Individual Differences in Cognitive Science: Conceptual and Methodological Issues

by

Zina Berry Ward

BA in Economics and Philosophy, Williams College, 2012

MPhil in History and Philosophy of Science, University of Cambridge, 2013

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Zina Berry Ward

It was defended on

May 20, 2020

and approved by

Colin Allen, Distinguished Professor, Department of History and Philosophy of Science

Mazviita Chirimuuta, Associate Professor, Department of History and Philosophy of Science

James F. Woodward, Distinguished Professor, Department of History and Philosophy of Science

David Danks, L.L. Thurstone Professor of Psychology and Philosophy,
Department of Philosophy (Carnegie Mellon University)

Dissertation Director: Edouard Machery, Distinguished Professor,
Department of History and Philosophy of Science

Individual Differences in Cognitive Science: Conceptual and Methodological Issues

Zina Ward, PhD

University of Pittsburgh, 2020

A primary aim of cognitive science is the investigation of psychological and neuroscientific generalizations that hold across subjects. Individual differences between people's minds and brains are pervasive, however, even among subjects considered neurotypical. In this dissertation, I argue that both scientific practice and our philosophical understanding of science must be updated to reflect the presence of such individual differences. The first half of the dissertation proposes and applies a philosophical account of what it takes to explain variation, while the second half identifies several methods in psychology and neuroscience that demand reform in light of existing individual differences.

Table of Contents

List of Figures

Preface

I am extremely grateful for the opportunity to become a philosopher. I didn't know it was a possibility until I met Bojana Mladenović at Williams. My advisors and friends in HPS at Cambridge turned the possibility into a genuine option. And my time at Pitt HPS made it happen.

I owe a great deal to Edouard Machery, who has been a source of insightful feedback and professional advice over the last few years. Edouard has an unusual knack for getting straight to the heart of an issue; at every stage, my doctoral work has benefited from his incisive comments. In our meetings, Edouard's openness to disagreement and willingness to change his mind have helped me become a more effective advocate for my views. I hope that my graduate seminars will one day be half as illuminating as his, which were a highlight of my time in Pittsburgh.

I am also thankful for the guidance of Jim Woodward, whose feedback substantially improved Chapters 2 and 4 of this dissertation. Jim was extremely generous with his time and advice during the job market process. Moreover, as I've worked to develop my scholarly voice, Jim's work has served me as a model of deep, careful philosophical thinking. I would also like to thank the other members of my dissertation committee: Colin Allen, Mazviita Chirimuuta, and David Danks. They provided valuable comments on many parts of this dissertation and their presence has made Pittsburgh an amazing place to do philosophy of cognitive science.

For written comments on various chapters, I thank Rachel Achs, Katie Creel, Dmitri Gallow, Gabby Johnson, Shivam Patel, Marco Viola, and Naftali Weinberger. Parts of the dissertation also benefited from conversations with Lena Kästner, Colin Klein, Amanda Sue Luby, Brendan McVeigh, Aaron Novick, Maria Olkkonen, and Jessey Wright. I have gotten some

of the most valuable feedback over the years from the HPS graduate students' work-in-progress series. I presented several of the dissertation chapters there and each one is better for it.

Thanks to Joyce MacDonald, Joann McIntyre, Katie Labuda, and Natalie Schweninger for their kindness and patient help with all of the logistics associated with getting a PhD. This past year, I was fortunate to receive a Dissertation Completion Fellowship from the American Council of Learned Societies and Andrew W. Mellon Foundation. During my doctorate, I have also received generous financial support from the Center for the Neural Basis of Cognition, Ruth Crawford Mitchell Memorial Fund, and Wesley C. Salmon Fund.

Like many, I can't say enough good things about the community of graduate students at Pitt HPS. While I've been in Pittsburgh, I've benefited from the wisdom and friendship of Mikio Akagi, Nuhu Osman Attah, Michael Begun, Nora Boyd, Dave Colaço, Haixin Dang, Josh Eisenthal, JP Gamboa, Mahi Hardalupas, Joe McCaffrey, Mike Miller, Aaron Novick, Elizabeth O'Neill, Evan Pence, Lauren Ross, Katie Tabb, Morgan Thompson, Tom Wysocki, and others. The Center for Philosophy of Science has also brought (unfortunately brief) good times with Ingo Brigandt, Andrew Buskell, Helen Curry, Alison Fernandes, Sara Green, Karen Kovaka, Maria Serban, Raphael Scholl, and Naftali Weinberger. My deepest gratitude goes to Siska De Baerdemaeker, Liam Kofi Bright, Katie Creel, Haixin Dang, and Jacob Neal for being the amazing cohort-mates, tennis partners, trivia teammates, practice interviewers, restaurant critics, coffee shop co-workers, concertgoers, secret Santas, soccer teammates, and reading group buddies that they are.

I have a wonderful partner who also happens to be a fantastic editor and sounding board. Shivam's feedback shaped almost every part of this dissertation. I've learned from him to write every word of exposition keeping in mind the argument to come. As the philosophical butcher, he's also tried to teach me to let go of what I've written if it's just not working, a lesson I'm still

trying to absorb. In the lows of the market, he dealt with my doubting and quiddit and lifted me up to face another cover letter, another interview. I don't know where I'd be without him.

I may never have found my way into academia without Grady, who, at age four, happily played along when his big sister gave him homework to do after daycare. These days, I can always count on him to celebrate my successes and offer an optimistic take on my failures. And finally, my parents' unconditional support and intellectual curiosity have shaped my path in more ways that I could probably ever understand. Their interest in what I was doing (no matter how unfamiliar and abstract it was!) gave me confidence to do whatever made me happy. I am exceptionally lucky.

1.0 Introduction

1.1 Psychology Divided

Lee Cronbach's (1957) "The two disciplines of scientific psychology" has become a classic diagnosis of a fundamental "schism" at the heart of psychology. Cronbach describes a discipline divided between experimental psychology and correlational psychology, each with its own "method, thought, and affiliation" (*ibid.*, 671). The two camps' differences are rooted in their divergent interests: "while the experimenter is interested only in the variation he himself creates, the correlator finds his interest in the already existing variation between individuals, social groups, and species" (*ibid.*, 671). Experimental psychology aims to exercise precise control over situational variables, permitting tests of hypotheses and claims about causation. Correlational psychology, pursued by a smaller and more diffuse group of inquirers, analyzes data from "Nature's experiments" (*ibid.*, 672). Although he notes that there is no antagonism between the two sides, Cronbach sees the division as fundamentally harmful. A united psychology would study both the variance among treatments and the variance among organisms.

More than sixty years later, some parts of Cronbach's paper are outdated, such as his claim that personality, social, and child psychologists rarely use experimental methods. Nevertheless, there is a central theme that remains just as relevant now as when the paper was published: many psychologists tend to ignore naturally occurring variation or view it as a distraction. As Cronbach explains, "individual differences have been an annoyance rather than a challenge to the experimenter," who tries to minimize within-treatment variation "by any possible device" (*ibid.*, 674). The experimentalist may use animal subjects with "controlled heredity and controlled

experience," recruit human subjects from the same culture, or use just a single subject, all in order to reduce variation and "get those embarrassing differential variables out of sight" (*ibid.,* 674). The struggle to minimize within-treatment variability and the neglect of such variability in mainstream theorizing coexist with the recognition that individual differences deserve study. Cronbach's paper is thus an enduring statement of the uncomfortable place that individual differences occupy in psychology and, one may argue, in cognitive science generally.

This dissertation is about such individual differences. By "individual differences" or "individual variation," I mean variability between human subjects that is stable over time. Unlike some authors (e.g. Chen et al. 2015), I do not include intra-individual differences (i.e., variation within subjects over time or between testing sessions) under this heading. I am interested in the ways in which individuals are consistently different from one another, where those differences are not intentionally induced by the experimental design. Although included under this definition, my emphasis is not on variation filtered through a demographic lens, such as sex or race differences, nor on variation studied by specific sub-disciplines of cognitive science, such as cultural psychology. Such "group differences" are a subset of all individual differences; there is arguably far more variability for which "the relevant classification is unknown" (Mollon et al. 2017, 5). Individual differences are also typically distinguished from differences between healthy and diseased (or neurotypical and non-neurotypical) subjects.[1] The term "normal variation" is accordingly sometimes used interchangeably with "individual differences." Hence, my focus is on variation that is not (or has not yet been) associated with existing demographic groups or clinical populations. Individual differences in this sense may be either "discrete" or "continuous"

[1] Nevertheless, there is a growing sense that some pathologies lie at the extreme end of a continuum that includes healthy but variable individuals, an understanding reflected in the dimensional approach of the NIMH's Research Domain Criteria (RDoC).

(Gangestad and Snyder 1985, Danks and Eberhardt 2009, Bartlema et al. 2014, Machery 2017): people may differ from one another qualitatively, as matter of kind, or quantitatively, as a matter of degree.

1.2 The Import of Individual Differences

Individual differences in cognitive science have only sporadically been the object of philosophical attention. Variation has occasionally been discussed in philosophical work on scientific explanation (Tabery 2009), human nature (Samuels 2012), and natural kinds (Buckner 2016). There is also a sense in which the topics of multiple realization and the robustness of biological systems relate to individual differences, concerned as they are with what sorts of differences between organisms make a difference (Shapiro 2000; Levy 2017). In my view, however, individual differences in cognitive science deserve more direct and sustained philosophical examination.

This is, first, because variation often poses a significant challenge to investigators seeking generalizations about the mind and brain. The discovery that neurotypical subjects differ from one another in their behavior, physiology, or neurological features can make it difficult to draw conclusions about how cognitive or neural systems work. That individual differences have the potential to lead scientists astray can be seen in the history of experimental psychology. It is now widely recognized that, given individual differences in performance on a task, a model that fits individual performance data may not fit data averaged from multiple subjects (and vice versa). This means that analyzing aggregate data only can be misleading. In the 1950s, several articles calling attention to the danger of averaging appeared in the psychological literature (Sidman

1952, Hayes 1953, Bakan 1954, Estes 1956). Despite these warnings, researchers investigating a variety of phenomena continued to fit curves to average data only. Because some of the phenomena were subject to significant individual differences, this led to serious scientific errors that persisted for decades.

For example, as Estes (2002) explains, in the mid-twentieth century, there was consensus among memory researchers that forgetting curves – which describe how many items from a list subjects forget over time – could be described by power functions. This became such a widely-recognized finding that it was sometimes called "the power law of forgetting." The same thing happened in the literature on skill (Newell and Rosenbloom 1981). It was generally accepted that curves of practice were best described by power functions, with "the power function [taking] on a basic role in models of skill acquisition" (Estes 2002, 6). There is now considerable evidence, however, that power functions' fits to such data are often purely artefactual. Forgetting curves may be best fit by an *exponential* function once the data are disaggregated and analyzed for each subject (Anderson and Tweney 1997, cf. Averell and Heathcoate 2011). In recent years, the causes of this "power law artifact" and the conditions under which it arises have been a topic of much discussion (Myung et al. 2000, Estes and Maddox 2005, Cohen et al. 2008). Individual differences are at the center of the story, since "performance differences among individuals" have been shown to be necessary for the artifact to occur (Myung et al. 2000, 836). Estes and Maddox (2005) summarize our current understanding of this problem: "there is danger…that individual differences among subjects with respect to values of a model's parameters may cause averaging to produce distorted inferences about true patterns of individual performance and the cognitive processes underlying them" (403).

Another reason individual differences are deserving of philosophical attention is that some authors have suggested that they are partly responsible for the so-called "replication crisis"

in parts of psychology, neuroscience, and the biomedical sciences. Variation among subjects detracts from statistical power, and underpowered studies can lead to replication failures in two ways. First, if research is systematically underpowered, there will be a high false positive rate in the scientific literature (Ioannidis 2005, Krzywinski and Altman 2013). False positive results are unlikely to replicate. Second, if replication attempts are underpowered, replications may fail to confirm the existence of real effects (Simonsohn 2013). Because individual differences can reduce the power of both original research and replication studies, they can contribute to both kinds of replication failures. Poline et al. (2010) have suggested that the lack of reproducibility in brain imaging is partly due to functional and anatomical variability in the brain (and the lack of analysis strategies for dealing with it). Bolger et al. (2019) likewise claim that some replication failures are due to neglect of causal effect heterogeneity, or individual differences "in the size and/or direction of a cause-effect link" (601). They argue that "adequately modeling heterogeneity can help protect researchers against underpowered studies, illusory findings, and replication failures" (*ibid.*, 602; see also Gelman 2015).

Even if one is skeptical that individual differences deserve much of the blame for the replication crisis, it is certainly true that they have generated a need for methodological reform in cognitive science. Because individual differences have been widely neglected, commonly used stimuli, experimental protocols, and analytical and theoretical approaches need to be modified to tackle variation. Hedge et al. (2017), for instance, have recently argued that individual differences cannot be effectively studied using popular paradigms like the Stroop and Posner cuing tasks. They claim that tasks for producing generalizations are often not suitable for studying variation in cognition, and that our historical focus on the former has led to a shortage of tasks tailored to the latter. Perceptual psychologists are in a similar boat, as their stimuli have been deliberately designed to elicit uniform responses. Standard protocols for color vision experiments, for

instance, involve "correcting" for individual differences in luminance sensitivity. The deliberate erasure of inter-subject differences via the tailoring of stimuli makes sense in certain research contexts, but new experimental protocols are needed when variation itself is the investigative target.[2]

Variation should also attract the attention of philosophers because we seem to be at the beginning of a surge of interest in individual differences in many areas of cognitive science. Olkkonen and Ekroll (2016) claim that color constancy research in perceptual psychology has largely ignored individual variation, but that a "new era" is beginning in which individual differences will be "the main focus of interest" (176). Lonsdorf and Merz (2017) argue that a "paradigm shift" is occurring in research on fear conditioning as researchers begin to appreciate "the role and opportunities of inter-individual differences in fear conditioning processes" (4). And Sporns (2013) notes a "resurgence of interest" in individual variation in brain connectivity (59). Others have recently advocated greater attention to individual differences in developmental psychology (Foulkes and Blakemore 2018), cognitive psychology (Navarro et al. 2006, Zeigenfuse and Lee 2009), psycholinguistics (Swets 2015), perceptual psychology (Wilmer 2008, Webster 2015, Emery and Webster 2019), and cognitive neuroscience (Seghier and Price 2018).

This surge in interest is due partly to scientists' increasing awareness that individual differences present not only inferential challenges, but also opportunities. The positive potential of variation was emphasized several decades ago by Benton Underwood in his classic (1975)

[2] More mundane methodological changes are required as well. Individual differences research demands a greater number of subjects than research aiming to uncover inter-individual regularities (Thirion et al. 2007, Swets 2015, Foulkes and Blakemore 2018); it may call for abandoning tried-and-true methodological strategies like the randomization of trial order (Mollon et al. 2017); and it can necessitate the adoption of methods that are rarely used in certain areas of cognitive science, like longitudinal studies or the testing of subjects on a variety of tasks.

paper, "Individual Differences as a Crucible in Theory Construction," but it has gained traction lately. Researchers now argue against conceiving of individual variation as "junk" (Navarro et al. 2006), and instead as a source of information (Kosslyn et al. 2002, van Horn et al. 2008, Kidd et al. 2017, Lonsdorf and Merz 2017, Bolger et al. 2019, Emery and Webster 2019). Kanai and Rees (2011), for example, suggest that more attention should be paid to individual differences in cognitive psychology and neuroscience. They call individual differences research a neglected but "potentially powerful approach" which "can be used as a source of information to link human behaviour and cognition to brain anatomy" (*ibid.*, 231). By studying the neural basis of variation in perception and motor behavior, they argue, scientists can shed light on how brain structures support various cognitive functions. The view of individual differences as an inferential tool may thus be poised to reshape research practice in several different fields.[3]

Hence, there are many reasons that individual differences in cognitive science are deserving of philosophical attention: they present both epistemic challenges and opportunities to practicing scientists; they necessitate methodological innovation and scrutiny; and they have lately garnered increased attention from scientists themselves. In the work that follows, I use philosophical tools to explore how to improve the treatment of variation in cognitive science.

[3] Of course, if you want to correlate, say, performance on a behavioral task with a neurological variable, you need individual differences at both levels. Correlation between two variables only works if there is variation in the values of both. So, even though there is more discussion nowadays of the ways in which variation can be used as an inferential tool, there is a sense in which it has been so used for a long time.

1.3 Dissertation Preview

The dissertation is comprised of four substantive chapters. The first two chapters develop and apply an account of what is needed to explain individual differences. The final two chapters discuss several ways in which scientific methodology should be modified to take account of variation.

In Chapter 2 ("Explaining Individual Differences"), I propose an interventionist account of what it takes to explain variation. On my view (SHRINK), causally explaining individual differences involves exhibiting causes that can be intervened on to reduce population variance in the trait or behavior in question. I show that this account is superior to an intuitive alternative proposal based on Kenneth Waters' (2007) notion of actual difference making, since it can capture the explanatory import of uniform variables. My account also reveals a new way in which the causal background must be kept in view when making claims about populations and suggests a novel means of distinguishing noise from variation.

Chapter 3 ("Hierarchical Bayesian Models: A Promising Approach?") uses SHRINK to examine the causal-explanatory potential of hierarchical Bayesian modeling in psychology. I reject the hyperbolic claims of some proponents of the hierarchical Bayesian approach, arguing that most models in the psychological literature today do not causally explain the variability in psychological parameter values that they capture. Nevertheless, there are ways of extending these models to enhance their explanatory power. A large part of the appeal of hierarchical Bayesian modeling is that it forges a middle path between the Charybdis of ignoring individual differences and the Scylla of treating every individual as unique. I call this the "uniformity/uniqueness dilemma," and suggest that its resolution is a central challenge in

research on individual differences. This chapter shows, however, that one may successfully navigate the dilemma without explaining variation.

Chapter 4 ("Human Variation and Rational Analysis") explores the potential of rational analysis to model psychological variation. Rational analysis is a modeling approach which tries to understand the mind by identifying the problems the mind is attempting to solve and deriving optimal solutions to those problems. Taking research on causal learning as a case study, I argue that practitioners of rational analysis usually ignore psychological variation or describe it in a way that cuts off further inquiry. Nevertheless, there are many ways of accommodating individual differences in rational models. I advocate for an approach I call the "rational analysis *of* variation," which seeks to illuminate how people's variable behavior makes sense given their different experiences, situations, and aims. Rational analysis is not alone in requiring reform: many scientific methods that worked well for producing psychological generalizations may founder when applied to the problem of understanding difference.

In Chapter 5 ("Registration Pluralism and Data Aggregation in Neuroscience"), I show that individual differences are methodologically consequential even when variation is not the object of study. Neuroscientists often aggregate brain data from multiple subjects by mapping the data from individuals onto a common spatial template. I argue that one of the key steps of this "cartographic approach," a process called registration, must be carried out in a way that is sensitive to the target of investigation. Registration aims to align homologous brain regions, but because of variability in brain organization, not all homologous regions can be aligned at once. As a result, no single registration method suffices for all neuroscientific purposes. I call this view "registration pluralism." Registration pluralism recommends several changes to neuroscientific practice and has philosophical implications for the transfer of evidence across epistemic contexts and the theory-ladenness of data aggregation.

In articulating criteria of success for the explanation of variation and examining the limitations of current scientific methods, this work contributes a philosophical perspective to nascent efforts to remedy the neglect of individual differences in cognitive science.

2.0 Explaining Individual Differences

2.1 Variation, Explanation, and #TheDress

In 2015, an overexposed photograph of a striped dress went viral.[4] Some people were certain that the dress had a blue body with black lace stripes, while others were just as sure it had a white body with gold stripes. The public was fascinated that people could have such different perceptions of the same object. Remarkably, researchers working on color perception were "almost equally surprised" by the phenomenon, which sparked a lively discussion on the Color and Vision Network mailing list (Olkkonen and Ekroll 2016, 176). The same year the photograph appeared on Tumblr, there was a trio of commentaries on it in *Current Biology*, and in 2017 the *Journal of Vision* dedicated a special issue to it. To date, the photograph has been the subject of more than 25 research papers (Martín-Moro et al. 2018). Scientists and laypeople alike want to know: why do people see such different things when they look at #TheDress?

The explanandum picked out by this question is different from many other scientific explananda in that it concerns variation rather than regularity. Psychologists often want to know things like what role working memory plays in language, why we are susceptible to the McGurk effect, and how people learn new perceptual categories. These explananda are about regularities across minds rather than differences between them. In this chapter, I'll spell out just what is different about explananda that concern variation and provide a philosophical account of what it takes to explain them. In doing so, I'll draw examples from ongoing research on #TheDress.

[4] The photograph can be seen at http://en.wikipedia.org/wiki/The_dress

Although my account focuses on psychology, its domain of application and philosophical implications are broad. Individual differences are of interest across the sciences, from medicine to education to economics. Especially in applied disciplines, variation can have practical consequences. It is important to know, for example, not just whether certain biomedical treatments or pedagogical interventions are effective on net, but also whether (and why) they have different effects on different individuals. Ignoring such variation can lead to one-size-fits-all approaches that cause harm to patients and students. Variation among other units of analysis matters too: we also want to be able to explain why different solar systems, muscle cells, bees, labor markets, and hurricanes are different from one another. Establishing regularities that hold across all hurricanes, but not understanding why different hurricanes follow different paths and development trajectories, would leave us with an impoverished meteorology.

Despite its importance, the explanation of variation remains philosophically undertheorized. With a few exceptions (Waters 2007, Tabery 2009), philosophers have not examined what it means for something to be a cause of individual differences or what is required to explain observed variation in a population. In this chapter, I'll argue that explaining variation requires identifying factors that could be intervened on to reduce the variability in the population. This account deepens our understanding of explanation and sheds new light on an underexplored dimension of science.

My plan is as follows. In Section 2.2, I delineate the subclass of explananda that I call "variation explananda." To formulate an account of what it takes to explain a variation explanandum, I adopt an interventionist framework, introduced in Section 2.3. I then articulate two interventionist theories of the explanation of individual differences, rejecting the first and endorsing the second. Section 2.4 argues against an account based on Waters' (2007) notion of actual difference making. On my view, presented in Section 2.5, explaining a variation

explanandum involves exhibiting causes of variation, which are variables that could in principle be intervened on to reduce population variance. My account provides criteria of success for the explanation of individual differences and reveals a new way in which the causal background must be kept in view when making claims about populations.

2.2 Variation Explananda

A primary aim of psychology is to understand how the mind works. Understanding the mind requires the formulation of generalizations about its operation that hold across subjects. Much psychological research tries to identify regularities of this kind. There is another dimension of psychological understanding, however, that is equally important: we want to know what differences exist between the minds of individuals and explain why they occur. Psychologists interested in color, for example, want to understand both the general mechanisms underlying color perception and why the photograph of #TheDress produces radically different perceptions in different people.[5]

There are different questions scientists can ask about individual differences. One can ask why variation is patterned in a certain way in the population, as in, "Why are individuals' susceptibilities to the irrelevant speech effect normally distributed (Ellermeier and Zimmer 1997)?" Or one can ask why there isn't *more* variation than there is, as in, "Why does #TheDress elicit just two different responses (Lafer-Sousa and Conway 2017)?" The most basic question to

[5] I'll talk loosely about people "seeing," "perceiving," or "interpreting" the dress differently. I intend to remain neutral about whether the phenomenon is perceptual or cognitive.

ask about variation, though, is why it exists at all: why there is (any) variability in the trait or behavior of interest? Questions that take this form include, "Why are some people faster than others at the Stroop task?" and "Why is language processing lateralized to the right hemisphere in some people and the left in others?" Such questions are the starting point for individual differences research and so will be my focus here.

Questions about the existence of variation pick out what I will call "variation explananda." Let variable *Y* represent a trait or behavior that ranges over individuals in a population *p*. A variation explanandum asserts that *Y* takes different values for (at least some of the) different individuals in *p*. In other words, it asserts that there are two or more values of *Y* represented in the population, rather than one.[6] Variation explananda can be about differences that are continuous (as in the question about reaction times in the Stroop task) or discrete (as in the question about language lateralization).

The relevant population *p* is occasionally obvious, like when we try to explain why there are individual differences among the 3.6 million people who participated in the online Twitter poll about the colors of the dress (Holderness 2015). More often, *p* must be inferred. When we ask why there is variation in how people interpret the colors of the dress, *p* is arguably comprised of all healthy adult internet users. Even when *p* is left implicit, it is an important component of a variation explanandum. There are no individual differences without a population of individuals. In what follows I'll focus on variation explananda that appear to be at the type level rather than the token level: explananda that concern kinds of (potentially recurring) individual differences, not those that are about a specific instance. Most variation explananda of interest in psychology

[6] In what follows, I will assume that a variation explanandum *only* states that there exists variation in *Y*. It does not give any further details, e.g., about other variables that correlate with *Y*.

(and arguably all the non-historical sciences) are type-level, since psychologists care most about explaining general, repeatable phenomena rather than particular events.

