of analysis[4] disregarded the complexity of the relationship between the neural and behavioral data. Specifically, their method led to an overestimation because the neural data from each unit was considered in isolation. In the neurocognitive process model, the neural data from each unit was considered not only in the context of the population of units, but also in the context of the larger neurocognitive process, which accounted for purely behavioral elements (i.e., the baseline bias of each mouse,

[4] 1,245 t-tests, no correction for multiple comparisons.

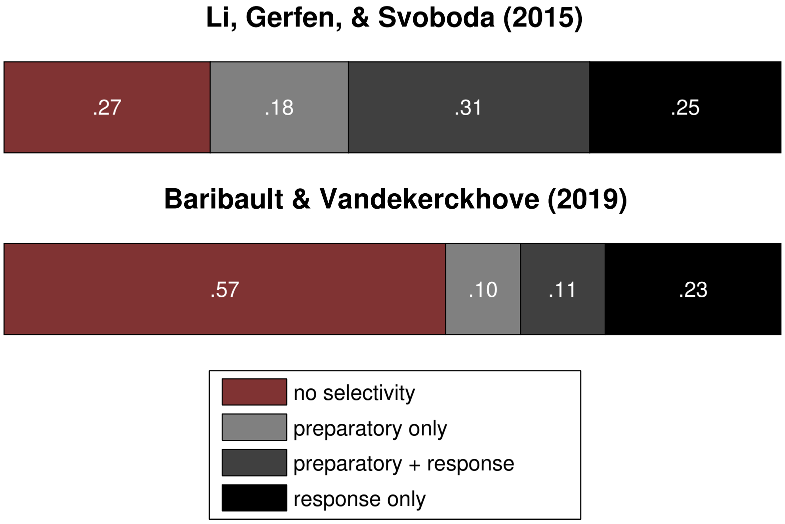


Figure 4.6. *Proportion of units that are selective during different trial epochs.* Although the proportion of units found to exhibit selectivity during the response epoch is similar, the proportion of units found to be selective during the sample and delay epochs (i.e., during the preparatory interval, before a response was made) is notably smaller in our analysis. As a result, we observe that overall far more units exhibit no selectivity at any point during the trial period than was observed in Li and colleagues' (2015) analysis.

π_0). As such, the model was able to provide a clearer assessment of what variability is due to the unit selectivity versus other factors, and thereby provide inferences that are arguably more valid.

Finally, we investigated whether individual mice exhibited response biases by examining the posterior distributions for the baseline bias parameters, π_0, for each mouse. (This parameter is interpreted similarly to the unit preference parameters, π.) If there were no response biases other than that which was accounted for by the neural model component, then the posterior distributions of the baseline bias for each mouse, π_{0_m}, should be centered on 0. This is not observed. While there is no population-wide bias ($\hat{\gamma} \approx 0$), the majority of individual mice have strong response biases regardless of trial type. As shown in Figure 4.7, 13 of 19 mice have credible intervals for π_0 that do not include 0. This suggests that mice exhibit notable individual differences beyond that which can be explained by the available neural data from ALM.

These secondary analyses show that, because we formulate neurocognitive process models

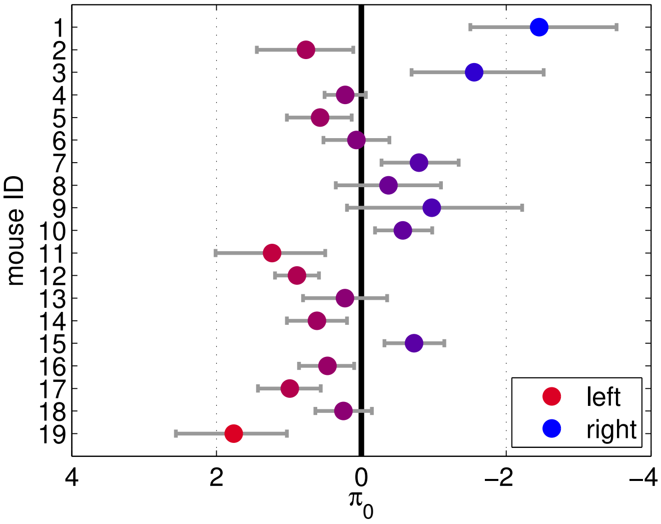


Figure 4.7. Posterior distributions of baseline bias parameters for each mouse. Circles represent point estimates, $\hat{\pi}_0$, and bars represent 95% credible intervals. The color of the point estimate marker indicates the strength of the bias. Red indicates the mouse has a strong bias toward a left response, blue indicates the mouse has a strong bias toward a right response, and purple indicates no or minimal bias.

in a hierarchical Bayesian framework, we can use the model to make novel types of conclusions about both cognitive aspects and neural aspects of the latent process. As such, our approach not only facilitates hypothesis testing, but also allows us to make novel statements about specific features of the underlying neurocognitive process through estimation-based analyses.