Variation explananda incorporate an implied contrast: they state that there is variation *rather than uniformity* in trait or behavior *Y*. This requirement is needed because claims about the presence of variation can be associated with different contrasts, not all of which make variation the explanatory focus. Consider the statement, "People's category representations differ from one another." In asking for an explanation of this, one might be asking, "Why are people's category representations, *but not their perceptual representations*, different?" or "Why do people have different *rather than the same* category representations?"[7] Only the latter targets a variation explanandum. The former is about differences between category and perceptual representations rather than differences between individuals. The use of explanatory contrasts also allows us to distinguish questions about why there is variation at all from other questions one can ask about variation, such as those mentioned above. For instance, "Why does the photograph of #TheDress elicit two different responses?" can be understood as asking, "Why does the photograph elicit two different responses *rather than one*?" or "Why does the photograph elicit *just two* responses *rather than many?*" The latter question also concerns individual differences, but it does not target a variation explanandum, as I define it.

The formulation of variation explananda can be laborious. Consider the explanandum, "People see the dress as either blue and black or white and gold." Color-matching studies were required to establish that people really do have different perceptual experiences when they look at the dress, and that they are not simply using color names differently (Brainard and Hurlbert

[7] This example is inspired by Shen and Palmeri (2016), but note that psychologists have not established that people's category representations differ while their perceptual representations do not.

2015, Chetverikov and Ivanchei 2016). The explanandum also remains controversial because some researchers believe the distribution of color perceptions produced by the dress is continuous rather than discretely bimodal (Gegenfurtner et al. 2015, Hugrass et al. 2017, Aston and Hurlbert 2017). Others think there are not two perceptions but three, since a small number of people report that the dress is blue and brown (Lafer-Sousa and Conway 2017, Jonauskaite et al. 2018).

To recap: variation explananda are a subclass of explananda that state that there is variation rather than uniformity in some trait or behavior type in a specific population. This gives us a grasp on our target. In addition to studying regularities across minds, psychologists are interested in explaining variation explananda. My aim in what follows is to propose a philosophical account of what this requires.

In doing so, I'll use the dress as an extended case study. Although individual differences in perceptions of the dress are especially striking, they have the same features as more mundane patterns of variation of interest to psychologists: they are measurable, predictable differences between people with high intra-individual reliability (Drissi-Daoudi et al. 2020). Assuming that the dress variation is discrete, it differs from the bulk of psychological variation that is (arguably) continuous, but this difference does not matter to the accounts of explanation that will be considered below. One could just as well formulate the discussion to follow in terms of individual differences in speech planning (Swets 2015), risky decision making (Glöckner and Pachur 2012), working memory (Unsworth et al. 2004), visual imagery (Reeder et al. 2017), or any other variable psychological phenomenon. Moreover, as we will see, potential explanations of variation in interpretations of the dress are not at all atypical, appealing to long-established psychological principles (e.g. de Lange et al. 2018). We can therefore safely generalize from an in-depth

examination of the investigation and explanation of variation in #TheDress to psychological variation more broadly.

2.3 Interventionism

In my view, the best current theory of causal explanation is interventionism (Woodward 2003, Woodward and Hitchcock 2003). Since interventionism has been defended elsewhere, I will not discuss its merits here. I will instead ask what the interventionist ought to say about what it takes to explain a variation explanandum, focusing on Woodward's (2003) account as the most thoroughly developed version of the view.

Interventionism conceives of causation as a relation between variables that take different values. Roughly speaking, *X* is a cause of *Y* if and only if *Y* can be changed by intervening on *X* while holding other factors fixed. The notion of an intervention here is critical and somewhat technical. Intervening is like conducting an ideal experiment to determine whether *X* causes *Y*: one sets the value of *X*, severing its connections with factors that usually influence it, in order to see whether the value of *Y* changes as a result. More precisely, an intervention on *X* with respect to *Y* is the taking of a specific value by an intervention variable *I* that acts as a switch for *X*. That is, when *I* takes on (a) particular value(s), the value of *X* is a function of *I* alone. *X* ceases to depend on any other variables. Moreover, *I* only influences *Y* through *X*; it does not exert an independent influence on *Y*. Interventions need not be practically feasible or even physically possible. They need only be logically or conceptually possible and well-defined (Woodward 2003, 127-132).

To generate causal claims from facts about interventions, one needs to decide where to hold fixed the aforementioned "other factors." Specifically, when considering a potential

intervention on X with respect to Y, values must be assigned to the variables not on a causal path from X to Y. (If off-path variables weren't held fixed, then changes to Y could be the result of changes to those variables rather than the intervention on X.) Woodward argues that the proper handling of off-path variables depends on whether we are interested in type or token causation. For type-causal claims, off-path variables can be fixed at any possible value.[8] As long as there is an intervention on X with respect to Y that changes the value of Y under *some possible assignment* of values to off-path variables, then X is a type cause of Y. Second, Woodward suggests that we can capture most of our token causal judgments by setting off-path variables at their actual values.[9] As long as there is an intervention on X with respect to Y that changes the value of Y when off-path variables are set to their *actual values*, then X is a token cause of Y. Part of my aim in what follows will be to show that neither of these proposals is adequate for causal claims about variation.

Colloquially, the interventionist view of causation is often presented in the language of "wiggling": X is a cause of Y iff there's some way of wiggling X that changes Y, with off-path variables held at the appropriate values. Rules for the assignment of values to off-path variables delineate the background conditions against which wiggling X must change Y for X to count as a cause of the relevant kind. (In this respect, limits on the assignment of values to off-path variables within an interventionist framework serve some of the same functions as Mackie's [1974] notion of a "causal field.") Explanations involve the provision of causal information,

---

[8] Here and elsewhere, when I discuss causes or causal claims, I mean to invoke Woodward's notion of a contributing cause. Other causal concepts (e.g., total cause, direct cause) require different treatment.

[9] Woodward concedes that this view doesn't generate the correct judgments in token cases of symmetric overdetermination. To accommodate such cases, he settles on a view that appeals to the notion of the "redundancy range" of a variable (Woodward 2003, 74-86; Hitchcock 2001). Since his initial account is far simpler and captures almost all of our judgments, and since cases of symmetric overdetermination are arguably quite rare, I will stick with the initial proposal that off-path variables be held at their actual values.

understood in interventionist terms. This means they show how the explanandum depends counterfactually on the explanans, where the counterfactuals in question describe the outcomes of interventions. Such counterfactuals are sometimes called "interventionist counterfactuals." Explanations can be used to answer what Woodward dubs "what-if-things-had-been-different" questions because they show us what difference it would have made to the explanandum "if the factors cited in the explanans had been different in various possible ways" (Woodward 2003, 11).

The rest of the chapter will develop an interventionist analysis of the causation and explanation of individual differences. Given the interventionist's commitments, what should she say about what it takes to explain a variation explanandum? The next section will consider and reject a proposal that appeals to Waters' notion of actual difference making. Informed by the shortcomings of this account, I'll then lay out my own view.

2.4 The ACTUAL Account

One potential interventionist account of how to explain a variation explanandum relies on a concept developed by Waters (2007). Waters claims that there is an ontologically special subset of causes called actual difference makers (ADMs) that are responsible for "actual differences" in a population. One might argue that the variables that are explanatorily relevant to variation are ADMs. In this section, I'll argue against this Waters-inspired account.

2.4.1 Introducing ACTUAL

According to Waters (2007), *X* is *an actual difference maker* with respect to *Y* in population *p* if and only if:

(i) *X* causes *Y*.
(ii) The value of *Y* actually varies among individuals in *p*.
(iii) The relationship *X* causes *Y* is invariant over at least parts of the space(s) of values that other variables actually take in *p*.
(iv) Actual variation in the value of *X* partially accounts for the actual variation of *Y* values in population *p* (via the relationship *X* causes *Y*).[10]

Loosely speaking, a variable *X* is an ADM for *Y* in *p* when *X* and *Y* both vary and *X* causally accounts for some of the variation in *Y*. Waters endorses interventionism and suggests that the causal concepts in his definition be explicated using Woodward's (2003) framework. Actual difference making is meant to capture the intuitive distinction between the causes of a trait and the causes of "actual differences" in the trait. For example, consider the actual differences in sun-related skin damage in the human population. The causes of these differences include differences in lifetime sun exposure, skin melanin concentration, and sunscreen usage. Such factors are all ADMs because they vary and are partly responsible for existing variation in skin damage among humans. One factor that is a cause of skin damage but not an ADM is the amount of ultraviolet radiation emitted by the sun. The sun's total UV radiation does not account for any of the variation among humans since it is the same for everyone.

Waters' paper is controversial because of how he uses the concept of actual difference making to reject causal parity theses, which claim that different causes in biology (such as genes and physiological background conditions) are ontologically on a par. Waters argues that

[10] Waters offers this definition of *an* ADM after his definition of *the* ADM, but I am focusing on the former because it is more widely applicable in living systems (as Waters recognizes; p. 571).

biologists only care about ADMs, not potential difference makers, and that genes are the most causally specific ADMs in biology. His aim is to defend gene-centric biology against the critiques of developmental systems theorists and others. Waters' argument has been roundly criticized: for claiming that biologists don't care about potential difference makers (Stegmann 2012, Currie 2018, Baxter 2019), for ignoring ADMs that aren't genes (Griffiths and Stotz 2013), for characterizing the distinction between ADMs and all other causes as ontological (Northcott 2009), and for caricaturing the views of those who endorse causal parity theses (Griffiths and Stotz 2013). These critiques have undermined Waters' conclusions about biology but have not challenged his definition of an ADM.

Actual difference making may seem to be just the tool we need to formulate an interventionist account of the explanation of variation explananda. The causes that Waters singles out as ADMs are those variables that account for individual differences in a population. One might argue, then, that ADMs are what explain variation:

ACTUAL The variables that explain a variation explanandum are actual difference maker(s) X with respect to Y in population p.

This account has prima facie plausibility. We do frequently explain variation by tracing differences in the trait or behavior of interest back to differences in some other property of individuals. Explanations that seem to appeal to ADMs abound in psychology, and #TheDress provides a ready illustration. One factor that strongly influences the perceived colors of the dress is how subjects interpret the illumination in the photograph (Chetverikov and Ivanchei 2016, Wallisch 2017, Witzel et al. 2017b, Toscani et al. 2017).[11] People who interpret the image as

[11] Although I'll talk about people "interpreting" or "making assumptions about" illumination, the process is almost certainly subpersonal. It is more accurate to speak of the illumination assumptions made "by the visual system," but for the sake of readability, I will stick with the personal level formulation.

showing the dress in shadow (or under blueish illumination) tend to see it as white and gold, while those who take it to be directly lit (by yellowish illumination) typically see it as blue and black. Wallisch (2017) explains why this is so: "Shadows overrepresent short wavelengths. In other words, shadows appear bluish. If someone assumes that the dress was in a shadow, color constancy mechanisms could be expected to discount the effect of the shadow, rendering the conscious appearance of the dress more yellowish" (5). Different subjects make different assumptions about illumination, leading color constancy mechanisms to "correct" the perception in different ways. Brainard and Hurlbert (2015) call this the "color constancy explanation" of individual differences in the perceived color of the dress. It is a proximal and partial explanation, to be sure, since it does not explain why people make different assumptions about illumination, but it does provide some explanatory information.[12]

Subjects' assumptions about illumination (X) seem to be an ADM with respect to the perceived color of the dress (Y). We can think of population p as including everyone who saw the dress when it circulated on the internet in 2015. Survey results have established that different individuals in p made different assumptions about the illumination in the photograph (Wallisch 2017), satisfying Waters' condition (ii) on actual difference making. There is strong experimental evidence that these assumptions causally influenced people's perceptions of the dress, satisfying (i) and (iii). Witzel et al. (2017b), for example, manipulate subjects' interpretations of the illumination by substituting realistic scenes which are either clearly sunny or clearly in shadow to serve as the background behind the dress. They find that when subjects are exposed to the

[12] Witzel et al. (2017b) and Metzger and Drewing (2019) argue that the dress provides an example of the cognitive penetration of perception: people's high-level beliefs about illumination shape their perception of color. This is another reason the dress is philosophically interesting (though I am personally skeptical of their claim).

image of the dress against an obviously sunny background, they tend to see the dress as blue and black; when the background is in shadow, they tend to see it as white and gold. Together with the survey results, such experiments suggest that variation in assumptions about illumination partially account for variation in perceived dress color, satisfying Waters' condition (iv). Hence, people's illumination assumptions, which partly explain individual differences in perceived dress color, appear to fulfill Waters' criteria for being an ADM, lending credence to ACTUAL.

2.4.2 The Explanatory Role of Uniform Variables

ACTUAL is ultimately unsatisfactory, however, because there are factors that are uniform in the target population that help explain variation. ADMs necessarily vary across subjects. Equating the explanation of individual differences with the exhibiting of ADMs therefore threatens to overlook an important class of explanatory variables.

For a variable to be an ADM, it must take different values for different individuals in population *p*. This is because of Waters' condition (iv), which states that variation in *X* partially accounts for variation in *Y* in *p*. Waters fleshes out the condition as follows: "*X* partially accounts for the actual variation of *Y* values in population *p* (via the relationship *X* causes *Y*) if and only if conditions (i)–(iii) [in the definition of an ADM] are satisfied and an intervention on *X* with respect to *Y* that changed the *X* values in one or more individuals in *p* to the *X* value that one of the individuals had sans intervention would change *Y* values in *p*" (Waters 2007, 571). Let's call this kind of intervention a "swapping intervention," since it involves swapping one individual's *X* value for the *X* value of a different individual in *p*. Condition (iv) is satisfied iff there is a swapping intervention on *X* that would change the value of *Y* for at least one individual in *p*. There can only be such an intervention if at least some of the individuals in *p* have different *X*

values. Waters embraces this feature of his account, emphasizing that no variable that "exhibit[s] uniform values in the actual population" can be an ADM for that population (*ibid.*, 570).[13]

The problem with ACTUAL is that there are variables that help explain individual differences but do not vary in *p*. Consider the following example. Many researchers studying #TheDress have noticed that the pixel chromaticities in the photograph differ primarily along the blue-yellow axis (Brainard and Hurlbert 2015). This means that the chromaticities closely mirror variation in natural daylight, tracking the so-called "daylight locus." Research has shown that this chromaticity profile is essential to the photograph's capacity to generate different perceptions in different people. Gegenfurtner et al. (2015) shift the RGB values in the photograph around the hue circle away from the daylight locus, making blueish pixels become pinkish and yellowish pixels become greenish. This adjustment makes the individual differences effectively disappear, with subjects reporting that the dress body is pink or red, and the lace green. It is not yet known why the photograph's blue-yellow chromaticity profile produces such stark individual differences. Building on work showing that people's color discrimination is poorest along the daylight locus (Pearce et al. 2014), Brainard and Hurlbert (2015) speculate that "there might be something especially ambiguous about images whose RGB values vary primarily in this way…[S]uch images might evoke a response shaped more than usual by implicit prior expectations about illumination spectra, which in turn might vary across individuals" (R553).

[13] My term "swapping intervention" may be slightly misleading, since it might suggest an *exchange* of two individuals' *X* values (a "swap" being an interaction that involves two people). On the reading I intend, however, only one individual's *X* value must be modified in a swapping intervention. Note also that, instead of the interpretation I suggest here, one could read Waters' condition (iv) as requiring that there be a swapping intervention on *X* that would change the overall *distribution* of *Y* values in *p* rather than the *Y* value of at least one individual in *p*. It is not clear which interpretation of (iv) Waters intends. Nevertheless, both readings imply that ADMs cannot be uniform, which is all that my argument here requires.

Follow-up work investigating this hypothesis has produced mixed results (Wallisch 2017, Witzel et al. 2017a, Witzel et al. 2017b, Aston and Hurlbert 2017, Wallisch and Karlovich 2019).

The chromaticity profile of the original photograph is not an ADM, but it does help to explain variation in subjects' perceptions of the dress. It isn't an ADM because it is the same for all subjects. Nevertheless, it determines whether other variables affect people's perceptions of the dress. Subjects' prior expectations about illumination would be causally irrelevant to the perceived color of the dress if the chromaticity profile of the photograph were different. Furthermore, as Gegenfurtner et al.'s (2015) experiment shows, there is an intervention on chromaticity profile that eliminates variation in the perceived color of the dress. Thus, appealing to the chromaticity profile helps to explain the individual differences even though it does not tell the whole story. To fully understand the variation, we would need to determine how chromaticity profile interacts with other (varying) factors to produce different interpretations of the dress. But we shouldn't get hung up on what is required for a complete explanation of variation, since science rarely traffics in complete explanations. The fact remains that the chromaticity profile of the dress does genuine explanatory work.[14]

Variables whose values are uniform across subjects therefore play an important role in explanations of individual differences. A Waters-inspired account of how to explain variation is missing something important. Contra ACTUAL, ADMs are not the only variables that contribute

[14] A defender of ACTUAL might try to argue that chromaticity profile is an ADM after all. What separates the photograph of #TheDress from photographs of other dresses – what makes its color unusually ambiguous – is partly its chromaticity profile. So chromaticity profile is an ADM with respect to the population of dress photographs. This may be true, but it shifts the goalposts. The explanatory questions about the dress that we have been considering ask why different *people* see this specific photograph differently. The population *p* is a group of humans, not dress photographs. Chromaticity profile is therefore not an ADM with respect to variation explananda about differences in a human population, though it may be an ADM with respect to differences between photographs.

to the explanation of variation explananda.[15] This objection not only refutes ACTUAL, but also undercuts the general intuition that difference must be explained by appeal to difference. One might have thought that, to explain why people's behavior varies, we must point to other ways in which those people differ, with the differences in the explanans being causally responsible for the behavioral differences in the explanandum. Although many causal explanations of variation do indeed have this structure, the above example shows that, perhaps surprisingly, some do not. As a result, no view that requires the presence of variability in the explanans variables will be adequate as an account of the causal explanation of individual differences.

2.5 The SHRINK Account

Waters' account of actual difference making is the primary extension of interventionism in the literature that directly concerns variation. The failure of ACTUAL, then, forces us back to the drawing board. In this section, I'll present my own positive account of how to explain a variation explanandum and lay out its virtues.

2.5.1 Introducing SHRINK

The idea behind SHRINK is that explaining a variation explanandum requires explaining why the population variance of Y is nonzero. This is accomplished by describing (an)

[15] In my view, there are additional issues with Waters' characterization of actual difference making. Even if the objection against ACTUAL raised here were unsuccessful, the account would still be inadequate because of technical problems inherited from Waters' definition of an ADM. See Appendix A for a summary of these problems.

intervention(s) on some variable X with respect to Y that would reduce (or "shrink") Var(Y). I call X a cause of variation in Y. The full account, which I'll explain below, is as follows:

SHRINK A variable X is a cause of variation in Y in population p iff there is a possible intervention on X with respect to Y that would reduce Var(Y) in p and in which the actual distribution of values of variables not on a path from X to Y in p is held fixed.

To explain a variation explanandum, one must exhibit a true causal generalization G relating change(s) in a cause of variation X with change(s) in Y that are associated with a reduction in Var(Y).[16]

SHRINK captures the explanatory import of uniform variables, avoiding the problem that plagued ACTUAL. Causes of variation, as defined by SHRINK, can be the same or different across individuals in p. For example, the chromaticity profile of the dress photograph qualifies as a cause of variation under SHRINK because there is an intervention on chromaticity profile with respect to perception of the dress that reduces the variance in the latter, when the distribution of off-path variables is held fixed (Gegenfurtner et al. 2015). Off-path variables in this case include things like viewing conditions (lighting, screen type, etc.) and people's prior expectations about illumination spectra. SHRINK therefore implies that the photograph's chromaticity profile helps to explain individual differences in how the dress is perceived. As I argued in the previous section, this is the correct verdict.

Several features of SHRINK require further elaboration. First, SHRINK is an account of the explanation of variation that applies in deterministic contexts (though a probabilistic extension, invoking the probability distribution over Var(Y), would be easy to formulate). The intervention on X with respect to Y that it requires must be an intervention for at least one of the individuals

[16] Population variance, written Var(Y) or σ^2, captures how much the values of Y differ from their mean $\bar{y}$ in a population with n members. It is standardly defined as follows:

$$\sigma^2 = \Sigma(y_i - \bar{y})^2 / n$$

in *p*. The account also assumes that there is causal homogeneity in the target population, meaning that the individuals in *p* conform to the same causal model. This may seem like an unwarranted assumption, but it is not as strong as it appears. First, it is a commonplace assumption in modeling and inference. Many statistical techniques require assumed homogeneity of some kind among the units. Second, individuals can differ from one another and still conform to the same causal model. They can have different values for any of the variables that range over individuals. Third, even if the individuals in *p* fall into (say) two qualitatively distinct groups, each associated with an entirely different constellation of causal relationships, they can all be modeled as conforming to the same model provided there is a variable in the model representing group membership.

There is also a question about what justifies the asymmetry at the heart of SHRINK. Why do we only count as causes of variation those variables that can be intervened on to *reduce* population variance? The characterization of variation explananda offered in Section 2.2 provides the answer. Recall that a variation explanandum comes with a contrast: it states that there is variation *rather than uniformity* in *Y*. The interventionist counterfactuals exhibited to explain any explanandum must be responsive to its contrast, in the sense that they must describe counterfactual outcomes that approach the contrast. This means an explanation of a variation explanandum has to show how changes in the explanans variable *X* would have produced greater uniformity in *Y*. Variation explananda are not explained by interventionist counterfactuals in which Var(*Y*) increases because they do not contrast the presence of variation with uniformity.

SHRINK's emphasis on the reduction of variance does not imply that the only evidence for causal claims about variation comes from experiments that reduce Var(*Y*). One must distinguish between the interventionist counterfactuals that make a causal claim true and the counterfactuals about experimental outcomes that support it. The latter may not mirror the former for practical reasons. For *X* to be a cause of variation under SHRINK, there must be an intervention on *X* with

respect to Y that reduces Var(Y). However, under certain circumstances, such as when intervening on X to reduce Var(Y) is not feasible, an experimental manipulation of X that brings about an *increase* in Var(Y) may be taken as evidence that X is a cause of variation in Y. Imagine that psychologists found that interfering with subjects' attention increased variance in perceptions of #TheDress. Under certain assumptions about attention, this might be taken to support the idea that attention is a cause of variation. The (sometime) inferential significance of interventions that increase Var(Y) is not a threat to the asymmetry in SHRINK, given that SHRINK is an account of the interventionist counterfactuals that make causal claims about variation true.

2.5.2 Off-Path Variables and the Causes of Variation

Another feature of SHRINK that requires explanation is its handling of off-path variables. Recall that, when considering an intervention on X with respect to Y, one must decide where to hold fixed the variables not on a path from X to Y. Woodward (2003) proposes two strategies, suitable for type- and token-causal claims, respectively: (i) set off-path variables at any possible combination of values, or (ii) set off-path variables at their actual values. In contrast, my account, focusing on causal claims about variation in a population, proposes to (iii) set off-path variables at values that maintain the *actual distribution* of values in the population. Unlike (ii), which holds fixed the properties of individuals, (iii) holds fixed population-level features. SHRINK claims that a variable X is a cause of variation in p iff there is an intervention on X with respect to Y that would reduce Var(Y) in which the actual distribution of values of variables not on a path from X to Y in p is held fixed. This means that there must be some way of wiggling X that reduces the variance in Y, where that wiggling is done against a causal backdrop in which off-path variables remain unchanged at the population level.

The intuitive motivation for this approach is that it guarantees that causal claims about individual differences in p reflect the characteristics of p. As was noted in Section 2.2, all claims about variation are restricted to a specific population. This population-relativity can be captured by assigning values to off-path variables in a way that preserves the actual distribution of values in the target population. X is reasonably called a cause of variation in Y in p iff wiggling X brings about the requisite changes in Y when other background conditions in p are kept constant.

Although it may be appropriate to adopt strategies (i) and (ii) for the handling of off-path variables when evaluating typical type- and token-causal claims, neither strategy is suitable for causal claims about variation. First, consider what would happen if we adopted (i), that is, if a cause of variation were any variable that could be wiggled to reduce Var(Y) against *any possible causal background* (i.e., with off-path variables set to any of their possible values). This proposal makes specification of the population largely inert, thereby failing to capture many intuitive judgments about the causes of variation in a population. Imagine that screen brightness (B) and screen type (T) are potential causes of subjects' color perceptions of #TheDress (D). Variable T can take one of two values: T=1 for conventional screens, and T=2 for brand new super screens. These super screens, we may imagine, eliminate color ambiguity when turned up to maximal brightness. We are interested in the causes of variation in perceptions of the dress among current adult humans, so what matters is whether wiggling B causes a reduction in Var(D). Let's stipulate that manipulating brightness has no effect on a subject's perception of the dress when she is using a conventional screen. Since super screens are not yet in commercial circulation, all of the individuals in p view the dress photograph on a conventional screen. However, if everyone were using super screens, it would be possible to increase screen brightness and in so doing cause everyone to see the dress in the same way.

I suggest that in this scenario, brightness is intuitively not a cause of variation in perceptions of the dress among current adults, since every individual in *p* is using a conventional screen. SHRINK delivers this verdict. To assess whether *B* causes variation in *D*, we hold fixed the actual distribution of *T* values in *p* and find that there is no way to wiggle *B* that reduces Var(*D*). So *B* is not a cause of variation in *D* in this population. However, had we adopted strategy (i) for handling off-path variables, we would have gotten the counterintuitive verdict that screen brightness *is* a cause of variation in perceptions of the dress. Under (i), we can set off-path variables to any possible values. If we set *T*=2, intervening on *B* reduces Var(*D*) to zero. So, *B* would count as a cause of variation in *D*. This example shows that adopting strategy (i) renders the population (and the values of variables which characterize it) largely irrelevant to the formulation of causal claims about variation. This is unacceptable: the causes of variation in a trait or behavior *Y* should reflect the characteristics of *p*.[17]

Defining causes of variation in the manner described by (i) therefore does not do justice to the population-restricted nature of causal claims about variation. What about strategy (ii), in which off-path variables are set to their actual values? Had we adopted (ii) instead of (iii), we would count *X* as a cause of variation in *Y* iff wiggling *X* reduces Var(*Y*) while keeping all off-path variables at their actual values. While this generates the correct verdict in the toy example just discussed, it is needlessly restrictive. It requires that the values of off-path variables for every individual in *p* be fixed at their actual values. But the presence of individual differences is a

[17] Note the important difference between the case of chromaticity and this toy example: the former concerned whether chromaticity, a uniform factor in the target population, can *itself* help explain variation in perceptions of the dress. (I argued that it can.) In this case, we're asking whether screen brightness can explain the variation given that the causal *background conditions* involve a variable, screen type, that is uniform in *p*. (I am suggesting that it can't.) The uniformity is relevant in the latter case but not the former because in the latter case it is an off-path variable that is uniform, while in the former it is the potential cause itself.

population-level, not individual-level, phenomenon. To ask for an explanation of variation is to seek information about a group rather than any of its individual members. Causes of variation are factors that have a certain kind of influence on the population; the impact they have on specific individuals does not matter in itself. What matters in the assignment of values to off-path variables, then, is that the causal background conditions in the population are held constant, not that every single individual remains the same.

For these reasons, extant strategies for handling off-path variables are not suitable for the definition of a cause of variation. SHRINK incorporates an alternative strategy, one that allows some flexibility in the assignment of values to off-path variables but ensures that the assigned values are representative of the target population. It is worth reflecting briefly on what this modification means for interventionism generally. Interventionists have shown that strategies (i) and (ii) successfully capture most of our causal judgments about paradigm type- and token-causal claims like "smoking causes cancer" or "the striking of the match caused it to light." However, I have argued that causal claims about variation require (iii), the holding fixed of off-path variables' distribution of values in the population. There is little reason to think that claims about variation are unique in requiring a more nuanced approach to the assignment of values to off-path variables. For example, I suspect that (iii) is apt for other causal claims that are restricted to a population, such as, "For American schoolteachers, smoking reduces stress." This claim seems to be true iff intervening on smoking lowers stress when the actual distribution of off-path variables characterizing the population of American teachers is held fixed. Let's imagine that wiggling smoking does indeed lower stress under these conditions. It is irrelevant that this might not be the case if, contrary to fact, American teachers were not underpaid and overworked. We must hold fixed the distribution of values of off-path variables to see whether smoking has an impact given the actual background conditions among American teachers.

Moreover, the fixing of a population-level distribution of values is not the only alternative way of handling off-path variables that is needed. Consider the claim, "Smoking causes cancer even when cigarettes do not contain tar." This is true iff there is an intervention on smoking with respect to cancer that changes the latter when an off-path variable representing the presence of tar in cigarettes is held fixed at the value for "no tar." If the only interventions on smoking that change cancer occur against a causal background in which tar is present, then the above claim is false. This suggests that assigning values to off-path variables must be sensitive to the content of the causal claim under consideration. Strategies (i) and (ii) may be adequate for the paradigmatic causal claims that have animated philosophical discussions about causation, but analyzing more complicated causal claims requires a subtler approach. Moreover, we cannot simply add strategy (iii) to our interventionist toolbox and then be done with it. The tar example shows that the way one assigns values to off-path variables must reflect the fine-grained content of the claim in question, shattering the hope that there might be a short, exhaustive list of strategies with which to analyze all causal claims.

Investigating the explanation of variation has therefore drawn our attention to a more general issue: the distinction between type and token causation cannot fully guide the assignment of values to off-path variables within the interventionist approach. Assigning values to off-path variables must be done in a way that reflects the content of the causal claim under consideration.

2.5.3 SHRINK and Scientific Practice

Let's now examine what can be said in favor of SHRINK, beyond the fact that it captures intuitive judgments about toy cases. As mentioned above, SHRINK, unlike ACTUAL, correctly implies that uniform variables like the chromaticity profile of the dress can help to explain

variation explananda. It also reflects the experimental logic of research on variation. Much scientific work on individual differences involves the same basic procedure: variation is identified, a hypothesis about its source is formulated, and an experiment is performed in which the hypothesized cause is manipulated with the aim of reducing the variation. If the manipulation succeeds, the causal hypothesis is favorably assessed. Research on the color constancy explanation conducted by Hesslinger and Carbon (2016) follows this procedure. Their Experiment 1 involves scrambling the image of the dress (i.e., dividing the image up into squares and randomly mixing them up) to give subjects less information about the illumination in the scene. They reason as follows: "If the color constancy explanation…is right, that is if interindividual differences in the use of illumination information significantly contribute to the differences in the perception of the dress, these differences will decrease when the illumination information of the image is reduced…by means of image scrambling" (*ibid.*, 3). They find that, indeed, the more scrambled the image, the less variation in subjects' color reports about the dress. They take this to support the hypothesis that divergent assumptions about illumination cause individual differences in perceptions of the (original) photograph.

SHRINK captures Hesslinger and Carbon's line of thought and the logic of their experiment. To assess whether assumptions about illumination are a cause of the variation in subjects' color judgments, Hesslinger and Carbon try to find an intervention on illumination assumptions (X) with respect to color judgment (Y) that reduces Var(Y), while keeping fixed the actual distribution of values of other variables. The authors' motivation for using image scrambling is that it disrupts illumination cues while "preserv[ing] the local color information of the image" (*ibid.*, 4). Image scrambling is an appropriate intervention, they claim, because it influences the variance in color judgments only through its effect on assumptions about illumination.

Experimental design is often subject to criticism, and SHRINK captures the logic of these challenges as well. Jonauskaite et al. (2018) cast doubt upon the evidential value of Hesslinger and Carbon's experiment by drawing attention to another potential cause that is affected by image scrambling. They point out that divergent interpretations of the texture of the dress could be responsible for variation in perceptions of its color. Since image scrambling reduces information about texture as well as illumination, the experiment does not provide strong support for the color constancy explanation relative to a texture-based explanation. Filtered through SHRINK, Jonauskaite and colleagues are arguing that the original authors fail to establish that illumination assumptions are a cause of the variation because image scrambling is not an intervention on illumination assumptions with respect to color judgments. A genuine intervention on X with respect to Y must not influence causes of Y that are not on a path from X to Y. Image scrambling changes textural information, an off-path cause of color judgment, so it is not an intervention of the requisite kind.

One might object to SHRINK on the grounds that it categorizes too many variables as causes of variation. Variables that exert an extremely minimal influence on Var(Y) count as explanatory under SHRINK, as do counterfactuals like, "If there had been a meteor strike that killed all life on earth, there would be no variation in perceptions of the dress." For a more realistic example, consider Hugrass et al.'s (2017) work showing that subjects' perceptions of the dress can be changed by inducing illusory shifts in the brightness of the photograph. They produce illusory dimming and brightening using an established technique in which rotating sawtooth gratings are superimposed on an image. When the photograph of the dress seems to be dimming, subjects are more likely to see it as blue and black; when it seems to be brightening, responses tend toward white and gold. Consider a continuous variable S that represents illusory brightness shift in the image of the dress ($S > 0$ for brightening, $S < 0$ for dimming). Assuming there is an illusory

brightness shift that would reduce the variance in perceptions of the dress, *S* counts as a cause of variation under SHRINK. This may seem strange. It doesn't seem explanatory to say that people perceive the colors of the dress differently in ordinary contexts because the apparent brightness of the photograph is not changing.

There are resources in the literature to counter this objection, which is a species of the broader worry that interventionism counts too many things as causes. First, interventionists are not committed to all explanations being equally informative. Deeper explanations "give answers to a finer-grained, more detailed, and wider-ranging set of what-if-things-had-been-different questions" (Woodward 2003, 222). The apparent brightness explanation does not answer a wide range of w-questions, which is partly why its explanatory power is limited. Second, we can invoke the notion of a "serious possibility" to rule out outlandish explanations. Woodward argues that, "if a change in a purported explanans is associated with some corresponding change in an explanandum, but the change in the explanans is not a serious possibility, then the information that the explanandum will change under this change in the explanans is typically not regarded as explanatory, or at least the purported explanation is not seen as satisfying or relevant" (*ibid.,* 227).[18] The absence of a meteor strike does not explain the presence of variation in perceptions of the dress because it is not a serious possibility. Similarly, the possibility that sawtooth gratings could be applied to the dress photograph is not one that we take seriously outside Hugrass and colleagues' laboratory. Finally, SHRINK can be defended by appealing to the philosophical literature on causal selection (Hesslow 1988). Even though SHRINK does indeed count many variables as causes of variation, there are strategies for selecting the most important

[18] Against the charge that the notion of a "serious possibility" renders interventionism overly subjective, Woodward argues that alternative accounts of causation require an analogous concept (*ibid.,* 89).

or explanatory causes from among the full set. Criteria that have been proposed for causal selection include causal strength, invariance, proportionality, and specificity (Hitchcock and Woodward 2003, Woodward 2010, Ross forthcoming).

2.5.4 SHRINK on Noise

One attractive upshot of SHRINK is that it suggests a principled way of distinguishing between noise and variation. Researchers frequently distinguish individual differences that are unstructured, uninteresting, or random, from differences that are systematic, interesting, or theoretically important. The former are called "noise" and the latter, "variation" (e.g. Finn et al. 2017, Seghier and Price 2018). Different definitions of noise have been offered in the scientific (Faisal et al. 2008), statistical (Upton and Cook 2008), and philosophical (McAllister 1997) literatures. One common theme is that noise, unlike variation, is fundamentally inimical to explanation. SHRINK can be used to make this idea precise. Noise, on this view, is the heterogeneity between individuals (or any units of analysis) that cannot be explained given the variables one is willing to countenance. When there are no causes of variation (as defined by SHRINK) within one's potential variable set, one is dealing with noise. No variable can be intervened on in the appropriate way to reduce the population variance. Genuine variation, meanwhile, is heterogeneity that is amenable to explanation in the manner laid out in SHRINK.

This way of cashing out the distinction leans heavily on the notion of a variable that one is willing to include in one's variable set. As a result, heterogeneity that counts as noise for one researcher (e.g., a cognitive psychologist who is interested in variables at the cognitive level) may count as variation for a different researcher (e.g., an electrophysiologist who deals with variables representing properties of neurons). There may be no variable that the cognitive psychologist

would entertain that could (in principle) be intervened on to reduce the heterogeneity, even though such a variable does exist for the electrophysiologist. While some might object to this relativity to variable set, in my view, scientific context does indeed partly determine whether heterogeneity qualifies as noise or variation (Cronbach 1957). Moreover, although the relevant variable set is determined by one's interests, once that set is determined, it is an objective matter whether the differences in question constitute noise or variation. Either there is a variable in the set that meets the conditions in SHRINK or there is not. Defending this approach to understanding noise would require a thorough comparison with alternative accounts. The suggestion itself, however, illustrates the applicability of SHRINK to topics beyond causation and explanation.[19]

2.6 Conclusion

I have argued that explananda about variation between minds constitute a distinct and undertheorized class of psychological explananda. Drawing from research on #TheDress, I explored what the interventionist ought to say about how to explain a variation explanandum. One plausible proposal, ACTUAL, based on Waters' (2007) well-known work on causal selection, was rejected. Explaining individual differences in some trait or behavior does not exclusively

[19] It has been pointed out to me that SHRINK could be used to define noise differently. One could argue that noise is heterogeneity that can't be reduced by intervening on any of the variables in *any possible* variable set. I am not tempted by this alternative account of noise, in part because there are infinitely many possible variable sets (since variables are cheap and can be recombined in countless ways). Moreover, I suspect that most of the heterogeneity in the world could be reduced by intervention on *some* variable, so very little of it would end up counting as noise. This runs counter to scientists' acknowledgment of noise in almost every domain of research. Thus, to the extent that we want our account of noise to be practically applicable, and to line up somewhat with scientific usage, the account of noise proposed above is preferable to the alternative considered here.

involve actual difference makers because explanatorily relevant variables may be uniform in the target population. Instead, I argued, explaining a variation explanandum requires exhibiting causes of variation, which are variables that can be intervened on to reduce population variance. This account, SHRINK, captures the explanatory import of uniform variables, reflects experimental practice, and suggests an appealing way of drawing the distinction between noise and variation.[20]

Although my focus has been on the explanation of individual differences in psychology, SHRINK can be applied in other domains as well. Explaining variation in a variable *Y* that ranges over any unit of analysis arguably involves identifying factors that could be intervened on to reduce variance in the way described by SHRINK.[21] Hence, with its broad applicability, my account enriches our philosophical understanding of explanation – not just in psychology, but across the sciences – by characterizing what is distinctive about explaining variation rather than regularity. It also highlights the need for subtlety in the application of interventionism. I have argued that we must attend to the background conditions in the population when analyzing causal claims about individual differences. Attention to variation therefore promises not only to shed light on an underexplored area of cognitive science, but also to contribute to the refinement of philosophical accounts of science.

---

[20] Since I have limited my discussion to causal explanation, there may be more to say about the non-causal explanation of variation explananda. Perhaps there are statistical, mathematical, or constitutive explanations of individual differences (Lange 2013a, 2013b; Yilkoski 2013; Lyon 2014). I suspect that non-causal explanations in science are less common than is usually supposed but cannot argue for this here.

[21] Woodward has a (1995) paper in which he argues that linear regression equations "typically explain actual variation in some specific population," assigning larger coefficients to factors that are of greater causal significance to variability in the dependent variable (46). Woodward's interpretation of linear regression models is consistent with (though not entailed by) SHRINK. An independent variable assigned a nonzero coefficient in a regression can be manipulated to reduce the variance in the dependent variable, so long as certain assumptions (e.g., about the direction of causation, the variable's manipulability) hold. If Woodward is right that regression equations pick out causes of variation in a population, then the scope of my account is very wide indeed, since regression is used throughout the sciences.

3.0 Hierarchical Bayesian Models: A Promising Approach?

3.1 The Uniformity/Uniqueness Dilemma

As interest in individual differences intensifies, cognitive science is increasingly faced with a dilemma: we want our models to neither assume that everyone is the same, nor treat every individual as entirely unique. Models that impose uniformity on the population can elide significant, theoretically meaningful differences between individuals. But modeling every individual completely independently is incompatible with scientific generalization and does not do justice to what we have in common. To acknowledge variation in a productive way, one must stake out an intermediate position between the poles of this uniformity/uniqueness dilemma.

Psychologists have developed a number of strategies for navigating the dilemma. Sometimes they divide the population into known or latent groups and model each group separately while assuming within-group homogeneity. In other contexts, they distinguish between different psychological "strategies" in use by different people at different times. Perhaps most commonly, they assume that individuals conform to the same mathematical model but possess different values for the model's psychological parameters. Bayesian modelers have developed a sophisticated form of this third approach, which includes the use of what I will call stochastic parameter models. In stochastic parameter models, individual parameter values differ but are all drawn from (typically the same) overarching distributions.

In this chapter, I will examine what is required for stochastic parameter models to causally explain psychological variation. I'll use SHRINK to reject the hyperbolic claims of some proponents of these models, arguing that at present, most models do not causally explain the variation in

parameter values that they capture. Still, there are ways of extending stochastic parameter models to enhance their explanatory power. The broader lesson is that escaping the uniformity/uniqueness dilemma is not the same thing as explaining variation. Some solutions to the dilemma are merely modeling devices. Overall, then, the chapter sounds a cautionary note: not all tools that help us represent individual differences also explain them.

In Section 3.2, I introduce hierarchical Bayesian models, or HBMs, the family of models which includes stochastic parameter models. I present Shiffrin et al.'s (2008) simple model of memory retention to illustrate the application of the hierarchical Bayesian approach to cognitive psychology. Section 3.3 clarifies the chapter's central question: can stochastic parameter models causally explain variation in individual subjects' parameter values? I argue in Section 3.4 that the causal-explanatory power of most current models depends on how we interpret their hyper-parameters. I entertain two possible interpretations, showing that neither permits most extant stochastic parameter models to satisfy SHRINK's criteria on the causal explanation of variation. The criteria can be fulfilled, however, if stochastic parameter models are extended, as discussed in Section 3.5. Section 3.6 shows that even HBMs that fail to explain variation may nevertheless have non-explanatory virtues. Finally, in Section 3.7 I discuss what this examination of HBMs tells us in general about modeling individual differences.

3.2 Hierarchical Bayesian Modeling in Cognitive Psychology

Hierarchical Bayesian modeling, which is becoming increasingly popular in cognitive psychology, is not Bayesian in the sense most familiar to philosophers. When philosophers think about Bayesianism in cognitive science, they typically think of models that assume that the mind

performs Bayesian inference (Chater et al. 2006, Griffiths et al. 2008). On this approach, one takes the mind to be updating a set of priors on the basis of new evidence in accordance with Bayes' theorem. Some interpret Bayesian models of cognition quite literally, taking them to describe cognitive processes occurring in the brain, while others argue that they provide non-mechanistic, rational analyses of behavior (Eberhardt and Danks 2011, Lee 2011; see Chapter 4). There is a second way of applying Bayesianism to psychological modeling, however, that simply involves using Bayesian methods for data analysis and model selection (Lee 2008). On this approach, the calculation of posteriors and Bayes factors is used to choose among competing models and estimate model parameters.[22]

Hierarchical Bayesian modeling is Bayesian in the second sense: it uses Bayesian estimation techniques to fit and assess hierarchical models. Although researchers in different fields define hierarchical models in different ways (Raudenbush and Bryk 2002, Gill and Womack 2013), modelers in cognitive psychology typically take them to be "models in which some parameters are partly determined (e.g. chosen from distributions defined by) other parameters" (Shiffrin et al. 2008, 1261).[23] Hierarchical models are best understood by contrast with non-

[22] Lee (2008, 2018) argues that there are three (rather than two) ways to apply Bayesianism to the study of the mind. He draws a distinction between Bayesian data analysis and Bayesian cognitive modeling, arguing that analyzing a model with Bayesian tools is not "just data analysis" because adopting priors amounts to changing the model itself. But he ultimately concedes that "modeling can be considered an elaborate form of data analysis" (Lee 2018, 38). I have accordingly collapsed over these two uses of Bayesianism in the paragraph above.

[23] Models with hierarchical structure were pioneered in education research in the 1980s, later adopted by clinical and social psychologists, and only recently introduced to cognitive psychology (Rouder et al. 2005). On the history of hierarchical modeling, Gill and Womack (2013) write, "it is often the case that fundamental ideas in statistics hide for a while in some applied area before scholars realize that these are generalizable and broadly applicable principles…hierarchical modeling follows this same storyline, roughly originating with the statistical analysis of agricultural data around the 1950s. A big step forward came in the 1980s when education researchers realized that their data fit this structure perfectly" (5). It took more time still for it to reach cognitive psychology.

hierarchical models. Consider the pair of generalized graphical models in Figure 3.1. Figure 3.1a shows a non-hierarchical model in which a set of parameters θ determine data y through function $f(\cdot)$. Most models in cognitive psychology today take this simple, non-hierarchical form. In a hierarchical model, by contrast, the first-order parameters θ are not taken as basic. Instead, the model captures how the parameters are generated.[24] In the hierarchical model in Figure 3.1b, the "hyper-parameters" ψ have been added upstream of θ. The parameters θ are generated from ψ by process $g(\cdot)$. Hierarchical models are also called "multilevel" or "nested" models because the model describing the relationship between ψ and θ is nested within the model describing the relationship between θ and y (Myung and Pitt 1997, Stephan et al. 2009).

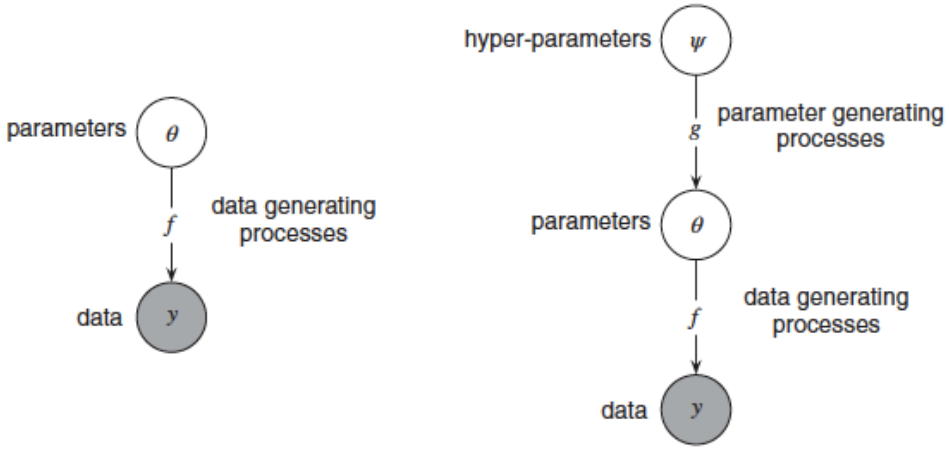


Figure 3.1. Non-hierarchical (a) and hierarchical (b) model structures (Lee 2018)

There are two families of techniques for estimating hierarchical models: maximum likelihood methods (Goldstein 2013) and Bayesian methods (Fahrmeir et al. 2013). When

[24] Some prefer to speak of parameter *values* being generated, rather than parameters themselves, but I will follow proponents of the hierarchical Bayesian approach in adopting the terminology of parameter generation.

hierarchical models first came into widespread use in education research, only maximum likelihood methods were available. Because it is impossible to compute the posterior distribution analytically for models of reasonable complexity, Bayesian estimation was of little use. The applicability of Bayesian estimation expanded dramatically with the invention of Markov Chain Monte Carlo (MCMC) methods, which allow one to simulate samples from the posterior distribution and hence approximate posteriors for complex models (Gill and Womack 2013). Perhaps the most popular MCMC algorithm among psychologists is Gibbs sampling. Gibbs sampling starts by randomly selecting initial values for the parameters and hyper-parameters. Then, for every iteration of the simulation, each first-order parameter and hyper-parameter is successively updated based on the data that has been collected and one's priors. This process repeats for thousands of iterations. Under many conditions and with enough time, the parameter estimates that one collects come to approximate the posterior distribution over the parameters and hyper-parameters (Gelman and Hill 2006, ch. 18). Hierarchical Bayesian modeling applies such techniques to hierarchical models. As Lee (2018) explains, "there is nothing inherently Bayesian" about hierarchical models, but the attraction of Bayesian estimation methods is that they "work in exactly the same way" for hierarchical and non-hierarchical structures (41).

Advocates of hierarchical Bayesian modeling in cognitive psychology argue that it has a variety of advantages over non-hierarchical, non-Bayesian approaches (Lee 2011). One of these advantages concerns the ability of HBMs to deal with individual differences. According to hierarchical Bayesian modelers, variation is poorly handled by non-hierarchical models. But, by adding hierarchical structure to a model and fitting it with Bayesian techniques, one can "reveal and account for individual differences in cognition" (Bartlema et al. 2014, 144). For example, Shiffrin et al. (2008) argue that their hierarchical model of memory retention, to be presented below, permits one to simultaneously model the parameters controlling memory retention and

the parameters controlling variation. Thus, according to its proponents, the hierarchical Bayesian approach "constitutes a deeper level of psychological theorizing, because it not only allows for individual differences, but imposes a model structure on those differences, and allows inference about parameters – like the group mean and variance – that characterize the individual differences" (Lee 2011, 3).

Following Myung et al. (2000), I use the term "stochastic parameter model" to pick out a type of HBM in which the first-order parameter values for each subject are randomly drawn from higher-order distributions. The model in Figure 3.1b would be a stochastic parameter model if the value of θ for each individual were sampled from a population-level distribution of θ values whose form was characterized the hyper-parameters ψ. The other main subtype of HBM, which I will not discuss, is what is sometimes called a "group model." Group models distinguish between two or more groups of individuals, either latent or observed, each with its own set of first-order parameter values.[25] I am putting group models to the side here in large part because recent enthusiasm for HBMs in cognitive psychology revolves around stochastic parameter models, not group models, which have been around for a longer time.[26]

Let us now take an in-depth look at a stochastic parameter model from Shiffrin et al. (2008) and Lee and Wagenmakers (2013). This example will serve to illustrate how typical stochastic parameter models in psychology represent variation, so I'll return to it several times throughout the chapter. Figure 3.2 presents a simplified model of memory retention which attempts to answer

[25] In addition, there are hierarchical models that combine the properties of both stochastic parameter and group models: multiple groups are postulated, each with its own group-level distributions from which the first-order parameter values for individuals belonging to that group are drawn (Bartlema et al. 2014).