Discussion

Through this case analysis, we showcased the unique advantages of the neurocognitive process modeling approach as a method of developing joint models of neural and behavioral data in a hierarchical Bayesian framework. The first such advantage we sought to highlight is the flexibility of the technique. A neurocognitive process model is defined by the inclusion of a deterministic statement linking the neural and behavioral components, however, the design of each of these components can and should be tailored to the experimental context, and likewise can and should be structurally altered to capture researchers' hypotheses about unobserved features of neural and cognitive processes.

Each neurocognitive process model component — and the linkage between the components — may be independently altered and extended. Here, we demonstrated how to alter the structure of the neural component to incorporate a different, rival theory of the underlying neural process. Another possible way to alter the neural component is to use a different likelihood function entirely. In this way, one could define a neurocognitive process model that accounts for fMRI, EEG, or other types of neural data in place of spike counts. Similarly, one could also swap out the behavioral component of the model in order to accommodate a different experimental context. The selection of the cognitive model component allows a neurocognitive process model both to account for new types of behavioral data, such as continuous accuracy (e.g., distance from a target position), and simultaneously, a different theoretical account of the underlying cognitive process (e.g., a decision tree, a memory process, or a categorization process).

The second advantage that we showcased here was the relative ease of conducting model comparisons. Often hypotheses concerning joint datasets are put in terms of mechanisms: One area connects to another in order to enhance, inhibit, or otherwise modulate its activity, a given area of cortex encodes one aspect of a stimulus or response, a particular computation is performed by neurons in a given an area. These hypotheses are difficult to assess with current Bayesian joint modeling techniques (e.g. Turner et al., 2013), as such methods are only intended to capture abstract linear relationships. However, the neurocognitive process modeling approach provides a structured method for instantiating complex theories in a model, even when the theory incorporates many abstract ideas and principles on both the behavioral and neural side.

For some, the high level of theoretical commitment required to formulate a neurocognitive process model may be seen as restrictive and undesirable. However, if multiple plausible hypotheses are captured in each of a handful of models, a model comparison might be performed across a large set of models to quantitatively determine which theories are comparatively more plausible based on the observed data. Furthermore, if a viable comprehensive theory is not currently available, and thus an exploratory analysis of a joint dataset is the goal, then the aforementioned work by Turner and colleagues (Turner et al., 2013) should be the preferred approach, as it is an excellent approach to theory generation. We see neurocognitive process models less as a direct competitor to other approaches, and more as a complementary technique.

It is our hope that the introduction of neurocogntive process models adds a useful new technique to our collective analytic toolbox. While we acknowledge that these models may be labor-intensive to specify and computationally resource-intensive to fit, we believe the opportunities they present for theory testing are worth the effort.

References

Forstmann, B. U., Brown, S., Dutilh, G., Neumann, J., & Wagenmakers, E.-J. (2010). The neural substrate of prior information in perceptual decision making: a model-based analysis. *Frontiers in human neuroscience*, *4*, 40.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-sma facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538–17542.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Li, N., Chen, T.-W., Guo, Z. V., Gerfen, C. R., & Svoboda, K. (2015). A motor cortex circuit for motor planning and movement. *Nature*, *519*(7541), 51–56.

Li, N., Gerfen, C. R., & Svoboda, K. (2014). *Extracellular recordings from anterior lateral motor cortex (ALM) neurons of adult mice performing a tactile decision behavior* [Data file]. Retrieved from `https://crcns.org/data-sets/motor-cortex/alm-1`. doi: 10.6080/K0MS3QNT

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, *23*(1), 103–123.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 221–247.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 485–493.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206.

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: Simultaneous modeling of eeg, fmri, and behavioral data. *Neuroimage*, *128*, 96–115.

Turner, B. M., Van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: the neural drift diffusion model. *Psychological review*, *122*(2), 312.

Turner, B. M., Wang, T., & Merkle, E. C. (2017). Factor analysis linking functions for simultaneously modeling neural and behavioral data. *NeuroImage*, *153*, 28–48.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., & Forstmann, B. U. (2011). Neural correlates of trial-to-trial fluctuations in response caution. *Journal of Neuroscience*, *31*(48), 17488–17495.

Van Veen, V., Krug, M. K., & Carter, C. S. (2008). The neural and computational basis of controlled speed-accuracy tradeoff during task performance. *Journal of Cognitive Neuroscience*, *20*(11), 1952–1965.