[26] Evaluating the explanatory contributions of stochastic parameter models and group models requires very different analyses. Whether group models explain the variation they capture depends on the contentious question of which groups or categories are explanatory (e.g. Glymour and Glymour 2014). Stochastic parameter models do not raise such thorny issues.

a question that has interested cognitive psychologists for over a century: what is the relationship between time and the number of items remembered from a list (Ebbinghaus 1913)? As mentioned in Chapter 1, many psychologists now believe that this relationship can be captured by an exponential decay function. One can construct a model that takes the probability θ_t that an item will be remembered after time t has elapsed to be a function of α, a parameter representing the rate of decay of information, and β, representing a baseline level of remembering:

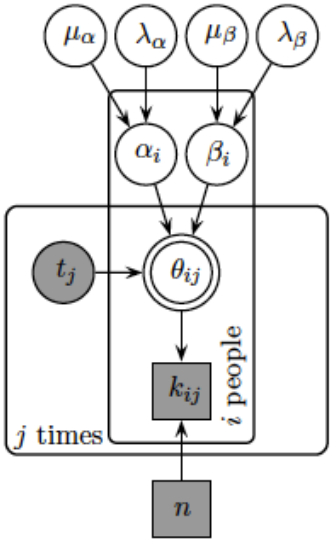
$$\theta_t = \exp(-\alpha t) + \beta$$

The (fictitious) data that Shiffrin et al. (2008) use to fit this model describe how many items from a list subjects are able to remember during a recall experiment. There are multiple data points for each subject after different amounts of time have elapsed. Before building a hierarchical model, the authors consider two non-hierarchical models. They first fit the exponential decay model to the data without incorporating individual differences; that is, they assume that all subjects have the same α and β values. The authors then fit a model in which each subject is given her own α and β values without constraint. Both of these non-hierarchical models are argued to be deficient. The first, "no individual differences model" is not descriptively adequate: the posterior predictive distribution does not correctly describe the behavior of several of the participants.[27] The second, "full individual differences model" fails to make informed predictions about new subjects and overfits the data by capturing noise (see Section 3.6).

A third, hierarchical Bayesian implementation of the model has neither of these faults. Each participant is allowed different α and β values, but these first-order parameters are taken to

[27] The posterior predictive distribution shows what data the updated model predicts. It is "a distribution over data, and gives the relative probability of different observable outcomes after data have been seen" (Lee and Wagenmakers 2013, 45). If the model assigns a low probability to many of the outcomes that have been observed, it is said to be descriptively inadequate.

be sampled from overarching Gaussian distributions characterized by hyper-parameters. This model, represented in Figure 3.2, is said to involve "*structured* individual differences" (*ibid.*, 1269; my italics).[28] The hyper-parameters μ_α, λ_α, μ_β, and λ_β are the same for all participants and describe the Gaussian distributions from which each subject's α and β values are drawn (μ = mean, λ = precision). The parameters α and β are placed inside the "i people" plate to indicate that they have different values for each person i. Parameters α and β determine the probability θ_{ij} that person i will remember an item at time j. This probability θ_{ij} determines whether or not an item is remembered by i at j, represented by k_{ij}.[29]



$$\mu_\alpha \sim \text{Beta}(1,1)$$
$$\lambda_\alpha \sim \text{Gamma}(.001,.001)$$
$$\mu_\beta \sim \text{Beta}(1,1)$$
$$\lambda_\beta \sim \text{Gamma}(.001,.001)$$
$$\alpha_i \sim \text{Gaussian}(\mu_\alpha, \lambda_\alpha)_{\mathcal{I}(0,1)}$$
$$\beta_i \sim \text{Gaussian}(\mu_\beta, \lambda_\beta)_{\mathcal{I}(0,1)}$$
$$\theta_{ij} \leftarrow \min\left(1, \exp\left(-\alpha_i t_j\right) + \beta_i\right)$$
$$k_{ij} \sim \text{Binomial}(\theta_{ij}, n)$$

Figure 3.2. Hierarchical exponential decay model of memory retention (Lee and Wagenmakers 2013)

---

[28] These authors follow standard conventions for graphical models in Figure 3.2: nodes represent variables and arrows represent dependencies. Circular nodes correspond to continuous variables and square nodes to discrete variables. Observed variables are shaded and unobserved variables are unshaded. Stochastic variables are represented by single-bordered nodes and deterministic variables by double-bordered nodes. The "plates" indicate replication.

[29] Following model comparison, Shiffrin and colleagues actually conclude that β values do *not* differ across subjects, whereas α values do. They then simplify the model, removing the hierarchical structure from the β parameter. I'll ignore this complication here, since the model in Figure 3.2 contains structured individual differences in both α and β.

Shiffrin and colleagues show that their hierarchical memory retention model succeeds in capturing the data of each subject and making informed predictions about the behavior of new people. Hence, in this analysis, "a theory of memory retention [is] combined with a theory of individual differences, to provide a more complete account of behavioral data from multiple subjects" (Lee and Wagenmakers 2013, 156).

3.3 The (Causal) Explanation Question: SHRINK, Applied

Many hierarchical Bayesian modelers have explanatory ambitions. Lee (2011), for instance, argues that HBMs "allow for the more complete *theoretical explanation* of data from a single task, letting different people use different cognitive processes, or letting the same people use different processes at different times" (6; my italics). Lee and Wagenmakers (2013) claim that, although machine learning approaches often produce more accurate predictions, the advantage of HBMs is that they "provid[e] details about the *underlying processes* assumed to produce the data" (95; my italics). This suggests that unlike machine learning models, which are merely predictive, HBMs are sometimes able to genuinely explain the data. Vandekerckhove (2014), too, claims that incorporating random effects in hierarchical models, as stochastic parameter models do, "can be used in the construction of models whose aim is to *explain* observed variance" (60; my italics). This is especially noteworthy given that Vandekerckhove demonstrates a sensitivity to the distinction between explanation and mere description (*ibid.*, 59).[30]

[30] There is also an instrumentalist contingent of hierarchical Bayesian modelers who shy away from explanation. Rouder et al. (2005) distinguish between statistical modeling and substantive modeling. They

Assessing whether stochastic parameter models causally explain variation is the central question of the remainder of the chapter. Let us now sharpen this question by specifying the explananda at issue and applying SHRINK's requirements for causal explanation to HBMs. In line with the previous chapter, the present inquiry will examine the ability of stochastic parameter models to causally explain variation explananda, i.e., claims to the effect that there is variation rather than uniformity in psychological trait or behavior Y in population p. Since we are considering hierarchical models, Y may stand for a variable at the level of behavior or at the level of the first-order parameters. In the memory retention model (Fig. 3.2), for instance, the explanatory target might be variation in k (the number of items remembered) or θ (the probability that a person will remember an item at a certain time), both of which are behavioral variables, or it might be variation in α (the decay rate) or β (the baseline level of remembering), the psychological parameters. Hence, there are two types of variation explananda we might be interested in explaining with a stochastic parameter model: explananda about why individuals' behavior differs (behavioral variation), or explananda about why individuals have different first-order parameter values (parametric variation).

In what follows, I focus on variation explananda concerning parametric variation. Hierarchical and non-hierarchical models are essentially on a par when it comes to causally explaining individual differences in behavior. With both kinds of models, behavioral variation is explained by appeal to the first-order parameters that generate the behavior. In both generalized graphical models in Figure 3.1, that is, parameters θ can be used to explain variation in data y. For example, in a social psychology experiment, we might explain variation in behavior y by

---

claim that, "[Statistical models] are used for estimation and inference, whereas [substantive models] are tested as truthful models of phenomena. We view our hierarchical Weibull model as best used in the spirit of the former" (*ibid.*, 197).

appeal to differences in a parameter θ representing one's degree of extraversion. The success of this causal explanation does not depend on whether there is an additional level in our psychological model, that is, on whether θ is conceived as the result of some process $g(\cdot)$. Hence, hierarchical and non-hierarchical models are equally capable of causally explaining variation in the behavior that their first-order parameters control.

If stochastic parameter models constitute an explanatory improvement over their non-hierarchical counterparts, then, it must be because they provide a new way of explaining individual differences in the first-order parameters. Non-hierarchical models cannot causally explain parametric variation because they take psychological parameters as basic. Recall from Figure 3.1a that there are no variables upstream of first-order parameters θ in non-hierarchical structures. By contrast, in Figure 3.1b, θ is modeled in terms of generating process $g(\cdot)$ and hyper-parameters ψ, suggesting that hierarchical models may have additional explanatory resources. Our central question, now sharpened, is this: can stochastic parameter models causally explain variation explananda concerning differences in individuals' first-order parameter values?

SHRINK, introduced in Section 2.5, gives us the tools to answer this question. Recall that SHRINK holds that causally explaining a variation explanandum about Y requires identifying (an) intervention(s) on a variable X that would reduce $\text{Var}(Y)$ in p. (This informal gloss, which neglects the complications about off-path variables discussed in Section 2.5.2, suffices for the purposes of this chapter.) Let us assume that the relevant population p is the population from which subjects have been drawn. Adopting the notation from Figure 3.1, our explanatory target Y is the psychological parameter(s) θ. SHRINK implies that, to explain a variation explanandum about individual differences in θ, an HBM must contain a variable that could be intervened on to reduce $\text{Var}(\theta)$ in the subject population. Our task is to determine when stochastic parameter models contain such causes of variation.

Doing so requires a clear understanding of the notion of an intervention. As discussed in Section 2.3, intervening is like conducting an ideal experiment to determine whether X causes Y: one exogenously sets the value of X to see whether the value of Y changes as a result. Woodward provides formal necessary and sufficient conditions on the notion of an intervention variable that can be given the following colloquial interpretation (Woodward 2003, 98). A manipulation constitutes an intervention on X with respect to Y iff:

C1. The manipulation causes a change in X.
C2. The manipulation detaches X from everything that usually causes it.
C3. The manipulation only affects Y through X. It doesn't directly affect Y or any causes of Y that aren't downstream of X.
C4. The manipulation is independent of any cause of Y that isn't X or downstream of X.

A manipulation of X with respect to Y that meets these four conditions counts as an intervention even if it is not practically feasible. An intervention must be "logically possible and well-defined," but need not be physically possible (*ibid.*, 128; see Section 3.4.3).

The following two sections will assess stochastic parameter models' causal-explanatory potential using SHRINK. Shiffrin et al.'s (2008) exponential decay model of memory retention will be used as an example of the most widespread sort of stochastic parameter model. Recall that one of the parameters in the memory retention model is α, representing the rate of decay of information. The question to be asked next is whether the memory retention model features a variable that could be intervened on (in principle) to reduce Var(α). Finding the answer to be "no," I'll then consider how such models might be modified so that they do display causes of variation. My argument will show that although most stochastic parameter models at present do not causally explain parametric variation (Section 3.4), they can be supplemented to satisfy SHRINK's criteria (Section 3.5).

3.4 Hyper-Parameters in Simple Stochastic Parameter Models

Most of the stochastic parameter models currently in use in psychology have a fairly simple structure: the hyper-parameters characterize population-level distributions over the first-order parameters, taking the same value for all subjects. They are not given any interpretation apart from their role in defining the distributions over the parameters. Moreover, the hyper-parameters are the only variables in the uppermost level of the model; there is nothing else upstream of the first-order parameters. Shiffrin et al.'s (2008) memory retention model has these features. Hyper-parameters μ_α, λ_α, μ_β, and λ_β are the only variables upstream of parameters α and β. They define the distributions over those parameters but are otherwise uninterpreted. In this section, then, the memory retention model will be used as a stand-in for the simple stochastic parameter models that currently dominate hierarchical Bayesian modeling in psychology.

At first glance, it might seem that models like the memory retention model easily explain parametric variation. Individual i's decay rate α_i is sampled from a higher-level Gaussian distribution with mean μ_α and precision λ_α. By intervening to increase the value of λ_α, one would thereby narrow the distribution of sampled values of α, reducing Var(α). It is therefore tempting to think that simple stochastic parameter models do satisfy SHRINK: they seem to contain hyper-parameters that can be manipulated to bring about a reduction in the population variance in the first-order parameters. I will now argue that this initial appearance is mistaken. The reasoning above is correct in focusing on hyper-parameters, since they are the only candidates for causes of parametric variation in simple stochastic parameter models. However, one cannot assume that it is possible to intervene on the hyper-parameters with respect to the parameters (e.g., intervene on λ_α with respect to α). I will entertain two possible interpretations of hyper-parameters in simple stochastic parameter models, arguing that neither supports the notion of an intervention

with respect to the parameters. Under both interpretations, then, these models do not causally explain parametric variation.

3.4.1 The Puzzle

Interpreting the hyper-parameters in simple stochastic parameter models is no easy task because modelers send mixed signals about what they are and how they relate to first-order parameters. As we have seen, it is common to think about parameter values as being "sampled" or "drawn" from population-level distributions. Hierarchical Bayesian modelers also frequently talk about "generation," as in Figure 3.1b: the parameters are said to be "generated" from the hyper-parameters, sometimes via a "parameter generating process" (Lee 2011, 2018). But it is not clear whether this generating process is internal to the model, so to speak, or in the world. Some remarks suggest the latter, such as Lee's (2018) claim that in a hierarchical model, "the basic model parameters…are themselves generated by a *psychological process*" (39; my italics). On the other hand, the different levels in an HBM are sometimes said to have different degrees of "abstraction." Lee (2011) claims that process *f* in Figure 3.1 depends on "a more abstract process *g*" (3). Abstracting is an activity of the modeler, not a process by which one psychological variable influences another.

The different ways of understanding the inter-level relationships in stochastic parameter models are visible in Shiffrin et al.'s (2008) characterization of the memory retention model. They claim, "just as the data have been assumed to be *generated by* the decay and baseline parameters combining in a memory process for individual subjects, the hierarchical model assumes that those parameters themselves are *generated by* more *abstract* latent [hyper-]parameters that describe group distributions across subjects" (*ibid.*, 154; my italics). Talk of "generation" and "abstraction"

is intermingled here. The authors also argue that their model allows "both the parameters controlling memory retention for individuals, and the parameters controlling individual differences [to] be considered simultaneously" (*ibid.,* 1280). Such claims again raise the question of whether the relevant notion of "control" is the control of the modeler or the control of one property by another.

Understanding the explanatory power of HBMs requires that we answer these interpretive questions. In my view, there are two plausible interpretations of the hyper-parameters in simple stochastic parameter models: we can take hyper-parameters to be (i) group-level descriptive summaries of the properties of individuals or (ii) basket variables that aggregate unnamed causes of the properties of individuals. The first interpretation implies that the connection between the parameters and hyper-parameters is model-internal, while the second holds that it represents real-world processes. Let's now take a closer look at each interpretation.

3.4.2 Hyper-Parameters as Summary Variables

The first interpretation takes the hyper-parameters in simple stochastic parameter models to be population-level summaries of individual parameter values. On this way of thinking, hyper-parameters provide an aggregate representation of the first-order parameter values belonging to individuals in the population. In the memory retention model, for instance, which assumes that α values are normally distributed, μ_α represents the average decay rate in the population while λ_α conveys how similar people's decay rates are. The first-order parameters are the focus of the model, while the hyper-parameters contribute a convenient summary of their values. They do not have an existence independent of the parameters. On this interpretation, then, the parameters and hyper-parameters are connected by a process of aggregation occurring within the model.

The interpretation of hyper-parameters as summary variables is in line with the claim that hyper-parameters "describe" or "capture" variation in the first-order parameters, as in Shiffrin and colleagues' suggestion that the hyper-parameters in the memory retention model "*describe* group distributions across subjects" (*ibid.*, 154; my italics). It is also compatible with the idea that the levels of an HBM are characterized by increasing degrees of "abstraction." Summarizing the properties of individuals involves abstracting away from the individuals themselves. A population-level summary variable is therefore more abstract than the individual-level parameters it summarizes.

Does this interpretation of hyper-parameters allow simple stochastic parameter models to causally explain parametric variation? As we have seen, the answer depends on whether one can intervene on the hyper-parameters to reduce the population variance in the first-order parameters. However, interpreting hyper-parameters as summary variables makes it impossible to intervene on the hyper-parameters with respect to the parameters *at all*. Any manipulation of the hyper-parameters will violate condition C3 on interventions, which holds that an intervention on X with respect to Y must only influence Y through X, and not directly affect Y itself. There is no way of manipulating a population-level summary of an individual property without manipulating that very same individual property. Any change to the hyper-parameters will also involve a change in the first-order parameters. In the memory retention model, there is no possible intervention on λ_α with respect to α because, to change the value of a variable that describes the dispersion of individual decay rates, one must manipulate the decay rates

themselves. One cannot change a summary variable without changing that which it summarizes.[31]

Intuitive examples help to show why condition C3 is an important requirement on interventions, and hence indirectly on causal explanation. If C3 were to be relaxed, then manipulations of X which directly and independently alter Y (as well as X) could count as interventions on X with respect to Y. Consider, for example, the relationship between the smell in one's kitchen (X) and the number of eggs in one's fridge (Y). Without C3, frying an egg would qualify as an intervention on X with respect to Y, supporting the causal claim that the smell in one's kitchen is a cause of how many eggs one has in the fridge. C3 blocks this absurd conclusion: frying an egg directly affects both the smell in one's kitchen *and* one's supply of eggs, so it isn't a genuine intervention on X with respect to Y. Relinquishing C3 would also lead to the misclassification of many non-causal relationships as causal: a student's overall GPA could count as a cause of her individual grades; a dog's barking could count as a cause of its making noise; and being an unmarried man could count as a cause of being a bachelor. Hence, interventionists have good reason to insist that interventions satisfy C3.

Given the current interpretation of hyper-parameters as summary variables, interventions on the hyper-parameters with respect to the parameters in a simple stochastic parameter model

[31] Note that I am not denying that there can be causal-explanatory relationships between population-level variables and individual-level variables. In my view, interventionism allows that causes can be at a higher "level" than their effects. For example, there can be psychological or behavioral causes of neural events, such as when intervening on psychological habits using cognitive-behavioral therapy affects brain activity (Porto et al. 2009). Similarly, a population-level variable representing some aspect of societal organization might be a cause of individual behavior. I am not denying the existence of such causal relationships. Instead, I am claiming that population-level variables cannot be causes of the very individual-level variables that they summarize.

are impossible.[32] SHRINK's criteria are not satisfied, so such models describe but do not causally explain parametric variation.

3.4.3 Hyper-Parameters as Basket Variables

To render simple stochastic parameter models causally explanatory, one would need to salvage an interpretation on which the hyper-parameters are causally connected to the first-order parameters. A strategy for doing so takes inspiration from Woodward's (2003) proposed understanding of the error term in linear regression models. Woodward argues that, to interpret a regression causally, one must take the error term to "represent causes of the dependent variable…that are unmeasured and not explicitly represented" in the regression equation (*ibid.*, 328). That is, we should understand the error term as "standing for something real and independent (omitted causes)" (*ibid.,* 325). Let us call variables that stand for the effect of a collection of unspecified causes, "basket variables." Basket variables represent the influence of one or more "unknown direct causes" on another variable (*ibid.*, 43).

The second interpretation of hyper-parameters takes them to be basket variables: a hyper-parameter represents the impact of a bundle of unspecified causes on the first-order parameters. In the memory retention model, hyper-parameters μ_α and λ_α stand for collections of factors that influence a subject's decay rate α. An individual with an α value of μ_α falls exactly in the middle of the population distribution. It is natural, then, to think that μ_α stands for an "average" set of

[32] Strictly speaking, intervention on a hyper-parameter with respect to the parameter *which it summarizes* is impossible. Some HBMs have pairs of hyper-parameters and parameters which are unconnected, so it is possible to intervene on one with respect to another (e.g. λ_α and β in the memory retention model). The pairs that are relevant to causal explanation, however, are those which are linked.

causes of α. Were these causes the only ones operative, everyone would have the same α value. Conversely, λ_α is a basket variable representing the influence of unspecified causes of dispersion in α, the factors that make individuals' decay rates different from one another.[33] Thus, simple stochastic parameter models decompose the causal influences on first-order parameters into a "typical" basket of causes and an "idiosyncratic" basket of causes. Under this interpretation, it might again seem possible to intervene on the hyper-parameters with respect to the parameters to reduce the variance in the latter. It appears that one could reduce Var(α) by manipulating λ_α, the collection of factors that cause subjects' α values to differ. Such an intervention would satisfy C3, avoiding the problem discussed above, because one can manipulate factors that make a difference to individuals' parameter values without directly changing the parameters themselves.

Despite appearances, this interpretation too does not permit simple stochastic parameter models to causally explain parametric variation. Briefly, my argument is this: because basket variables are placeholders, there is no coherent notion of what it would be to intervene on them. So, hyper-parameters understood as basket variables cannot be causes of variation.

Woodward (2016) distinguishes between two different notions of an intervention. For a "possibility constrained intervention" on X it must be conceptually and mathematically possible to intervene on that which is represented by X. For a "setting intervention," it need only be possible to intervene on the value of X. The notion of a setting intervention is weaker than that of a possibility constrained intervention because there are some variables whose values can be manipulated even though the changes represented by those manipulations are not coherent.

[33] The variable λ_α, representing precision, is the inverse of variance, σ_α. It may be more intuitive to say that σ_α represents the causes of dispersion. Of course, from a mathematical point of view, it doesn't matter whether a model is specified in terms of λ_α or σ_α.

Woodward offers an example: "Suppose that we introduce a variable 'animal' which takes the values {lizard, kitten, raven}. By construction, this variable has more than one value," and so is amenable to *setting* interventions. He continues: "but if, as seems plausible, we have no coherent idea of what it is to change a raven into lizard or kitten, there will be no [possibility constrained] intervention for this variable" (*ibid.*, 39). Woodward holds that such variables are not causes. He argues that interventionists should adopt the possibility constrained notion of an intervention rather than the setting notion in part because statements like "ravenness causes blackness" are not genuine causal claims. Woodward also suggests that adopting the more stringent notion of an intervention captures the view, shared by many physicists and philosophers of physics, that causal concepts do not apply in parts of fundamental physics.

I contend that manipulations of basket variables are akin to setting interventions. The value of a basket variable can be changed, but the actual manipulation that such a change represents is not intelligible. Just as the value of the variable 'animal' can be changed from 'lizard' to 'kitten' without our having a coherent idea of what it is to change a lizard into a kitten, so too can the value of a basket variable be altered despite there being "no coherent description" of that intervention in actuality (*ibid.*, 17). This is because, in representing the effect of an unspecified collection of causal factors, basket variables lack determinate content. They are essentially placeholders. Hence, although a setting intervention on a basket variable is possible, a possibility constrained intervention on a basket variable is not.[34]

---

[34] As an exegetical matter, it is not clear whether Woodward would endorse my reading of what is required for a possibility constrained intervention. In his (2003) book he states, "for something to be a cause we must be able to say what it would be like to change or manipulate it" (112). There he calls possibility constrained interventions "well-defined interventions" and claims that, for an intervention on X to be well-defined, we must have a "coherent conception" of what it is to change X (*ibid.*, 13). On the reading I am pursuing, "coherent" here means roughly "understandable," such that possibility constrained interventions must correspond to conceivable changes in the world. However, one might instead interpret "coherent" to just mean something like "consistent," i.e. not contradicted by mathematical or conceptual constraints. On this

On the version of interventionism which incorporates the latter, more robust notion of an intervention, hyper-parameters construed as basket variables cannot be causes. Although we can modify the values of hyper-parameters, we do not have a coherent idea of the manipulations that correspond to such modifications. Assume that λ_α is a basket variable representing the influence of an indeterminate set of causes of dispersion in α. We can change the value of λ_α, but we do not know enough about the causes of dispersion to conceive of changing that which λ_α represents. On the current interpretation, then, there is no possibility constrained intervention on λ_α that would reduce the population variance in α. The stochastic parameter models in which such hyper-parameters are embedded therefore do not causally explain parametric variation.

One might object that interventionism should adopt the notion of a setting intervention rather than the more demanding notion of a possibility constrained intervention. In my view, though, the latter is deeply rooted in an interventionist understanding of what causal explanation is in the first place. For the interventionist, to explain something causally is to show how it would be different were other factors made different. In this respect, causal explanation is intrinsically linked to acting in the world: causal explanations provide information about what one would have to do to bring about alternative states of affairs. Possibility constrained interventions generate interventionist counterfactuals that serve this goal. They describe worldly changes one could (in principle) make to produce different outcomes. By contrast, interventionist counterfactuals about setting interventions may not be action-guiding, since they may describe manipulations that are not coherent. For instance, we have no clear conception of what it would be to satisfy the antecedent of the counterfactual, "if an animal were changed from a raven into a

---

more conservative reading, it might indeed be possible to perform a possibility constrained intervention on a basket variable.

kitten, then it would purr." Such counterfactuals therefore do not serve the aim, central to causal explanation, of providing information relevant to manipulation and control. (There is undoubtedly much more to be said about why the interventionist should adopt the notion of a possibility constrained intervention. For several other arguments against a version of interventionism based on setting interventions, see Woodward [2016].)

3.4.4 Parametric Variation, Unexplained?

This section has proposed two interpretations of hyper-parameters in simple stochastic parameter models. I have argued that, if we interpret hyper-parameters as summary variables or basket variables, manipulations of hyper-parameters fail to qualify as genuine interventions. I did not discuss the merits of either interpretation because my aim was not to adjudicate between them but to draw out their common implication. Of course, I have not argued that the two interpretations are exhaustive. There may be another way of understanding hyper-parameters that makes them amenable to intervention. However, the burden now lies with the hierarchical Bayesian modeler to formulate such an alternative interpretation.

Nowadays most stochastic parameter models take the simple form illustrated by Shiffrin et al.'s memory retention model. The only variables in their uppermost tier are hyper-parameters which characterize distributions over the first-order parameters. The conclusion that such hyper-parameters cannot causally explain parametric variation therefore amounts to a pessimistic appraisal of the causal-explanatory purchase of present-day stochastic parameter models.

This chapter has inherited the previous chapter's exclusive focus on causal explanation. My analysis may overlook ways in which existing stochastic parameter models explain parametric variation non-causally. Given that there are now many types of non-causal

explanation enumerated in the philosophical literature, I cannot rule out that simple stochastic parameter models instantiate one of them. But it is notable that some of the leading forms of non-causal explanation are non-starters. For instance, considering the statistical character of HBMs, one might suspect that they provide mathematical or statistical explanations of parametric variation. A standard way of conceiving of mathematical or statistical explanations takes them to "sho[w] the explanandum to be more necessary…than ordinary causal laws could render it" (Lange 2013b, 485). Simple stochastic parameter models certainly do not show that the existence of individual differences in first-order parameter values is necessary in any strong modal sense. (It was not mathematically necessary for humans in memory experiments to have different decay rates.) Unificationist explanations, which operate by exhibiting common patterns in disparate phenomena, may also be classified as non-causal (Kitcher 1989). The clearest way in which stochastic parameter models unify different instances of parametric variation is purely mathematical: in each case, individual differences are modeled as draws from a population-level distribution with a common (e.g. Gaussian) form. Yet, assimilating disparate phenomena to a common mathematical framework does not provide the kind of substantive unification that is presumably required for unificationist explanation.

At least at first glance, then, simple stochastic parameter models do not conform to two popular models of non-causal explanation. This is not the final word, since there are other proposed species of non-causal explanation. It is ultimately an open question whether more could be said to defend the (non-causal) explanatory power of simple stochastic parameter models.

3.5 Extending Stochastic Parameter Models

Most simple stochastic parameter models draw individuals' first-order parameter values from distributions entirely defined by their hyper-parameters. Yet this simple structure is not an inevitability. It is often possible to supplement simple stochastic parameter models so that their distributions over first-order parameters are partly characterized by additional variables *which can (in principle) be intervened on*. Such extended models include factors that can be legitimately understood as causes of parametric variation.

I am aware of very few stochastic parameter models in psychology that have been so extended. One extended model by Nunez et al. (2015) can be seen in Figure 3.3. Nunez and colleagues propose a hierarchical diffusion model which captures reaction time and accuracy data in a speeded perceptual decision-making task. There are three first-order parameters in their model: drift rate δ_j, representing subject j's evidence accumulation during his decision process; diffusion coefficient ζ_j, representing the amount of variability in evidence accumulation in one trial; and non-decision time τ_j, representing the amount of time it takes for everything in the subject's response process except decision making.

A simple stochastic parameter model would assume that individual δ, ζ, and τ values are randomly sampled from population-level distributions. But Nunez and colleagues supplement the highest level of their hierarchical model with an independent measure of visual attention, steady-state visual evoked potentials (SSVEPs). SSVEPs, which are measured by EEG, are enhanced when a stimulus is attended. The measured SSVEPs for subject j are represented by x_j in Figure 3.3. Instead of sampling drift rates from a common distribution, each subject's drift rate

δ_j is drawn from a normal distribution with mean $\alpha_{(\delta)k} + x_j\gamma_{(\delta)}$ and variance η.[35] Note that this distribution is individual-specific, since its mean depends in part on x_j. When the authors fit the model to the data, they find that variation in drift rates was "partially explained by individual differences in noise suppression as measured by SSVEPs. Participants who better suppressed noise at high frequencies during both the preparatory period (noise interval) and the decision period (response interval) were able to accumulate correct evidence faster, which led to more accurate, faster response times" (*ibid.*, 11). Thus, the model appears to explain individual differences in the speed of evidence accumulation across subjects.

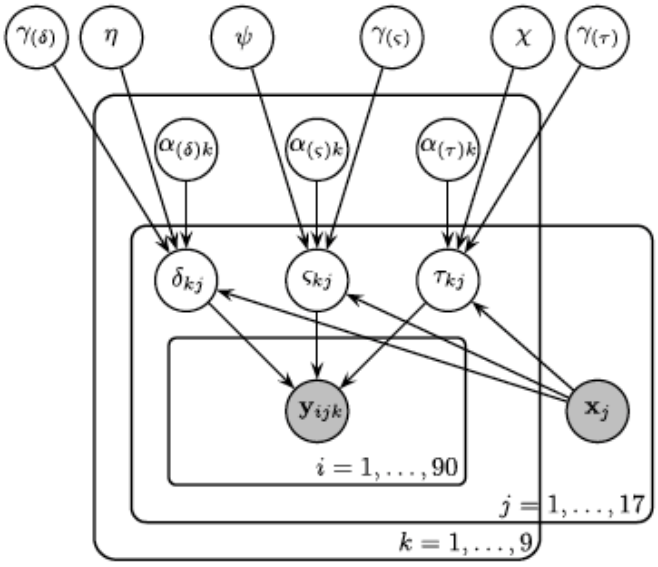


FIGURE 5 | Graphical representation of Model 3. Drift rates δ_{jk}, diffusion coefficients ς_{jk}, and non-decision times τ_{jk} were assumed to vary over both conditions and participants. Each of these parameters are assumed to be drawn from normal distributions with means of the form $\alpha_k + \boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\gamma}$, where $\boldsymbol{x}_j$ is the vector of SSVEP responses of subject j, and with variances that did not vary across conditions. As an example, $\alpha_{(\tau)k}$ is the condition effect on the non-decision time and $\gamma_{(\tau)}$ reflects the change in non-decision time (seconds) due to a one SSVEP unit difference across two participants.

Figure 3.3. Extended stochastic parameter model (Nunez et al. 2015)

[35] I focus here on drift rates, but similar conclusions hold for the other first-order parameters in the model.

SHRINK confirms that the model does indeed provide a causal explanation of parametric variation in the drift rate. The drift rate parameter δ is controlled not just by a suite of shared hyper-parameters ($\gamma_{(\delta)}$, η, and $\alpha_{(\delta)k}$), but also by x_j, a variable representing individual attention. Unlike the hyper-parameters, it is possible to intervene on x with respect to δ. There is a potential intervention that would set individuals' visual attention at the same level and thereby reduce the population variance in the drift rate. So, the model shows that differences in the extent to which subjects suppress noise during the preparatory and decision periods of the experiment partially explain differences in their drift rates.

It is often possible to add variables that represent causes of parametric variation to a stochastic parameter model. One way of doing so involves finding a (manipulable) individual covariate of the first-order parameter, as Nunez and colleagues did. The memory retention model could perhaps be extended in this way by incorporating an independent measure of individual motivation m_i. This extended model, patterned after Nunez et al.'s, would hold that individual α values are drawn from a normal distribution whose mean is partly determined by m_i, with greater motivation leading to lower decay rates via a leftward-shifted distribution of α values. The model might imply that intervening on subjects' motivation (e.g., by changing the monetary payout for accurate performance) would reduce the variance in their decay rates. Were the model empirically supported, it would partially explain parametric variation in decay rates: information decays for different people at different speeds because they have different degrees of motivation.

A second type of model extension involves incorporating variables that characterize a feature of the task or population and so take the same value for all subjects. (As suggested in Chapter 2, uniform variables can help explain individual differences too.) For instance, imagine there is more variability in people's ability to recall images than in their ability to recall words. The memory retention model could be extended to capture this by incorporating an indicator

variable z representing whether the stimuli are images or words. Individual α values could then be drawn from a distribution partly determined by z. Assuming one could intervene on z in a way that reduces Var(α), the extended model would partly explain some variation explananda about α. The model would show, for instance, that certain memory experiments find highly variable decay rates among their subjects because they use imagistic stimuli.

These examples show that it is not difficult to imagine extensions to simple stochastic parameter models that would, if successful, causally explain parametric variation. By adding covariates and other variables that influence the amount of variation in the population, stochastic parameter models' explanatory power can be enhanced.

3.6 Beyond Explanation

I have now argued that most stochastic parameter models in psychology today do not causally explain parametric variation, though they have the potential to do so when extended. This suggests that the ability of HBMs to shed new light on individual differences in psychology has thus far been somewhat overblown. Still, it is worth remembering that causal explanation is just one of the aims of modeling. Even if existing stochastic parameter models are limited in their causal-explanatory power, they may have other advantages in the realms of description and prediction. I'll now briefly discuss four of these potential virtues, from the least controversial to the most. The first two advantages are the result of stochastic parameter models' hierarchical structure, and so apply equally well to hierarchical models estimated with maximum likelihood methods. (The model structure, not the estimation technique, is "doing the work.") By contrast,

HBMs are sometimes claimed to have a slight edge over hierarchical models estimated with maximum likelihood methods when it comes to manifesting the final two virtues.

The first advantage of stochastic parameter models over non-hierarchical models is that they provide a perspicuous description of individual differences in the population. Once estimated, stochastic parameter models characterize parametric variation quantitatively via their hyper-parameters. This is so regardless of whether the hyper-parameters are interpreted as summary variables. Second, by characterizing individual differences in the population, stochastic parameter models draw attention to parametric variation. If psychologists are to explain parametric variation, it must first be constituted as an explanandum: the contours and boundaries of the variation must be understood prior to any attempt to understand its causes (Bogen and Woodward 1988). By providing population-level descriptions of variation, stochastic parameter models help identify the explananda to be targeted by future research.

Third, and more controversially, there is some evidence that HBMs provide more accurate estimates of first-order parameter values than non-hierarchical models and/or hierarchical models estimated with maximum likelihood methods. Most of the evidence comes from simulation studies in which researchers simulate data from an experiment, analyze the data using different modeling approaches, and assess which model recovers the data-generating parameters most accurately. For example, Nilsson et al. (2011) compare a hierarchical Bayesian implementation of prospect theory with a non-hierarchical model (with no individual differences) estimated with maximum likelihood methods. They find that the hierarchical Bayesian model recovers the data-generating parameters "somewhat more accurately… and with less variability" (*ibid.*, 89).

The primary reason stochastic parameter models sometimes outperform non-hierarchical models is that, since they draw individual parameter values from a common distribution, their

parameter estimates are biased toward the population mean. This is called "shrinkage." Shrinkage is thought to lead to greater accuracy because some of the differences in individuals' behavior is due to noise, and some to genuine variation in parameter values. Shrinkage reduces overfitting to the noise, bringing parameter estimates closer to their true values (Rouder and Haaf 2019). Note that shrinkage is a result of hierarchical model structure. There is some debate about whether using Bayesian techniques to estimate hierarchical models provides an additional accuracy boost. Researchers such as Rouder et al. (2005) claim that Bayesian methods produce the most accurate parameter estimates. In a follow-up simulation study, however, Farrell and Ludwig (2008) find that hierarchical models perform better than non-hierarchical models, but it doesn't matter whether they are estimated with Bayesian or maximum likelihood methods. The results of stimulation studies are highly sensitive to the way the simulated data have been generated. There are also theoretical reasons to be skeptical about what exactly simulation studies show us (Lee 2018). So it is appropriate to be cautious in asserting that hierarchical models produce more accurate parameter estimates than non-hierarchical models. Nevertheless, because of shrinkage, it is very likely that they have an accuracy advantage in at least some contexts.

Finally, stochastic parameter models can also give rise to more accurate predictions about the behavior of individual subjects. These models are often in a predictive sweet spot between non-hierarchical models with no individual differences and non-hierarchical models with full individual differences. Models with no variation fail to make good predictions for individuals because they estimate one set of parameter values for everyone. Models with full individual differences may make poor predictions for previously observed subjects, since their parameter estimates reflect noise as well as true variation (i.e., they do not undergo shrinkage). They also fail to make informed predictions about new subjects, since they treat every individual as unique.

Stochastic parameter models fall in the middle, predicting people's differences while not overfitting the data (Lee and Wagenmakers 2013).

Hence, even those stochastic parameter models that do not causally explain parametric variation may advance the study of individual differences in other ways.

3.7 Conclusion: Representing versus Explaining Individual Differences

This chapter has used SHRINK to explore the causal-explanatory potential of stochastic parameter models. Pushing back against modelers who suggest that the hierarchical Bayesian approach enhances our understanding of individual differences in cognitive psychology, I have argued that the simple stochastic parameter models currently in use do not causally explain the parametric variation they represent. Nevertheless, extending such models can make them capable of causally explaining individual differences in parameter values.

Much of the appeal of the hierarchical Bayesian approach comes from its elegant resolution of the uniformity/uniqueness dilemma. In stochastic parameter models, variability is acknowledged by allowing individuals to have different first-order parameter values, while a degree of uniformity is preserved by drawing those parameter values from overarching distributions. Non-hierarchical, "full individual differences" models lie too close to the "uniqueness" pole of the dilemma: they estimate each person's parameter values entirely independently, and as a result mistake noise for genuine variation. Adding hierarchical structure constrains the individual parameter assignments, nudging psychological models back toward the "uniformity" pole and sometimes improving parameter estimation. The error that proponents of HBMs threaten to make is in thinking that this descriptive advantage is also an explanatory one.

Stochastic parameter models do indeed help us assign individual parameter values in a principled fashion, striking a sensible balance between uniformity and uniqueness. But they are not guaranteed to causally explain individual differences in the parameters. While all stochastic parameter models escape the uniformity/uniqueness dilemma, only some causally explain parametric variation.

This lesson applies broadly. Psychologists want to understand why people's minds differ from one another. In studying individual differences, though, it can be especially difficult to distinguish modeling approaches that provide an answer from those that merely give us a better handle on the question. Overcoming the uniformity/uniqueness dilemma is a representational challenge, one that requires us to capture variation in a way that recognizes both difference and similarity. Finding a fruitful, empirically adequate way to describe a pattern of variation is a significant accomplishment. But tools that merely help us represent individual differences should not be confused with tools that truly explain them.

4.0 Human Variation and Rational Analysis

4.1 Methodologies of Variation

For decades, the study of psychological variation was mostly confined to what Cronbach (1957) called correlational psychology. Now, though, individual differences are going mainstream. As we saw in Chapter 1, the last few years have witnessed growing appreciation of individual differences in language processing (Swets 2015), visual categorization (Shen and Palmeri 2016), fear conditioning (Lonsdorf and Merz 2017), mental imagery (Reeder et al. 2017), and much else. This groundswell of interest raises the question: are the scientific methods we have used so far to generalize about the mind up to the task of tackling individual differences?

In this chapter, I will address this question by focusing on a common modeling approach known as rational analysis. Rational analysis involves identifying problems that the mind is attempting to solve and deriving optimal solutions to them. The solutions are then used to build models and make predictions about human behavior. Rational analysis emphasizes the nature of the environment and problem structure over mechanistic information. Many of its practitioners claim to be building computational rather than algorithmic models of the mind (Marr 1982). Although rational analysis has arguably been quite successful in many areas of psychology, its ability to shed light on individual differences has not yet been systematically examined. My goal in this chapter is to provide such an examination.

I argue that, although rational analysis has thus far largely failed to account for variation, it has the potential to become a fruitful tool for the study of individual differences if it is applied differently. The rational analysis of causal learning is used as a case study. After introducing

rational analysis in more detail in Section 4.2, I argue in Section 4.3 that traditional rational analysis (that is, rational analysis as it has typically been practiced) has dealt with individual differences quite poorly: most of the time, practitioners of rational analysis ignore variation, modeling only the average or modal behavior of subjects. When they do acknowledge variation, it is usually in normatively loaded terms, with one behavior being labeled rational, and all others irrational. This approach, I argue, is both epistemically and ethically flawed.

But rational analysis doesn't have to be this way. As I show in Section 4.4, the rational framework has the resources to capture variation. What we need is a research program for the study of individual differences based on rationality principles: a rational analysis *of variation*. Characterized in Section 4.5, the rational analysis of variation traces individual differences in behavior back to variation in subjects' goals, environments, or constraints. Like traditional rational analysis, it assumes that individuals are behaving rationally, but it recognizes that what is rational is different for different subjects. Section 4.6 addresses two potential objections to the rational analysis of variation. The first is that it introduces new sources of flexibility into the rational approach, exacerbating existing concerns about overfitting and falsifiability; the second is that rational models have limited explanatory potential because they do not tell us why people behave optimally. Section 4.7 concludes with a discussion of what, exactly, is "rational" about the rational analysis of variation.

4.2 Traditional Rational Analysis

Rational analysis is a research program in psychology that takes as its starting point the assumption that human subjects behave (approximately) optimally or rationally. Named and

championed by John Anderson (Anderson and Milson 1989, Anderson 1990, Anderson 1991a), rational analysis revives an older tradition from the early days of probability theory and formalized logic of tightly coupled normative and descriptive theorizing (Oaksford and Chater 1998). The methodology of rational analysis involves identifying a problem that the mind is attempting to solve, characterizing the environment, and specifying minimal processing constraints under which the mind is operating. The modeler then derives an optimal (or approximately optimal) solution to the problem given the constraints and uses that solution to make new behavioral predictions and build models that capture human behavior.

Rational analysis is often contrasted with a mechanistic approach to the mind (Sakamoto et al. 2008). As Anderson (1990) explains, rational analysis "focuses on what is outside the head rather than what is inside" (23). Practitioners of rational analysis concentrate on understanding the structure of the problems that humans face and the environments in which they solve those problems. They wager that "we can understand a lot about human cognition without considering in detail" the mechanisms that are responsible for it (*ibid.*, 3). This is an advantage of the rational approach, advocates claim, because it is easier to test assumptions about the environment than about the brain.[36] It is important, however, not to overstate the independence of rational analysis from mechanistic information, as is sometimes done. Advocates of rational analysis have long recognized that it "does require some assumptions about mechanism to establish the costs and define the constraints under which optimization takes place" (Anderson 1991b, 511). Mechanistic constraints play a role in rational analysis even though elucidating mechanisms is not its goal.

[36] This purported advantage rests on shaky ground. Thanks to technological advances, it is becoming increasingly feasible to test assumptions about the brain. Meanwhile, it remains difficult to gather information about the prehistoric environment in which the human brain evolved or lifetime environments in which individual brains develop.

The core methodology of rational analysis consists of a six-step procedure laid out by Anderson (1990, 1991a) and endorsed by subsequent practitioners of the approach (Chater and Oaksford 2000). The procedure is summarized in Figure 4.1. The first three steps involve the "framing of the information-processing problem" that the cognitive system faces (Anderson 1991a, 474). In Step 1, the goals being optimized by the system are identified. These must be independently motivated; they can't be goals whose only support comes from the behavior to be explained. In Step 2, the structure of the environment to which the behavior is optimized is specified. Anderson explains that this can be done using existing scientific theory about the relevant environment, statistical studies of the environment, or plausibility arguments. It is best to rely on existing theory, and plausibility arguments should only be used if scientific theory and statistical information aren't available. In Step 3, minimal assumptions about the computational constraints on the optimization problem are made. This is where mechanistic information comes in. The modeler is to make as few assumptions as possible while capturing the limits of humans' optimization capacities.[37] Step 4 involves figuring out what the optimal behavior is given the assumptions made in Steps 1–3. Sometimes this is not an analytically tractable problem, so one has to run simulations or make simplifying assumptions to derive an optimal solution. In Step 5, the researcher assesses whether subjects' actual behavior can be predicted from the model, and thus whether the model is empirically confirmed. If not, in Step 6, the theory is refined.[38]

---

[37] Anderson (1990) calls Step 3 "the potential Achilles' heels of a rational approach" (32). The more complex and arbitrary the computational constraints, the less behavior is predicted from the structure of the environment rather than the structure of the mind, and the "less explanation there will be in appealing to optimization" (*ibid.*, 28). Recently, Lieder and Griffiths (2019) have advocated incorporating more robust cognitive constraints into rational analysis, an approach they call "resource-rational analysis" (see Section 4.5.4).

[38] Anticipating that this final step would be controversial, Anderson (1990) argued that iterative theory construction occurs throughout science. The rational approach is no more dependent on *ad hoc*

Table 1 *Steps in developing a rational theory*

1. Specify precisely the goals of the cognitive system
2. Develop a formal model of the environment to which the system is adapted.
3. Make the minimal assumptions about computational limitations
4. Derive the optimal behavioral function given 1–3 above.
5. Examine the empirical literature to see whether the predictions of the behavioral function are confirmed
6. Repeat, iteratively refining the theory

Figure 4.1. The procedure of rational analysis (Anderson 1991a)

There are two senses of "rational model" that are widely conflated in cognitive science. On the first, rational models are models that show that human behavior in a particular domain is substantively rational or optimal. The label "rational" refers to the thing being studied, since these models purport to demonstrate that individuals are acting rationally. On the second reading, rational models are models constructed through rational analysis. The label "rational" refers to a particular kind of discovery process rather than the modeling target. These two senses of "rational model" come apart because a particular model may qualify as rational under the first reading but not the second or vice versa. For instance, researchers in the fast and frugal heuristics research program have shown that the use of heuristics leads to optimal or near-optimal performance in some circumstances (Gigerenzer et al. 1999). However, they firmly reject the method of rational analysis, since they believe that it is naïve to think that humans solve problems optimally. Some heuristics-based models, then, are "rational models" in the first sense but not the second.

maneuvering than any other part of psychology, he argues. He even claims that, in his experience, mechanistic modeling requires more iterative refinement than rational modeling.

Conversely, one may engage in iterative rational analysis and end up with a model that does not show people to be substantively rational. For instance, practicing rational analysis is compatible with positing that humans have irrational goals (Step 1) or that their behavior is limited by computational constraints (Step 3). A rational model in the second sense need not be rational in the first.

In this chapter, I will adopt the second understanding of "rational model," using that term to refer to models constructed in rough accordance with Anderson's procedure of rational analysis (Fig. 4.1). On my usage, a model's rational status depends on how it was built, not what it ultimately says about human behavior. I will say more to defend this use of the term "rational" in Section 4.7.

Many rational models in psychology are Bayesian in character: they posit probabilistic representations of uncertainty and Bayesian updating over those representations. (Such models are Bayesian in the *first* sense described in Section 3.2). However, it is important to keep in mind that there are rational models that are not Bayesian. For instance, Nosofsky's (1998) contribution to Oaksford and Chater's (1998) collection on rational analysis concerns an exemplar model of classification based on the Generalized Context Model (GCM). Nosofsky explains that "the theme of optimal performance has always played a central part in theorizing involving the GCM" (Nosofsky 1998, 218). The model he proposes assumes that subjects distribute their attention over the different dimensions of exemplars in a way that optimizes classification performance. Another family of non-Bayesian rational models are diffusion models of decision-making, which conjecture that response times are minimized for a given level of accuracy (Ratcliff et al. 2016).

Most of what has been said so far to characterize rational analysis is uncontentious among its practitioners. There are, however, several issues on which the community is divided. Psychologists who adopt the rational approach often appeal to Marr's (1982) levels to explain the

theoretical status of their models. Most often, they claim that rational models are computational-level models: they capture problems that cognitive systems solve. At other times, rational modelers suggest that rational analysis can be used to construct models between or at multiple of Marr's levels (Chater and Oaksford 2008a). A related issue is whether subjects actually perform the computations described in the models or it is just "as if" they do (and if so, what exactly "as if" means; van Rooij et al. 2018). Critics of Bayesian models argue that Bayesians are unclear and inconsistent about whether the mind is literally performing Bayesian computations (Jones and Love 2011, Bowers and Davis 2012). Anderson (1990), for his part, argues that the "rational level" isn't "psychologically real" (22). The mind only performs computations that "approximat[e] or realiz[e] the optimal ideal" (*ibid.*, 251).[39]

Rational analysis has been applied in many different psychological domains. Anderson (1990, 1991a) builds rational models of memory, categorization, causal inference, and problem solving. Oaksford and Chater's (1998) collection contains papers proposing rational models of memory, categorization, induction, reasoning, and search. In this chapter, I will focus on rational analyses of causal learning. Many causal learning experiments have the same setup: subjects are presented information about the co-occurrence of events, either in sequence or all at once, and asked to make judgments about the causal relationships between the events. Psychologists are

[39] Rational models (in both senses of that term) are a subspecies of optimization models, which are used in many areas of science, from chemistry (Woody 2019) to economics (Hausman 2018) to biology (Beatty 1980). Optimality approaches are controversial, nowhere less so than in evolutionary biology, where biologists and philosophers of biology have long argued about whether optimality modeling relies on an overly strong form of adaptationism (Orzack and Sober 1994, Potochnik 2009). Although there are similarities between rational modeling in psychology and optimality modeling in biology, there are substantial differences as well. In biology, the fact that some trait can be successfully captured by an optimality model is taken to be evidence that the trait was produced by natural selection; no such inference is made for rational models. In biological optimality models, fitness is always the thing being optimized; in rational models, what is being optimized varies. These dissimilarities mean that most criticisms of optimality modeling in biology do not straightforwardly apply to rational models in psychology.

interested in modeling the relationship between the information presented and subjects' responses to causal probes, both as subjects learn and in the long run (Danks 2007). Causal learning is a good test case for assessing how rational analysis can be used to study individual differences. There are several prominent rational models of causal learning, including the Δp model and power PC theory (Allan 1980, Cheng 1997). Moreover, we know there is variation in what strategies and statistical information subjects use on causal learning tasks (Kao and Wasserman 1993, Anderson and Sheu 1995). Many of the problems with how modelers have handled individual differences, which I will discuss next, appear in the literature on causal learning, as do the seeds for a new, more productive approach to variation, to be discussed in Section 4.5.

4.3 Variation and Traditional Rational Analysis

The remainder of the chapter will examine whether rational analysis is a fruitful approach for studying individual differences. I'll begin by examining how practitioners of traditional rational analysis – that is, rational analysis as it has typically been practiced in the last three decades – have dealt with inter-subject variation, focusing on the case of causal learning. I'll argue that traditional rational analysis has a poor track record: most of the time, variation is ignored; and when it isn't, it is understood through an epistemically stultifying and ethically troubling normative lens.

4.3.1 Traditional Approach #1: Ignore Variation

Most applications of rational analysis do not address individual differences at all. Models are built to capture the average or modal behavior of subjects. The behavior of individual subjects is not modelled and the possibility of subgroups within the population is not considered. Anderson's (1990) model of causal inference is an example of this homogenizing approach. Following the procedure laid out in Figure 4.1, Anderson posits that the goal of a system performing causal inference is to optimize the accuracy of its predictions about the future. Specifically, the system tries to predict the likelihood that an event *E* occurs in the future based on cues *C* that it observes now, using causal rules that connect cues and events. The rational model that Anderson proposes to accomplish this task uses Bayesian estimation of the relevant probabilities.

One of the selling points of Anderson's model is that it can account for a central finding in causal learning: the asymmetric weighting of cells in contingency tables (Fig. 4.2). Contingency tables represent the information presented to subjects about the co-occurrence of two events, *X* (a potential cause) and *Y* (an effect). The cells in the table show the number of trials in which the effect is present or absent, with or without its potential cause. On the basis of this trial frequency information, subjects are asked to judge whether (and how strongly) *X* causes *Y*. A key empirical finding is that, in judging the causal relationship between *X* and *Y*, most subjects do not weight the four types of trials equally. Typically, they place the most weight on cell *a* and the least on cell *d* (Schustack and Sternberg 1981, Kao and Wasserman 1993).

	Y is present	*Y* is absent
X is present	*a*	*b*
X is absent	*c*	*d*

Figure 4.2. A typical 2x2 contingency table

Anderson's model captures this effect. He estimates the model using data from Schustack and Sternberg (1981), showing that under certain conditions, asymmetric weighting of the cells in a contingency table can increase one's predictive accuracy. Anderson's focus is on capturing and rationalizing average behavior, not the behavior of individual subjects. In fitting his model and discussing its advantages, he appeals to *average* causal judgments and ignores variation. This is so even though the data he is using come from a paper that explicitly documents and discusses individual differences in contingency learning (Schustack and Sternberg 1981, 114; see Section 4.6.1). Moreover, in Anderson's (1990) single original experiment on causal learning, the only result that is reported is the mean percentage rating. No individual data are shown and it is unclear how much subjects in the experiment differ from one another (*ibid.*, 184).[40]

In its neglect of individual differences, Anderson's model illustrates a common trend in traditional rational analysis. The literature on causal learning contains a variety of proposed

[40] The only variation that Anderson discusses in the chapter on causal inference is developmental. To model Siegler's (1976) finding that 8-year-olds are more cautious in ascribing causality than 5-year-olds, Anderson assigns the two groups of children different response thresholds.

models, some rational and some not: the Δp model, power PC theory, linear combination models, Rescorla-Wagner model, and probabilistic contrast model, to name a few (Danks 2007). There are many papers assessing how well two or more of these models account for human causal learning data. Most of these comparative assessments only consider whether the different models correctly describe the modal, average, or median responses in causal learning studies. Shanks (1995), for example, considers only mean judgments in his defense of the probabilistic contrast model (Cheng and Holyoak 1995). Cheng's (1997) argument in favor of power PC theory, which makes use of an enormous array of published results, shows that the predictions of power PC theory conform well to mean causal ratings and estimates of causal strength.[41]

Ignoring variation in causal learning is becoming increasingly untenable. Evidence of significant individual differences has been accumulating since at least the 1990s (Schustack and Sternberg 1981, Kao and Wasserman 1993, Anderson and Sheu 1995, White 2000, Steyvers et al. 2003, Buehner et al. 2003, Osman and Shanks 2005, Rehder 2014, Mayrhofer and Waldmann 2016). These studies, several of which will be discussed below, show stable and systematic differences in how people infer causal relationships from contingency information. This variation is deserving of study. Moreover, as discussed in Section 1.2, it has long been understood that averaging in the face of individual differences can be theoretically misleading. In one of the simplest examples, Hayes (1953) points out that if subjects learn a simple association suddenly but at different times, averaging over their learning curves produces a gradual learning curve. A researcher who only looks at the average data will mistakenly conclude that learning is gradual.

[41] Elsewhere these authors do acknowledge variation; but in these papers, the focus is exclusively on aggregate data.

Because of the importance of individual differences and the dangers of averaging over variation, it is a mistake to build rational models to fit aggregate data only.

4.3.2 Traditional Approach #2: Characterize Variation Normatively

When practitioners of traditional rational analysis do acknowledge variation, it is in normatively loaded terms, with one behavior being labeled rational and all others irrational. In Kao and Wasserman's (1993) study of contingency learning, for example, they explicitly take the ΔP rule to be the normative standard.[42] They divide their subjects into two groups based on their collected data: ΔP users and heuristic strategy users. The authors discuss these groups in explicitly normative terms, explaining that the heuristic strategy users' judgments were "systematically biased" (Kao and Wasserman 1993, 1371). The two groups are repeatedly compared in terms of accuracy and correctness, with the heuristic strategy users cast as deficient. White (2000) provides an even more striking example of how non-dominant responses in causal learning tasks are characterized as irrational. He distinguishes between two subgroups of subjects which he labels the "idiosyncratics" and the "normatives" (*ibid.*, 422). The "idiosyncratics" are those subjects who do not "conform to the prescriptions of either the ΔP rule or the associative-learning models" (*ibid.*, 416). In his Experiment 1, there were 8 cases of "idiosyncratic" reports by 6 subjects, out of 40 total subjects.

Even when such language is not used, it is common for practitioners of traditional rational analysis to assume that a single behavior is normatively acceptable and all deviations constitute

[42] The ΔP rule states that causal strength between a cause x and an effect y is $p(y \mid x) - p(y \mid \sim x)$. It can be calculated from a contingency table as $a/(a + b) - c/(c + d)$.

irrationalities. Understanding individual differences through this normative lens is both epistemically unproductive and ethically suspect. It is epistemically unproductive because it does not promote the scientific understanding of variation. Simply labeling a behavior irrational does not itself explain the behavior. This may seem obvious, but psychologists do occasionally mistake the christening of phenomena ("effects" or "biases") for the explanation of those phenomena. Calling a certain pattern of behavior "The Hawthorne effect," for example, does not explain *why* people modify their behavior in particular ways when they realize they are being observed. Similarly, merely characterizing a particular pattern of behavior as a manifestation of irrationality does not explain it.

Furthermore, when one is engaged in rational analysis, labeling a behavior irrational places it beyond the scope of investigation and explanation. The foundational, heuristic assumption of the rational approach is that behavior is rational in some sense. The work of modeling is to figure out precisely how this could be so. Without the assumption of rationality, rational analysis offers no way to proceed; the procedure laid out in Figure 4.1 *requires* that we suppose subjects are behaving optimally given some task specification. Viewing variation through a normative lens is therefore an epistemic dead end when one is practicing rational analysis. Once one declares some subject's behavior irrational, one gives up on trying to understand the behavior using the rational approach. Scientists ought to characterize phenomena in ways that facilitate explanation when it is possible to do so. This applies to variation as much as anything else: we should try to describe variation in a way that enables us to investigate and understand it. Given that rational analysis is fundamentally unable to account for behavior categorized as irrational, rational modelers studying individual differences should be wary of characterizing variation in terms of irrationality. We will almost always learn more by adopting

a heuristic assumption of rationality than by immediately dismissing a subset of behaviors as irrational.

In addition to these epistemic concerns, there are also ethical reasons to avoid viewing variation through an explicitly normative lens. The problem is not that normative terms like "bias" and "idiosyncratic" should be kept out of science entirely. Philosophers of science have shown that thick concepts play an ineliminable and sometimes beneficial role in many parts of science (Longino 2004, Dupré 2007). Rather, there is something particularly problematic about conceiving human psychological variation in terms of irrationality. Labeling some behaviors rational and others irrational collapses behavioral variation onto a single, normative dimension, and obscures the potential benefits of cognitive diversity. What researchers call irrational patterns of responding may in fact be advantageous in certain circumstances or with respect to alternative rational benchmarks. Recognizing that there are many different ways of classifying behaviors, and many dimensions along which they can be evaluated, is a more egalitarian way of conceiving variation. We saw above that researchers using traditional rational analysis occasionally characterize not only behaviors, but also subjects themselves, as irrational. This is an especially worrisome trend that threatens to encourage the reification of different types of people, some of whom are more capable than others.

Using irrationality of either behaviors or people as the default frame for understanding human variation is fundamentally uncharitable. It condemns those who don't conform to researchers' expectations, rather than recognizing that different behaviors might make sense for different people or in different circumstances. Approaching human variation with an attitude of charity and humility is to be preferred on ethical grounds: when we try to understand the behavior of others, disparaging behaviors as irrational should be done sparingly. Practitioners of

rational analysis who hastily conclude that behaviors that do not conform to their model are irrational fail to exhibit a charitable attitude toward all of their subjects.

We have now seen that traditional rational analysis either focuses exclusively on average or modal responses or characterizes variation in terms of rational and irrational people or behaviors. The former approach is inadequate and potentially misleading, while the latter is an epistemic dead end that also raises ethical concerns. The question is: can rational analysis do any better?

4.4 Rational Resources for Capturing Variation

When a rational model labels a subject's aberrant behavior "irrational" or ignores it altogether, it fails to capture that behavior. Some authors seem to think such failures are an inevitable part of the rational approach: rational models simply can't capture variation. This is hinted at in Jones and Love's (2011) contention, for example, that the rational modeler's goal is to "determin[e] the – *presumably unique* – optimal pattern of behavior" (178; my italics). However, their presumption is too quick. The practitioner of rational analysis need not claim that only a single behavior is optimal, rendering all observed deviations unaccounted for. In this section, I'll catalogue the resources available for capturing individual differences within the rational framework and argue that, in fact, the principle of charity on which rational analysis is based is especially appropriate for the study of psychological variation.

4.4.1 Anderson's Strategies and Beyond

Anderson (1990) anticipates that variation will be taken to be a problem for rational analysis. He imagines an objector asking, "If human behavior is optimized, does this not imply that all people should behave the same, optimal way?" (254). His reply is a decisive "no," because the practitioner of rational analysis is not committed to "uniqueness of solution" (*ibid*., 254). He claims there are two families of ways to incorporate variation into rational models. The modeler can either say that subjects pursue different strategies or assign them different parameters. There are two kinds of parameters that can vary between subjects: environmental parameters, which describe the structure of the environment, and computational parameters, which describe the costs of computation or optimization. If their environmental or computational parameters differ, subjects who behave differently may nevertheless all be describable by a rational model.

In fact, Anderson's first tactic of positing different strategies reduces to the second tactic of assigning different parameters. Anderson only briefly mentions the first one in his (1990) book, but we can get a feel for what he has in mind by looking elsewhere. Anderson and Milson (1989) propose a rational model of memory which treats memory as an information-retrieval problem. The model assumes that people search through memory items in order of "need probability," or the probability that a given item is required, and stop searching once they reach some threshold. At the end of the paper, Anderson and Milson acknowledge that, "it is well documented that subjects engage in numerous strategies for processing to-be-learned material, and these strategies can substantially affect their memory performance" (*ibid.*, 714). Some subjects, for example, repeat memory items over and over, while others pursue a strategy of elaboration in which they add meaningful information to the items to be remembered. Anderson and Milson propose to incorporate variable memory strategies into their rational analysis by assuming that one's

strategy alters the *input* to one's memory system. An experimenter can give two subjects exactly the same items to remember, but if they are using different memory strategies, they will have different "strategy-determined experiences" (*ibid.*, 714). This allows Anderson and Milson to preserve the assumption that all subjects are behaving rationally: by using different strategies, "subjects can essentially manipulate the input to their memories… [So] human memory, blind to the intentions of the subject and to the fact of a deliberate manipulation, behaves as rationally as it can, given the statistics of the input it receives" (*ibid.*, 714). They deliberately leave it open whether certain strategies (that is, certain manipulations of inputs) are themselves more rational than others.

Although Anderson and Milson do not end up modifying their model to incorporate multiple memory strategies, their claim that those strategies alter the lists that reach one's memory system suggests that differential strategy use is ultimately to be cashed out in terms of different environmental parameters. Their distinction between the information being presented to subjects and the information that reaches the system of interest also suggests that there are two ways of assigning different environmental parameters to different subjects. Several authors have noted that there is ambiguity about whether the environmental parameters in rational analysis ought to be interpreted "objectively" or "subjectively" (Becker 1991). On the one hand, Anderson emphasizes that a central advantage of rational modeling is that it focuses on what is outside the head, so its "parameters have meaning in the external world" (Anderson 1991b, 508). This implies that environmental parameters represent "objective" features of subjects' environments. On the other hand, one could argue that what really matters is how the environment is *perceived* by the subjects to be modeled, so environmental parameters should be "subjective." Anderson and Milson (1989) arguably adopt this subjective approach in their treatment of memory strategies. They argue that, instead of "simply deriv[ing subjects'] behavior from an analysis of the

information the experimenter is presenting to them," we must examine how the strategies "transform" the inputs (*ibid.*, 714). Their idea seems to be that a rational model of memory should be based on what subjects think the inputs are, rather than what subjects are actually presented. This approach is consistent with Oaksford and Chater's (2009) claim that "the core objective of rational analysis…is to understand the structure of the problem *from the point of view of the cognitive system*" (72).[43]

Permitting the subjective interpretation of environmental parameters introduces another way of accounting for individual differences in rational models. Subjects who behave differently from one another can be modeled as perceiving their environment in different ways. This way of accommodating variation has occasionally been used in causal learning research, where psychologists have shown that behavioral variability on some tasks can be traced back to variable uptake of the presented information. Buehner et al. (2003) gave their subjects information about a number of "patients," indicating whether each patient had taken a particular drug and whether they had developed headaches. The researchers were interested in subjects' estimation of the causal strength of the drug causing or preventing headaches. In a variant of their original experiment, they also asked subjects to report whether headaches had occurred more often, less often, or equally often in the experimental group compared to the control group. The authors found that some of the individual differences in causal strength estimates could be explained by

[43] It is not clear whether Oaksford and Chater intend this passage to endorse the subjective approach. Jones and Love (2011), who are less sympathetic to rational analysis, also argue for subjective variables: "as far as predicting behavior is concerned, all that should matter is what the subject *believes* (either implicitly or explicitly) are the true [environmental] probabilities. Decoupling information encoded in the brain from ground truth in the environment allows for separation of two different tenets of the rationalist program. That is, the question of whether people have veridical mental models of their environments can be separated from the question of whether people reason and act optimally with respect to whatever models they have" (184).

variability in subjects' perceptions of the frequency information. Some subjects simply misperceived or misremembered the frequencies in the patient data. Such studies indicate that rational modelers interested in capturing variation can look beyond the objective information presented, to the potentially variable uptake of that information by different subjects.

Adopting subjective environmental parameters does undermine one of the original motivations for rational analysis, which promised to shed light on behavior without wading "inside the head." As discussed in Section 4.2, however, the rational approach has never been able to avoid internal mechanisms altogether, since it incorporates processing constraints. Interpreting environmental parameters subjectively is just another step in the same direction. There are really two ways, then, of pursuing Anderson's strategy of assigning subjects different environmental parameters: one can accommodate behavioral variation by tracking differences in subjects' objective environments, or by identifying differences in how subjects perceive their shared environment.

Another way of accounting for variation, conspicuously absent from Anderson's list, is to recognize when subjects have subtly different aims. We have seen that Anderson entertains the possibility of variation in environmental parameters (corresponding to Step 2 in Fig. 4.1) and in computational parameters (Step 3), but he neglects potential variation in subjects' goals (Step 1). This way of capturing variation is likely to be important in causal learning. In general, subjects may have different goals in the same experimental task if they have different understandings of what they are being asked to do. One potential source of such differences is ambiguity in the wording of probe questions. There is considerable evidence that subjects' behavior on causal learning tasks is influenced by the exact formulation of the question they are asked. Matute et al. (1996) find that the degree of competition between causes (i.e., how much causal ratings are diminished by the presence of multiple potential causes) depends on whether subjects are asked

a *causality* question of the form "Is *C* the cause of *E*?" or a *contiguity* question of the form "When *C* is present, does *E* occur?" Following up on this study, Collins and Shanks (2006) consider the effects of *counterfactual* probes, which ask about what would happen in an alternative, possible scenario (Buehner et al. 2003). The primary lesson from such research is that "apparently minor presentation changes can give rise to very different judgmental patterns" in causal learning tasks (Perales and Shanks 2008, 1484). Subjects' responses are driven by perceived task demands, which are highly sensitive to the wording of the probe question.

Thus far, research on causal probe wording has focused on average effects. Experiments have shown that certain types of causal probes lead to higher or lower average responding. These results raise the possibility that there could be also individual differences in causal probe interpretation. Given that people seem to be sensitive to "apparently minor" wording changes, it is plausible that different subjects interpret the same probe in different ways. Of course, the presence of between-probe effects is no guarantee of between-subject effects, but it is at least suggestive. Variability in probe interpretation may be common, with different interpretations leading subjects to pursue different aims. Another strategy for modeling variation can therefore be added to Anderson's list. In causal learning and elsewhere, the possession of different goals, perhaps engendered by different interpretations of the experimental task, can be invoked to explain behavioral variability. (Section 4.5.2 will present an example of this strategy in action.)

4.4.2 Bayesian Tools for Capturing Variation

Despite Anderson's insistence that rational analysis can accommodate individual differences, he does not model variation in his work. All of the models in his classic (1990) book "tr[y] to predict behavior by deriving a single solution" (254). To see rational modelers actively

grappling with variation, we must look to recent Bayesian work on causal learning. There we find several distinctively Bayesian strategies for accommodating individual differences.

First, variable behavior can be captured by assigning different subjects different priors. Mayrhofer and Waldmann (2016) pursue this strategy in their study of how people figure out which variables are causes and which are effects given co-occurrence information. They hypothesize that people have abstract prior beliefs that influence their inferences about causal structure. The authors distinguish between two types of priors: "a sufficiency prior (i.e., preference for high causal strength) and a necessity prior (i.e., preference for low base rate of effect)" (*ibid.*, 2141). Mayrhofer and Waldmann divide their subjects into three groups based on the data they collect: subjects whose behavior is consistent with a Bayesian selection procedure using a sufficiency prior (54%), a Bayesian selection procedure using a necessity prior (14%), and random guessing (26%). A handful of subjects could not be uniquely assigned to any of the three clusters (6%). The authors argue that these clusters represent stable individual differences, since 92.4% and 92.0% of the participants' behavior in the sufficiency prior cluster and necessity prior cluster are consistent with the predictions of a Bayesian model with the respective prior. The authors also show that they can manipulate people's priors, producing changes in behavior that are consistent with the predictions of their model.[44]

A second way to accommodate individual differences within a Bayesian framework is to show that there is variation in the assumptions people make about sampling (which, in turn, influence their assumed likelihood functions). Navarro et al. (2012) explore this strategy with a

[44] It is interesting to consider how Bayesian tactics for accommodating variation (like assigning variable priors) relate to the strategies Anderson proposes. Anderson (1991b) argues, "my view of the priors…is that they incorporate the experience from the *evolutionary and personal history* of the individual" (513; my italics). If one takes this history to be the history of environments, then assigning individualized priors seems to be an instance of capturing individual differences using variable environmental parameters.

Bayesian model of inductive generalization. In induction tasks, subjects must make an assumption about how the sample provided by the experimenter was generated: "in strong sampling, data are assumed to have been deliberately generated as positive examples of a concept, whereas in weak sampling, data are assumed to have been generated without any restrictions" (*ibid.*, 187). If subjects assume strong sampling, the generalizations they make will depend on the number of observations, but if they assume weak sampling, they will not. Strong and weak sampling are at two ends of a continuum, with many intermediate sampling assumptions in between. Navarro and colleagues' primary result is that subjects' responses accord well with their Bayesian model on the assumption that individuals have different beliefs about how the data were generated. Variation in how people behave can be traced back to variation in their likelihood functions, and from there to differences in their assumptions about the evidentiary value of presented data.

4.4.3 Rational Analysis and the Principle of Charity

The previous two subsections have shown that, far from being committed to homogeneity (cf. Jones and Love 2011), rational analysis has the resources to accommodate individual differences. Subjects can be modeled as facing different environments, either real or perceived; having different computational limitations; or pursuing different goals, perhaps because of different interpretations of the experimental situation. In Bayesian models, they can be assigned different priors or likelihood functions. Behavioral variability in an experiment is compatible with the assumption that humans find optimal solutions to the problems they face so long as we allow that those problems differ slightly between people.

Moreover, a rational approach to variation could be especially valuable given its foundational commitment to charity toward subjects. Historically, rational analysis has been used as a corrective to psychological research claiming to demonstrate pervasive human irrationality. Practitioners of rational modeling have aimed to show that behaviors elicited in the lab and claimed to be irrational actually involve the deployment of cognitive capacities that work well in less artificial settings (Oaksford and Chater 1998, Chater and Oaksford 2000). Rational modelers take behaviors that do not seem to make sense and show that they are reasonable once one considers computational constraints, typical features of the environment, and what subjects' goals might be. Griffiths (2009) describes the ethos: "Rather than deciding that people solve a problem poorly, we should consider the possibility that they are solving another problem well, and try to determine what that problem might be" (89).

Rational analysis therefore embodies a kind of principle of charity, not unlike those discussed by philosophers of language. Chater and Oaksford (2000) recognize this parallel:

> [We should] ai[m] to 'do the best' for human everyday reasoning strategies – by searching for a rational characterization of how people actually reason. There is an analogy here with rationality assumptions in language interpretation (Davidson 1984; Quine 1960). We aim to interpret people's language so that it makes sense; similarly, the empirical approach to rationality aims to interpret people's reasoning behavior so that their reasoning makes sense. (105)

The rational models of reasoning that Chater and Oaksford have in mind, however, are primarily models of modal or average human behavior. Like many rational modelers, they apply the principle of charity only to the *typical* responses on reasoning tasks.

But the charitable approach these authors advocate can and should be applied to variable behavior as well. Indeed, a scientific principle of charity seems like an especially good starting point for the study of individual differences in psychology. Instead of ignoring or stigmatizing variation when we encounter it, if we are guided by a principle of charity we will ask: how might

people's different behavior *make sense* given their different experiences, situations, or goals? Adopting this charitable heuristic is both epistemically and ethically advantageous. It is likely to be scientifically fruitful because it can make new lines of inquiry salient and lead to the formulation of successful explanations. As I argued above, labeling aberrant behaviors "irrational" is an epistemic dead end; entertaining the possibility that they could be rational, if only we understood the problem structure more clearly, is the opposite. Adopting a scientific principle of charity for the study of psychological variation is also ethically appealing for reasons discussed above. Instead of condemning subjects who behave in unusual ways, being charitable involves trying to see things from their perspective. Different behaviors might make sense for different people.

Thagard and Nisbett (1983) raise several objections to applying a strong principle of charity in social science, one of which concerns its ethical appeal. They argue that despite its egalitarian veneer, "[a] rampant principle of charity preempts the possibilities of criticism and improvement. If we cannot assume actions and judgments to be irrational, then we cannot hope to educate and improve choice strategies and inferential procedures" (*ibid.*, 263). This line of argument is also a response to my suggestion in Section 4.3.2 that cavalier attributions of irrationality are ethically irresponsible. I would argue, however, that criticism and improvement do not require irrationality diagnoses. It is possible to adopt a charitable attitude and capture a subject's behavior with a rational model but still critique that behavior. For example, even if a subject is acting appropriately given her understanding of the task, she may have failed to retain information presented to her, as in the study by Buehner et al. (2003) discussed in Section 4.4.1. Such a subject is criticizable for this mistake. The charitable spirit of rational modeling is therefore compatible with attempts to critique and improve subjects' behavior.

Given the principle of charity at the heart of rational analysis, and the resources at its disposal for accommodating variation, the seeds of a more promising approach to modeling individual differences are already present in the existing literature. But this potential needs to be highlighted and cultivated. To overcome the poor treatment of individual differences by traditional rational analysis, we need a new agenda for rational modeling that places individual differences front-and-center: a rational analysis *of variation*.

4.5 Rational Analysis of Variation

The rational analysis of variation, partially but not yet fully actualized, is a research program that applies a rational approach to the study of individual differences in psychology. In this section, I'll characterize the rational analysis of variation and discuss how it has been and should be applied in the study of causal learning.

4.5.1 General Approach

The rational analysis of variation involves looking for variations in problem structure that lead to multiple rational behavioral variants on (what appears to experimenters to be) the same task. Upon encountering individual differences, the practitioner of rational analysis of variation asks: how might people's different behavior *make sense* given their different experiences, situations, goals, and so on? Her initial assumption is that variability in behavior reflects variability in the problems that subjects are solving. Just as traditional rational analysis has been used to rationalize purportedly irrational performance on reasoning tasks, rational analysis of

variation can vindicate human diversity by showing that optimal behavior differs across subjects and contexts.

Rational analysis of variation starts with an acknowledgement of variability in the behavioral data to be modeled, followed by an examination of task structure with an eye toward potential sources of difference. Any of the strategies for accommodating variation discussed in Sections 4.4.1 and 4.4.2 can be used in model-building. The modeler looks for evidence that subjects are facing different environments, either real or perceived; that they are working with different computational constraints; that they have different goals, perhaps because they interpret the instructions differently; or that they have different priors or likelihood functions, if the model is Bayesian. Once the researcher has identified a hypothesized source of variation, she builds a rational model, deriving optimal solutions to each of the problems she suspects her subjects are confronting. As in Anderson's original procedure (Fig. 4.1), these solutions are compared against the actual pattern of variation found in the data and the model is iteratively refined.

4.5.2 Methodological Principles

Responsible application of the rational analysis of variation requires that modelers adhere to several methodological principles.

(1) **Independent Measurement:** Every effort must be made to obtain independent measurements of the explanans variables.

Modeling individual differences with rational analysis should not be like adjusting a dial until the desired results are achieved. When a modeler hypothesizes that a certain variable is the source of observed individual differences, he should make every effort to measure that variable in order to confirm that it does indeed differ across subjects and that the differences are correlated

with differences in the observed behavior. If he hypothesizes that behavior is variable because subjects' environments are slightly different, for example, he can try to measure the environmental features in question.[45] For example, recall that Buehner et al. (2003) hypothesized that their subjects were responding differently to the causal probe because they were perceiving or remembering the co-occurrence data differently. To confirm this, the authors asked subjects what the relative frequencies of different types of trials had been. They found that some of their subjects possessed mistaken beliefs about the data, rendering unnecessary any deeper explanation of the variation in terms of fundamentally different causal learning processes. This conclusion would have been *ad hoc* had the authors not separately solicited subjects' beliefs about the frequencies in the data.

Sometimes it is not possible to measure the hypothesized sources of variation as Buehner and colleagues did. In these cases, the second and third methodological principles become even more important.

(2) **Model Comparison:** Rational models of variation must be tested against (i) non-rational models of the variation, and (ii) models that assume the variation is random noise.

Rational analysis of variation does not involve an unshakeable commitment to the rationality of every individual subject. It simply begins with the defeasible hypothesis that subjects' behavioral responses are optimal with respect to a goal. To ensure that the heuristic assumption of rationality does not become dogma, rational modelers must compare their models of variation to two sorts of alternatives.

[45] Bayesian modelers have been criticized for neglecting to measure features of actual environments (Jones and Love 2011, 180). When environmental parameters are varied to account for individual differences, direct measurement of (real or perceived) environments becomes even more critical.

First, the best rational model of the variation in question should be compared to the best non-rational model. Bayesian modelers are often criticized for failing to make such comparisons. Bowers and Davis (2012) argue that Bayesians rarely present head-to-head comparisons of their preferred models with non-Bayesian (and non-rational) models of the same behavior. They agree with Jones and Love (2011) that "competing alternatives need to be explicitly recognized and compared" (184). Practitioners of rational analysis of variation should test rational models against non-rational models of variation. If a rational model performs substantially worse than a non-rational model, the latter is to be preferred. Note that even if a particular rational model of variation should be rejected, the rational approach itself need not be abandoned. There may be an alternative rational model that fits the data better. Moreover, even if one has a highly successful non-rational model, that does not mean the variation cannot also be successfully rationally modeled. A non-rational model can "fill in" the mechanistic details about how behavior depends, in a manner captured by the rational model, on some feature of the task at hand. Rational and non-rational models of the same phenomenon are compatible.

Second, rational modelers should also assess whether the variation in question is simply noise. There are different ways of conceiving of noise, and hence different ways of understanding this methodological principle.[46] One heuristic for assessing whether behavioral variation constitutes noise involves examining whether it is random with respect to all the variables that characterize the structure of the task. Non-zero correlations between the behavior of interest and

[46] Recall that the definition of noise that I favor, discussed in Section 1.5.4, holds that noise is heterogeneity that cannot be reduced by intervening on any of the variables in one's variable set. The rational modeler's variable set is comprised of variables that describe the structure of subjects' task, broadly construed. This includes variables representing subjects' goals, environment, or computational limitations. On this conception of noise, the alternative hypothesis that the target variation is noise amounts to the claim that the variation cannot be reduced by intervening on variables that characterize the structure of the task.

the task parameters suggest that the variation is not just noise. If the variation to be modeled does not bear a systematic relationship to any of the variables characterizing the task, it is likely to frustrate rational analysis, because the rational approach must account for variation by appeal to differences in problem structure across subjects.

This second methodological principle, which acknowledges that the rational approach might not be appropriate in all circumstances, is in keeping with the original spirit of rational analysis, if not always its practice. Anderson (1990) writes, "I don't know yet how successful a rational analysis will ultimately prove to be for different aspects of human cognition" (xii). When one applies rational analysis to some psychological capacity, one may find that "the given aspect of human cognition is not optimized in any interesting sense" (Anderson 1991a, 472). Oaksford and Chater (2009b) concur, saying that they are "happy to agree with commentators who suggest that there are cognitive phenomena for which purely rational considerations provide an incomplete, or indeed incorrect, explanation" (112). Practitioners of the rational approach to variation should actively compare their rational models of variation to both non-rational models and models that treat the variation as noise. Only the second type of comparison can force us to abandon the rational analysis of variation in a dataset altogether; the first type merely justifies the rejection of particular rational models.

(3) **Repeated Subject Measures:** If possible, data should be obtained from the same subjects over multiple testing sessions.

Variables that have the potential to account for individual differences in rational models fall into two categories: those assumed to stay relatively stable over time for a given individual, and those subject to transient fluctuations. Variables that are expected to stay relatively constant over short timescales include one's computational constraints (holding fixed other computational

demands), the set of environments that one has encountered over the course of one's life, and sometimes one's priors and likelihood functions.

Psychologists can learn a great deal from repeated behavioral measurements of the same subjects, ideally from testing sessions on different days. Taking repeated measurements allows researchers to test the hypothesis that behavioral variation is due to intra-individually stable factors. If modelers suspect that individual differences in task performance are caused by variation in lifetime exposure to a particular kind of environment, for example, they can administer the task over multiple testing sessions to check that people's behavior stays reasonably constant. Of course, there are significant practical obstacles to testing the same subjects multiple times, so this guideline should be seen as an ideal rather than a requirement.[47]

4.5.3 Rational Analysis of Variation in Causal Learning

I have argued that the rational analysis of variation represents a departure from traditional rational analysis, offering a more epistemically fruitful and ethically appealing way of approaching individual differences in psychology. This is not to say that all practitioners of rational analysis have been pursuing the traditional approach. Some of the causal learning research discussed in Sections 4.4.1 and 4.4.2 lays the groundwork for the research program I am advocating.

[47] It can also be helpful to measure subjects' performance on different kinds of tasks. Osman and Shanks (2005), for example, show that people's sensitivity to base rates is relatively constant across both a causal learning task and a decision-making task. This supports efforts to attribute behavioral variability to differences in base rate sensitivity.

Buehner et al. (2003) is arguably the best example of rational analysis of variation in action. We have already seen how Beuhner and colleagues determined that some of the variation in their data was caused by differences in how subjects perceived or remembered the frequencies they had been shown. This is just one of several instances in their paper of the rational analysis of variation. Buehner and colleagues endorse power PC theory as the correct normative theory of causal learning (Cheng 1997). Upon encountering individual differences, however, they do not ignore them, nor do they write off those subjects who deviate from the predictions of power PC theory as irrational. Instead, they formulate and test hypotheses about potential sources of variation.

In Experiment 1, Buehner et al. present subjects with data about the relationship between being vaccinated with a new vaccine and developing a viral disease (preventive cause scenario), and about viruses being exposed to certain rays and subsequently mutating (generative cause scenario). The experimenters ask subjects to judge "how strongly each vaccine prevented the disease related to the virus in question" or "how strongly the particular rays cause mutation," both on a scale from 0-100 (Buehner et al. 2003, 1123). Subjects' responses differ. The authors speculate that ambiguity in the wording of the causal probes could be the reason. They point out that the questions afford two interpretations: "What difference does the candidate cause make in the current learning context, in which alternative causes already produce [the effect] in a certain proportion of the entities?" or "What difference does the candidate cause make when alternative causes never produce [the effect]?" (*ibid.*, 1126). Buehner and colleagues hypothesize that some subjects interpreted the causal probe question in the first way, and some in the second. They perform a K-means cluster analysis, sorting their subjects into two groups based on behavioral similarity. They find that the two groups line up with the two potential interpretations: one group's ratings are normative with respect to the first question, while the other group's ratings

are normative with respect to the second. To further support their hypothesis, in Experiment 2 the authors eliminate the ambiguity by adopting a causal probe with counterfactual wording. They make several other protocol changes too, including adopting a different scenario about the connection between taking a certain medicine and getting headaches. Subjects who indicated that the medicine caused headaches were asked, "how many out of 100 people who did not have headaches would have a headache if given the medicine" (*ibid.*, 1131)? Sure enough, much of the response variability was eliminated when this change was implemented.

This work illustrates what it takes to build rational models of variation. Confronted with individual differences in subjects' causal ratings, Buehner and colleagues hypothesized that the variation was due to variability in subjects' goals, brought about by different interpretations of the causal probe. They looked for and found multiple lines of evidence for this hypothesis. This showed that subjects who behaved differently were nevertheless responding rationally given how they understood the task.

Buehner et al.'s sequence of experiments conforms reasonably well to the methodological principles formulated above. Although they do not independently measure subjects' interpretations of the causal probe as recommended by the first principle, they eliminate the probe ambiguity in Experiment 2 and find that it reduces the behavioral variability. This ensures that their hypothesis about the source of individual differences is not *ad hoc*. Their use of cluster analysis instantiates the second methodological principle. If the behavioral variation were just noise, the cluster analysis would not have picked out groups corresponding to the two hypothesized interpretations of the causal probe, nor would the modification of the experimental protocol in Experiment 2 have succeeded in reducing the variation. Buehner et al. did not take repeated measurements from the same subjects at different testing sessions, but this is not a major

shortcoming of the study because there is no strong reason to expect subjects' interpretation of an ambiguous causal probe to be stable over time.

4.5.4 Resource-Rational Analysis

Another contemporary variant of rational modeling, "resource-rational analysis," has recently been developed by Lieder and Griffiths (2019). Resource-rational analysis is a combination of Anderson-style rational analysis with a more intensive focus on cognitive constraints. (It retains the general procedure of Figure 4.1, in other words, but expands the mandate of Step 3.) The heuristic assumption of resource-rational analysis is that people make rational use of limited resources. Lieder and Griffiths argue that resource-rational analysis represents the happy marriage of top-down rational principles with bottom-up mechanistic information.

When it comes to modeling variation, it is possible for resource-rational analysis to fall into the same traps as its traditional predecessor. If resource-rational modelers assume that all subjects are solving the same problems, their models will fail to capture variation. However, it is also possible to combine the resource-rational modeler's heightened emphasis on resource constraints with the goal of accounting for individual differences.[48] Still, when practitioners of resource-rational analysis seek to model variation, they should not assume that all behavioral variability can be traced back to differences in individuals' resource constraints. As we have seen, there are many potential sources of individual differences within the rational framework besides

[48] Because the features of resource-rational analysis and the rational analysis of variation are largely independent, Lieder and Griffiths' proposal is orthogonal to mine. Arguably the best approach would be to adopt both refinements of traditional rational analysis.

computational limitations. Focusing too much on people's variable capacities threatens to stigmatize and oversimplify variation, just like under the traditional rational approach. With sufficient attention to these issues, the resource-rational modeler can be a practitioner of the rational analysis of variation as well.

4.6 Objections to Rational Analysis of Variation

A critic of the rational analysis of variation might argue that the approach is too flexible or explanatorily toothless. I'll now respond to each of these objections in turn.

4.6.1 The Flexibility Objection

One might worry that rational analysis of variation is too unconstrained: given the many ways of accommodating variation, rational models of individual differences are all but guaranteed to be available. This flexibility undermines their usefulness. Even traditional rational models are frequently charged with being overly flexible. Sakamoto et al. (2008) argue that "rational explanations, although illuminating and satisfying, largely serve as just-so stories that are constructed after interesting behavioral findings present themselves" (1064). Snow (1991) claims that rational modelers have access to so many degrees of freedom that "success in coming up with a model seems unsurprising" (506). The approach proposed here is even more vulnerable to this objection. While traditional rational models are at least constrained to model all subjects in the same way, the rational analysis of variation adds several new dimensions of flexibility, exacerbating worries about overfitting and falsifiability.

Adopting the methodological principles formulated above helps to mitigate these worries. If researchers take independent measurements of the explanans variables, test their rational models against random and non-rational models of variation, and obtain multiple measurements from the same subjects, the chance of overfitting is slim. The flexibility objection only gains traction when researchers don't or can't comply with the methodological principles. And even in those cases, the objection underestimates the constraints inherent in rational modeling. When one proposes a rational model of variation, one commits to a particular rational skeleton: a basic picture of rational behavior in the domain in question that applies to all subjects. Individual subjects may vary, but only along certain dimensions within that general framework. Furthermore, major modifications to a model must be normatively justified. One cannot make structural changes to rational models simply to save the phenomena. (Parameters are sometimes changed without justification, but not model structure.)

Because rational models commit to a rational skeleton, rational models of variation are very often *less* flexible than non-rational models of variation. In Section 4.3.1, we briefly encountered Schustack and Sternberg's (1981) study of contingency learning, which established that subjects do not weight the four cells of contingency tables equally. What I neglected to mention was that Schustack and Sternberg propose a non-rational linear model to account for this finding. The authors ask subjects to evaluate causal hypotheses in three experiments using contingency information. They perform a linear regression on the causal ratings with five independent variables: four representing frequency information about the four types of trials encountered, and a fifth representing "the average strength of the two alternative causes that were the best competing explanations for the outcome" (*ibid.*, 1981, 110). On their model, subjects' causal ratings are a linear combination of cells *a*-*d* in a contingency table, plus one additional variable. Initially, Schustack and Sternberg estimate the model for the population of subjects.

Then they fit the model to the individual data and estimate the coefficients separately for each subject. (This is a "full individual differences" model; cf. Chapter 3.) The authors argue that the model is quite successful, since it captures more than half of the variance ($R^2 = 0.64$).

Schustack and Sternberg's non-rational, linear model of individual differences in causal learning raises serious concerns about overfitting. The parameters are estimated for each subject entirely independently. They do not have an interpretation that extends beyond the immediate experimental context, so they are little more than data-fitting devices. This model is not an isolated case. Other authors have proposed similar linear models of causal or contingency learning, though they typically fit the models to the population rather than individual data, running afoul of the problems with traditional rational analysis discussed in Section 4.3.1 (Perales and Shanks 2008, White 2009). Such linear models demonstrate that non-rational models can be more flexible than rational models of the same phenomena.

The lesser flexibility of rational models compared to non-rational models occasionally gives them explanatory advantages. Consider again the asymmetric weighting of cells in contingency tables. When Schustack and Sternberg estimate their model with no individual differences, they find that the parameter with the highest average absolute magnitude is the one that represents cell *a*, and the lowest is cell *d*. The model captures the phenomenon of asymmetric weighting, but only because its parameters have been estimated from the data. It does not explain why the cells are weighted unequally. By contrast, Cheng (1997) shows mathematically that her rational model of causal learning, power PC theory, predicts that *a* and *b* are weighted more than *c* and *d*: "unlike linear models…the power PC theory *explains* the differential weighting without the use of any parameters" (Cheng 1997, 394; my italics). Sewell et al. (2011) argue that this is a general feature of the rational approach:

> One area where Bayesian perspectives appear particularly more illuminating than mechanistic approaches is in explaining individual differences….A significant limitation of a mechanistic approach is that the solutions have been built into the models. By contrast, recent Bayesian modeling… has showed that many aspects of the individual differences observed empirically emerge naturally if one assumes that people are trying to learn about their environment in a rational manner. (212)

Hence, in some cases, rational models have explanatory advantages compared to non-rational, mechanistic models.

Finally, to take a slightly broader view, the flexibility objection should be considered in the context of the uniformity/uniqueness dilemma, a tension inherent to all research on individual differences (see Sections 3.1 and 3.7). On the one hand, researchers cannot treat every subject as unique, or else they will overfit the data and lose projectability to new subjects. On the other hand, they cannot treat all subjects as the same if their behavior is reasonably variable. Hierarchical Bayesian modeling is one tool for finding a middle path between surrendering to and ignoring variation. Rational analysis of variation is another. Its commitment to a shared rational skeleton means it does not treat every individual as unique. But its tolerance of variants in the task specification permits (and, in many cases, explains) a degree of variability in behavior.

4.6.2 The Explanation Objection

A second potential objection against the rational analysis of variation concerns its explanatory potential. There are worries, particularly among philosophers, that rational models do not provide explanations of the behavior they capture (Godfrey-Smith 1991, Danks 2008, Brighton and Olsson 2009, Danks and Eberhardt 2009, Reijula ms). Critics argue that showing that a behavior is optimal does not (by itself) explain the existence of the behavior. To explain "*why* [the mind] is wired up the way it is," a rational model must be supplemented with an

empirically supported causal story about the origins of optimality (Godfrey-Smith 1991, 496). As Danks (2008) puts it, to provide an "optimality-based explanation" of some behavior *X* – to explain why *X* occurs by appeal to its optimality – the rational modeler needs to establish that "people do *X because* it is optimal" (61; my italics). Without such a causal story, the fact that *X* is optimal does not explain why people do *X*, since "there are many other reasons why [the behavior] might occur" (*ibid.*, 62). At least two kinds of stories meet this requirement: one can show that *X* is an evolutionary adaptation, optimized by natural selection; or one can show that learning mechanisms ensure that *X* is optimal.

Critics argue that rational analyses rarely explain the behavior they model because they rarely provide an account of how the behavior came to be optimal. Indeed, practitioners of rational analysis sometimes actively eschew such causal information. Anderson (1990), who claims that "the thesis in [his] book is not about evolution," expresses hope that discussion of rational modeling does not "degenerate into evolutionary arguments" (247). Oaksford and Chater (2009b) likewise decline to offer a causal backstory for their models.[49] As a result, Danks (2008) argues that "many current rational analyses…offer data summaries and potentially predictions, but essentially no additional explanatory power" (59).[50]

What does this critique mean for the rational analysis of variation? One might argue that rational models of variation are explanatory *only if* a causal story about the origins of optimality

[49] Oaksford and Chater (2009b) argue that the optimality-based explanations that their models provide are teleological and so "distinctively non-causal" (113). I happen to believe that this mistaken, and that critics like Godfrey-Smith and Danks are correct that, for the fact of optimality to do explanatory work, one must determine why humans behave optimally in the relevant domain. In my view, Oaksford and Chater's (2009b) protestations to the contrary are rooted in a misunderstanding of teleological explanation as non-causal (Wright 1976). I cannot argue for this here, however.

[50] Danks (personal communication) notes that the word "current" is key here; he agrees that there are now models that are explanatory in the sense to be described shortly.

is provided. In my view, this goes slightly too far. Even if critics are right that optimality-based explanations require a causal backstory about how optimality arose, other kinds of explanations do not. There are many explanatory why-questions, including some about individual differences, that can be answered by rational models that remain non-committal about the origins of optimal behavior. Rational models of variation can answer questions like: "Why do some subjects do x_1 and others do x_2?" and "Why does subject *a* do x_1?" even without information about whether behavior was shaped by natural selection or individual learning. This is because successful rational models capture causal dependencies between behavior and problem structure. They show how manipulating features of the task at hand would change what people do (Rescorla 2018). Note that the designation of behavior *as optimal* is doing no work in such explanations. What matters is that the behavior causally depends on specific properties of the task; its rationality or optimality is beside the point.

For a concrete example, let's return to Buehner et al. (2003). Buehner and colleagues show that individual differences in their original causal learning task are caused by different interpretations of the probe question. Subjects who understand the probe in one way respond optimally given that interpretation, while subjects who read it the other way respond optimally given the other interpretation. The authors do not provide any evolutionary or ontogenetic evidence about why subjects behave optimally, so at least on Danks' view, they do not provide an optimality-based explanation of subjects' causal judgments. But their account still has explanatory purchase. It explains why there was variability ("because subjects interpreted the prompt in different ways") and why each subject responded the way they did ("because they adopted the first/second interpretation of the causal probe").

Hence, rational models can represent causal relationships even when the origins of optimality are unknown. Elsewhere Danks (2008) acknowledges this, claiming that rational

models "enable us to characterize the relevant factors for some cognitive problem, even if we do not know precisely how that factor is used in the cognitive system" (66). In fact, he argues, rational models are sometimes "better able to highlight the relevant features of a situation, precisely because they are not committed to a mechanism" by which the features influence the system (*ibid.*, 66). Thus, rational models of variation can represent causal dependencies between features of a task and behavior that can be used to explain various explananda. This does not require information about why people behave optimally.

4.7 Conclusion: What's Rationality Got to Do With It?

In this chapter, I have argued that rational analysis must be modified if it is to be a fruitful approach to studying variation. Traditional rational analysis papers over individual differences or describes them in a way that is scientifically stultifying and ethically problematic. Although many rational analyses of causal learning exhibit these flaws, others contain the seeds of a superior approach. There are many ways for rational models to accommodate variation: one can posit different real or perceived environments, goals, computational limitations, priors, or likelihood functions. Moreover, the rational approach embodies a scientific principle of charity that is especially appropriate when trying to understand individuals' behavior. Combining and magnifying these elements, I argued for the adoption of rational analysis of variation, an approach that encourages us to ask: how might people's different behavior make sense given their different experiences, situations, and goals? I also recommended that practitioners of this approach adopt certain methodological principles to ensure that they build appropriately constrained models of variable behavior.

One might be puzzled that I have retained the label "rational" for the approach I have advocated. Some authors argue that the claim to rationality is surrendered once subjects are modeled as operating under computational constraints (Sloman and Fernbach 2008). Jones and Love (2011), for example, claim that "characterizing capacity limitations is essentially an exercise in characterizing the mechanism, which represents a departure from rational principles. Once all capacity limitations are detailed, notions of rationality lose force" (183). These authors would presumably argue that the term "rational" does not apply to rational analysis of variation, since it not only incorporates computational constraints, but also allows that different behaviors can be rational given different perceptions of the environment, sampling assumptions, and so on. If a rational model of variation can capture the behavior of a subject who is computationally constrained, who may be misperceiving stimuli or acting on a mistaken understanding of his environment, one might wonder: in what sense is his behavior really "rational" and what makes the model a "rational" one?

The answer, I think, is simply that the subject is behaving rationally in that his behavior makes sense given what he thinks is going on. In my view, this is a reasonable deployment of the concept of rationality. Even if the subject is seriously constrained and mistaken about what he has been told, we should not ignore the fact that he is doing something right. He is behaving optimally given his understanding of the task and the resources at his disposal. Whether this thin sense of "rational" is the one that matters for philosophy or policy or law is a separate issue.[51]

On my usage, a rational model is just a model that was produced via a process of rational analysis (see Section 4.2). Some authors have suggested that many such models do not merit the

[51] Note that it is also thinner than the concept of rationality deployed in the first sense of "rational model" discussed in Section 4.2.

label "rational." Rehder (2011), for example, recommends that we "drop the label 'rational' for these sorts of models and call them what they are, namely, probabilistic models" (210). He goes on to claim, "freeing probabilistic models from the burdens of rationality clarifies both their virtues and obligations" (*ibid.,* 210). I disagree with Rehder's recommendation. It makes sense to retain the label "rational" because it reflects the distinctively normative flavor of the modeling approach. Model building in traditional rational analysis proceeds by examining how the average subject would behave if she were solving a problem *optimally*. Rational analysis of variation applies the same attitude toward individual subjects, as it tries to see their behavior as *making sense* from their perspective. In both traditional rational analysis and rational analysis of variation, then, the construction of models is driven by normative thinking. This makes the label "rational" apt.

Furthermore, the label distinguishes rational analysis of variation from other approaches to modeling individual differences. One alternative way of studying variation would be to decompose the psychological processes responsible for the behavior of interest and then look for variation within those sub-processes (Cummins 1975). An even more distant alternative would be a "bottom-up" approach which seeks to model behavioral variability in mechanistic, neurophysiological terms. There are countless bottom-up studies in which researchers try to correlate behavioral variability with the volume or functional activity of particular brain regions (e.g. Osaka et al. 2003). Their hope is that understanding the neurophysiological basis of behavioral variability will help us explain it. Rational analysis of variation is markedly different from both process-based and bottom-up approaches to modeling variation. The label "rational" picks out what is distinctive about it.

I have not tried to compare the rational approach to other ways of studying variation. Different approaches have different merits, and each may be well-suited to different domains or

types of variation. Just as there are patterns of individual differences that cannot be explained by rational analysis, so too are there types of variability that are unlikely to be illuminated by a process-based or bottom-up approach. Pursuing the rational analysis of variation in some psychological domain is therefore to make a bet: to wager that people's variable behavior makes sense under different task descriptions. Psychology is best served if different researchers make different such bets and so adopt different approaches to the study of individual differences.

5.0 Registration Pluralism and Data Aggregation Across Brains

5.1 Data Aggregation in Neuroscience

Neuropsychology has a long history of drawing conclusions from lone, anomalous cases (think of Phineas Gage or H.M.), but most neuroscientists agree that collecting data from multiple subjects is preferable when it is feasible. The brain is a notoriously noisy organ, and using multiple subjects helps to distinguish the signal from the noise. It also ensures that one's findings are not hostage to the idiosyncrasies of a single brain. But the use of multiple subjects brings with it the challenge of data aggregation: how are the data from different people to be combined and analyzed?

One dominant strategy for dealing with data aggregation in neuroscience is what I will call "the cartographic approach." On the cartographic approach, cross-brain comparison and aggregation are accomplished by placing whole-brain data from multiple subjects into a common reference frame or onto a template (Toga et al. 2006). This mapping (or "registering") of individual data into a shared space allows brains to be compared and group-level statistics to be computed. Although alternative aggregation strategies exist,[52] the cartographic approach has been widely adopted since at least the mid-twentieth century. In that time, it has evolved substantially: from the visual inspection of paper atlases, constructed from post mortem examination of stained sections; through the invention of stereotactic spaces and early landmark-

[52] Alternative strategies include the functional localizer approach (Poline et al. 2010), temporal alignment (Zhang et al. 2017), and hyperalignment (Haxby et al. 2011).

based alignment methods, which were designed for neurosurgery but coopted for neuroimaging research in the 1980s; to the construction of digital brain atlases and the proliferation of automated and semi-automated registration methods, which continues to the present (Toga and Mazziotta 2002, Toga et al. 2006, Evans et al. 2012).

In this chapter, I will characterize the present-day cartographic approach and argue against a tempting view about registration, one of its key components. The view I reject, which I call "registration monism," maintains that all brain data should be registered to spatial templates in the same way. The registration monist takes it to be a problem that different researchers currently use different registration methods and believes that eventually neuroscientists should or will converge on the single best one. I'll argue that this view of the cartographic approach is mistaken. Instead, we ought to embrace "registration pluralism," which claims that the best way to register data to a brain template depends on the phenomenon under investigation. Registration pluralism asserts the in-principle impossibility of ever finding a single spatial mapping suitable for all neuroscientific purposes. This impossibility is a consequence of the substantial individual differences that exist in the organization of the human brain. Registration pluralism therefore highlights the methodological significance of individual differences even when they are not an explicit object of study.

I begin in Section 5.2 by describing the fundamental components of the cartographic approach. Section 5.3 introduces registration pluralism and Sections 5.4 and 5.5 defend it. I then flesh out its scope in Section 5.6. Finally, I explore its possible methodological consequences in Section 5.7 and its broader philosophical significance in Section 5.8.

5.2 The Contemporary Cartographic Approach to Aggregation Across Brains

True to its name, the cartographic approach to aggregation involves mapping information about the brains of individuals to a shared spatial reference frame: whole-brain data from different subjects are projected onto a two- or three-dimensional template or into a stereotactic space (Fig. 5.1).[53] Statistical analysis is then conducted on the aggregate data. Often this involves the use of an atlas to divide the template brain into distinct regions. To understand the cartographic approach in its contemporary form, it is important to understand these central components.

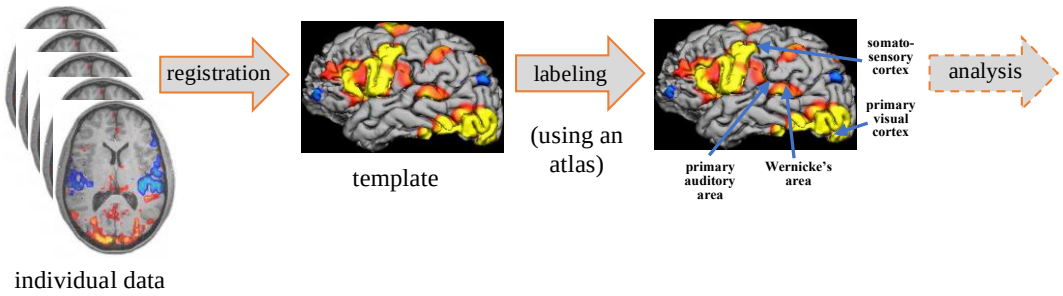


Figure 5.1. Schematic representation of the cartographic approach

A *stereotactic space* is a coordinate system used for specifying locations in the brain relative to internal or external landmarks (Roland and Zilles 1994, Rahman et al. 2009). Stereotactic spaces may be two- or three-dimensional (Tucholka et al. 2012), and they may be applicable to the entire brain or just a part of it. Each stereotactic space comes with rules about how a brain is to be

[53] The cartographic approach, and the problem of aggregation which it solves, can be found in many medical imaging contexts, not just in neuroscience (Crum et al. 2003). The issues to be discussed here may therefore have analogues in other parts of physiology and biomedical science. How far the analogies extend, including whether registration pluralism applies to organs besides the brain, is an empirical matter (see Sections 5.5 and 5.6).

positioned in the space: it specifies where the origin lies and how the axes are oriented. Once a brain is mapped to a stereotactic space, specific points can be labeled with stereotactic coordinates. A groundbreaking stereotactic coordinate system in neuroscience was introduced by Talairach et al. (1967). The three-dimensional "Talairach space" uses the inter-hemispheric fissure and the anterior and posterior commissures, subcortical structures that are relatively invariant across individuals, for orientation.

A brain *template* is a representation of a brain onto which other brains are mapped. Templates may also be called targets, references, or baseline images (Crum et al. 2004). Some studies use a brain scan from one subject chosen at random to be the image to which all others are normalized. Others use templates that have been constructed by averaging images from multiple subjects. Many stereotactic spaces are associated with templates. Talairach and Tournoux (1988), for example, published a template to go along with the Talairach space some twenty years after it was introduced. Like stereotactic spaces, a template can be two- or three-dimensional (Saad and Reynolds 2012).

An *atlas* is distinct from a template in that it has labeled parts. Atlases are representations of the brain that partition the volume (if it is three-dimensional) or surface (if it is two-dimensional) into discrete, labeled regions (Gholipour et al. 2007). The regions that atlases pick out may be cytoarchitectural, macroanatomical, functional, histological, or chemoarchitectural. There are a variety of brain atlases in use today: MNI, ICBM, Harvard-Oxford, and Freesurfer, to name a few (Evans et al. 2012). Even though some authors use "template" and "atlas" interchangeably (Toga 1998, Dickie et al. 2017), it is important to distinguish them because the construction of atlases raises a host of issues about how to divide the brain into parts that templates alone do not. Selecting a representative template is a different problem from partitioning a template in a scientifically useful way.

Registration is the process of transforming a target image in order to relate positions in the target to positions in a template or stereotactic space. This chapter will deal only with cross-subject registration, not the co-registration of multiple images from the same subject, so I will use the term "registration" interchangeably with "normalization." To register or normalize an image is to determine a "mapping," a "warping," or a "spatial transformation" from the image to the template.[54] A registration method typically consists of three components: a similarity measure, an optimization measure, and a mapping (Crum et al. 2004). The similarity measure provides a way of assessing how well the image matches the template. (For example, a very crude similarity measure for two images of the same size would be the sum of the difference in intensity values of every corresponding pair of pixels.) An optimization procedure is used to choose a transformation that maximizes the similarity measure. The transformation that is selected is then applied to the image to register the image to the template.

It is common to distinguish between two broad kinds of registration methods: intensity-based and feature-based approaches.[55] Intensity-based approaches employ a similarity measure that assesses the difference in image intensity between the target and the template. Feature-based approaches, by contrast, represent distinct "elements in each of the scans to be matched… includ[ing] functionally important surfaces, curves, and point landmarks. [These] elements are each parameterized and matched with their counterparts in the target scan, and their correspondences guide the volumetric transformation" (Toga 1998, 4). In other words, intensity-

[54] In what follows I will primarily discuss registration to templates, but my conclusions apply equally well to registration to reference spaces.

[55] This distinction goes by many names, including "geometric vs. intensity approaches" (Crum et al. 2004), "model-based vs. intensity-based approaches" (Toga 1998), "label-based vs. non-label-based approaches" (Friston et al. 1995), and "photometric vs. geometric approaches" (Hellier et al. 2003).

based approaches select transformations that make the image look visually similar to the template, while feature-based approaches aim to bring specific landmarks into alignment. Intensity-based approaches have gradually been losing ground to feature-based approaches over the last several decades, partly because the range of usable features has expanded (Ashburner 2012). Early feature-based registration was based on gross macroanatomical landmarks; now, there are feature-based methods that incorporate information about curves, major sulci/gyri, microstructure, and even function (see Section 5.7.2).

I will only be discussing registration methods that map data onto a two- or three-dimensional template with a standard spatial interpretation. (I'll sometimes call this "spatial registration.") There are other kinds of registration, used by non-cartographic approaches, to which my arguments do not apply. For instance, one can align some kinds of neuroscientific and psychological data temporally (Zhang 2017). Another non-cartographic data aggregation method is hyperalignment, which involves projecting individual data into an abstract, high-dimensional space (Haxby et al. 2011). Both alternative kinds of registration fall outside the scope of my discussion.

Philosophers of neuroscience have done a good deal of work on topics related to brain atlases, such as the problem of identifying the brain's parts or functions (van Orden 1997, Klein 2012, Anderson 2014). They have paid relatively little attention to apparently prior questions about how neuroscientific data is aggregated in the first place.[56] Despite its neglect by philosophers, registration has been the object of intense scientific activity. Over the last several decades, scientists have developed an increasingly diverse and sophisticated set of registration

[56] As we will see below, the task of alignment only appears to be prior to the identification of parts in the brain. Registration pluralism implies that selecting an appropriate registration method requires one to pick out certain brain locations as salient, which would seem to require some type of neuroscientific ontology.

methods. The simplest transformations that one can apply to a brain image include translation (moving the image up/down or left/right), scaling (changing the image's overall size), and rotation (rotating the image around the origin). Early registration techniques used these kinds of simple transformations. The original Talairach method, for example, divides an individual brain image into 12 rectangular regions, each of which is individually scaled and positioned using piecewise affine transformations (Toga 1998, Chau and McIntosh 2005). By contemporary standards, this is a very crude procedure. Simple rigid transformations have now been largely replaced by affine and non-linear transformations, which have significantly more degrees of freedom (Crum et al. 2004, Toga and Thompson 2007). Unlike the Talairach method, which can only reposition and rescale the brain in limited ways, newer registration techniques allow neuroscientists to move, dilate, stretch, scale, and rotate brain images in a highly complex fashion.

5.3 Registration Pluralism

There is now a wide array of spatial registration methods available to the neuroscientist who adopts the cartographic approach. Some researchers have expressed concern about this variety of methods (van Essen and Dierker 2007) or implicitly assumed that it is temporary, to be winnowed down over time as we move toward the single best registration method. Such an attitude is arguably implicit, for example, in the literature on the validation and comparison of different registration methods. Many papers that compare multiple methods conclude with an overall recommendation about which method provides "the optimal alignment" or is "the best registration method," without obviously indexing these phrases to a specific neuroscientific context (Klein et al. 2009, Robinson et al. 2014). Improvements to the optimization methods used

to select transformations, the increasing mathematical complexity of mappings, and the inclusion of more features in cost functions may be taken to suggest that we are getting ever closer to the ideal, universally applicable registration procedure. Let's call the view that there is a single best registration method toward which all current methods strive "registration monism."[57]

In what follows, I aim to show that registration monism is mistaken. There can be no universally applicable registration procedure because, given the nature of the brain, different target phenomena require different methods. I intend to defend registration pluralism:

> *Registration pluralism:* There is more than one appropriate way to register a brain to a spatial template. The best registration method depends on the phenomenon under investigation.

Registration pluralism is an analog of pluralism about atlases, a far more visible and popular view. Neuroscientists frequently acknowledge that there is no single best way of partitioning the cortex, and that different atlases may be appropriate for different purposes (Arslan et al. 2017, Dickie et al. 2017). Bohland et al. (2009), for example, argue that, "it is highly unlikely that the neuroscience community will, or even should, adopt a single scheme for partitioning the brain or for labeling its pieces" because "the motivations underlying the construction of one atlas can be different from another" (11). Many express this ecumenical attitude about atlases (Toga and Thompson 2001, Brett et al. 2002, Shattuck et al. 2008, Amunts et al. 2014). Registration pluralism is in the same spirit. Different registration methods, as well as different partitioning schemes, ought to be applied in different scientific contexts.

[57] I will try to show in Section 5.7 that a monistic attitude toward registration is at least implicit in several neuroscientific practices. I hope that this will satisfy the skeptical reader that, even if there are not many scientists who explicitly endorse registration monism, it is still a worthwhile target for criticism.

The plausibility of registration pluralism can be illustrated with a simple example. Consider two hypothetical research projects. The first concerns the relationship between sleep/wake cycles and functional activation in areas of the brain that specialize in language processing. The research question is: how does the BOLD signal in language processing regions change over the course of the day? The second research project examines the effects of mercury on the brain (Azevedo et al. 2012). This researcher asks: do people who have had more mercury exposure have more GABA-A receptor activity in cortex surrounding the calcarine sulcus? Let's imagine that both researchers are committed to using the cartographic approach to aggregate their multi-subject data. My contention is that the two researchers should register their data in different ways. For the first project, the researcher should try to align brain regions thought to be involved in the same language-related functions. For the second, the researcher should try to align the calcarine sulcus and its surrounding brain tissue. Given that brains are slightly different from one another, a transformation that aligns language processing regions will not perfectly align points surrounding the calcarine sulcus. Hence, different registration methods are required in the two cases. Diversity in our neuroscientific projects necessitates registration pluralism.

The following two sections will generalize the reasoning in this simple example to provide an argument for registration pluralism. Section 5.4 will characterize the goal of registration and Section 5.5 will show that the goal cannot be achieved with a single method.

5.4 Homology and the Goal of Registration

The first step of the argument for registration pluralism requires understanding what registration aims to do. Fortunately, neuroscientists who employ the cartographic approach are

quite explicit on this point. Nearly every author who writes about registration claims that the goal of registration is "to maximize the genuine homology of points that are brought into correspondence by the transformations" (Mazziotta et al. 2001, 1301). Or, to put it a different way, "the objective is to warp the images such that homologous regions of different brains are moved as close together as possible" (Ashburner et al. 1997, 350-1). Registration to a template succeeds to the extent that "homologous cortical regions in different subjects have been brought into register by the registration transform" (Toga and Thompson 2001, 4). I see no reason to doubt this consensus, so I take it that the aim of registration is to align homologous locations across brains.

The real challenge is understanding what this means. It may be tempting to think that the homology concept at work here is simply a homology concept from biology applied to cognitive science. I believe this is mistaken: neuroscientists discussing the cartographic approach do not use "homology" in the same way as biologists or philosophers of biology (Brigandt 2002; Wagner 1994, 2014).[58]

The concept of homology originated with Richard Owen (1843), who famously defined a homologue as "the same organ in different animals under every variety of form and function." Owen was a comparative anatomist, interested in how animals from different species seem to instantiate the same "archetypes" (Panchen 1994). Although the concept of homology has changed since Owen, biologists continue to understand homology primarily as a cross-species notion. This is the first reason that a biological interpretation of "homology" in the context of brain registration is inappropriate. Registration is used to align the brains of individuals from the same species. When neuroscientists talk about points in the brain being homologous across people, they are not using "homology" to talk about inter-specific relationships, as a biologist

[58] This is not to say that neuroscientists never use "homology" this way (Liebeskind et al. 2016).

would. Even the concept of serial homology in biology bears no resemblance to the neuroscientist's concept, since it involves the repetition of a part within one and the same animal. Hence, there is no inter-individual but intra-specific sense of "homologous" that neuroscientists could be importing directly from biology.

One might think it would be perfectly natural to apply biologists' homology concept(s) to within-species relationships even if biologists do not. After all, it seems possible to talk about homologies across different dog breeds, even though all dogs belong to the same species. Even granting this, there are other indications that neuroscientists are using "homology" differently from biologists. In contemporary biology and philosophy of biology, a distinction is frequently made between two homology concepts: genealogical homology and developmental homology (Wagner 1989, Brigandt 2002, Ramsey and Peterson 2012). Homologous parts in the genealogical sense share a phylogenetic origin, while homologous parts in the developmental sense are subject to the same developmental constraints and underwritten by common ontogenic mechanisms (Wagner 1994). Neither a genealogical nor a developmental homology concept is reflected in the neuroscientific usage. Very few of the neuroscientists who adopt the cartographic approach are investigating the evolution of the nervous system, so it would be strange to interpret their use of the phrase "homologous parts" in terms of phylogeny. When neuroscientists use the cartographic approach, they do not discuss the evolutionary history of the places in the brain they are attempting to align, nor do they try to establish that the places share a common phylogenetic origin. The genealogical homology concept is therefore ill suited to the context of registration. A developmental homology concept can be ruled out for similar reasons. Neuroscientists do not need to know anything about the development of the brain locations they intend to co-register. Their aim during registration is to align present structure or function across individuals. They make no effort to show that the co-registered locations share a developmental trajectory or are

subject to the same constraints, as would be expected if they were deploying a developmental homology concept.

For these reasons, the phrase "homologous locations in the brain" should not be understood by appeal to biologists' homology concept(s). Instead, neuroscientists should be read simply as using homology as a synonym for sameness. Neuroscientists call places in different brains "homologous" when they are the same. Importantly, however, sameness of brain locations across people cannot be specified tout court because there are many different kinds of sameness that one might be interested in (Goodman 1972). In the hypothetical case above, one researcher cares about points that have the same language-processing functions while the other cares about points in the same position relative to the calcarine sulcus. Sameness of functional capacity is distinct from sameness of sulcus-centered positioning. Because sameness is always sameness-in-some-respect, which type of sameness is relevant in any particular scientific context is determined by the target of investigation. Sameness of location in the brain, and therefore homology of brain locations, is purpose-relative.[59]

This should not surprise us since it is just an instance of a general phenomenon. Sameness of location within wholes is purpose-relative whenever the wholes are qualitatively different from one another. Consider a simple inanimate example: imagine there are two similar houses, and a location in one house has been singled out as a point of interest. We might ask the question:

[59] Some biologists have claimed that homology is actually "context dependent" in just this way (Abouheif 1997). Wagner (1994) argues that "homology is a scientific conceptualization of th[e] perception of 'sameness'" (274). He thinks the reason homology has been such a tricky concept is that there are different "aspects" of sameness that are prioritized by different biologists: "the same structural organization, the same developmental origin, the same developmental constraints, the same (genetic) information, and common phylogenetic origin" (Wagner 1994, 274). If Wagner is right, the analogy between biological and neuroscientific uses of "homology" need not be rejected after all: both are purpose-relative. I thank Aaron Novick for discussion of this point.

where is the same location in the other house? The answer clearly depends on what we are interested in. There are many places in the second house that one could identify as the "same" as the place in the first: the location that is the same absolute distance from the front door; the location that is the same distance, proportionally, from the two ends of the house; the location that contains the same piece of furniture; and so on. Sameness of location in houses is purpose-relative. Brains are no different from houses in this respect.

In line with this, several neuroscientists acknowledge that different types of homologies in the brain exist simultaneously. Mazziotta et al. (2001) explain that "various criteria can be used to define homology" (1301). They distinguish between "anatomical [and] functional homologues" and then proceed to make finer grained distinctions within these categories, arguing that "homologies based on function and cytoarchitectonics are more fundamental to neuroscience…than homologies based on sulcal and gyral anatomy" (Mazziotta et al. 2001, 1316, 1301). Likewise, Uylings et al. (2005) claim that a "critical issue" for any study is "defining the criterion of correspondence, i.e. homology" (424). This implies that different criteria are available. At least some neuroscientists agree, then, that homology in the brain is purpose-relative. In what follows, I use the term "homologous" to mean sameness-in-some-respect, in keeping with this standard scientific usage.

5.5 Organizational Variation and Failures of Simultaneous Alignment

We have now seen that the goal of registration is to align locations across brains that are the same. Furthermore, there are different ways for brain locations to count as "the same." By

itself, this does not establish registration pluralism. To understand why, let's return to the house analogy. Imagine that that an architect is trying to register two houses' blueprints to a single template. If the houses were built by the same developer in a cookie-cutter suburban neighborhood, they may be essentially identical. In this situation, it may be possible to register the blueprints to the template in a way that preserves all the homologous relationships the architect cares about. A single transformation will be sufficient to align points that are the same distance from the front door, locations with the same furniture, walls with the same load-bearing capacity, rooms with the same practical functions, and so on. Registration, for the architect, need not be sensitive to the feature of interest.

Human brains are too different from one another for this to work in neuroscience. Spatial alignment of one kind of homologous brain location will not align homologous locations of all other types. A registration method that brings the sulci and gyri of two different brains into alignment, for instance, will not perfectly align cytoarchitectural regions, and vice versa. This is because the brain's organization is variable across people: different types of regions do not stand in constant spatial relationships to one another. For example, the position of one person's occipitotemporal sulcus relative to his Wernicke's area may not be the same as the relative position of another person's occipitotemporal sulcus and Wernicke's area. Consequently, by aligning two subjects' occipitotemporal sulci, you may not succeed in aligning their Wernicke's areas, and vice versa. Organizational variation prevents there being a single way of spatially aligning all homologous brain regions at once.[60]

---

[60] As some authors have noted (Klein et al. 2009), a neuroscientist using the cartographic approach has to assume that the locations he is attempting to align are present in all brains. This assumption might not always be justified. For example, Nieto-Castañón and Fedorenko (2012) argue that it is unlikely that exact correspondences exist between ocular dominance columns in V1 of different individuals. They claim that it would be a mistake to try to co-register such fine-grained functional regions. When is it safe to assume

Evidence for this claim comes from studies showing individual differences in the relative positioning of different types of brain regions. First, it is well known that macroanatomical features like sulci and gyri are variably positioned relative to cytoarchitectural boundaries (Uylings et al. 2005, Amunts et al. 2007). Amunts et al. (1999) demonstrate this variation in their classic examination of Brodmann's areas 44 and 45. They use a computerized technique to identify cytoarchitectural borders on stained brain sections and then compare the cytoarchitectural regions identified with macroanatomical landmarks. The authors find significant inter-individual differences in the location of cytoarchitectural regions relative to macroanatomical features: "one and the same cytoarchitectural border was located in a sulcal fundus in some [individuals'] hemispheres but on one or the other wall of the sulcus or at the top of the gyrus in others" (*ibid*., 335). Scheperjans et al. (2008) reach the same conclusion about superior parietal cortex. Using a similar technique, they find that "the locations of [cytoarchitectonic] borders are not reliably associated with macroanatomical landmarks" (*ibid.*, 2152).

There is also considerable evidence that macroanatomy and function are not predictably related to one another. For instance, Watson et al. (1993) examine intersubject variability in the location of V5, a visual motor area, in relation to sulcal and gyral patterns. They define V5 functionally by comparing PET data collected while subjects saw a moving or stationary checkerboard. They find that the position of V5 can "vary by as much as 27 mm in the left hemisphere and 18 mm in the right" relative to macroanatomical features (*ibid.*, 79). Others have stressed that although some functional areas, like the frontal eye fields, are strongly related to

---

that the same location exists across brains, and that it is therefore appropriate to apply the cartographic approach? This is an important question that deserves more attention.

macroanatomy, others, like the fusiform face area, are not (Frost and Goebel 2012). The divergences are emphasized by critics of the cartographic approach, who argue that registration methods based on anatomical features often fail to align functional regions of interest (Fedorenko and Kanwisher 2009, Nieto-Castañón and Fedorenko 2012).

What of the relationship between cytoarchitectural and functional regions? It is usually thought that functional differences between brain regions are underwritten by cytoarchitectural differences, and hence that there is a close correspondence between cytoarchitectural and functional areas. If this is correct, cytoarchitectural regions do generally stand in constant spatial relationships with functional regions; indeed, they are coextensive. However, not all functional divisions are marked by changes in cytoarchitecture. Weiner et al. (2017) examine the relationship between cytoarchitectural regions and functional regions in human ventral temporal cortex (VTC). As predicted, they observe that face- and place-selective regions in VTC have different cytoarchitectural properties. They also find, however, that there is a "many-to-one mapping" between functional regions of interest (fROIs) and cytoarchitectural regions (cROIs), with several fROIs contained within a single cROI (*ibid.*, 155).

Hence, there is considerable empirical evidence that macroanatomical, cytoarchitectural, and functional brain regions do not stand in constant spatial relationships with one another. This indicates that organizational variation in the brain prevents the simultaneous spatial alignment of every type of homologous region. Given that registration aims to align homologous regions, different registration methods should be used in different scientific contexts. Registration pluralism follows. Note that this argument is not based on the shortcomings of our current methods. That there is no way to spatially align all homologous brain regions at once is a consequence of the nature of the brain, not the limits of our current tools for studying it. As such,

registration pluralism captures a permanent feature of the cartographic approach rather than a temporary obstacle.

The importance of purpose-relative registration is supported by work aimed at comparing different registration methods. Crivello et al. (2002), for example, register the same fMRI and PET data to a Human Brain Atlas template using four different registration methods and then assess the methods' success using several different metrics. They calculate the degree of spatial overlap between the template and the normalized individual MRI volumes for grey matter, white matter, and cerebrospinal fluid (CSF). Two of the normalization procedures Crivello et al. use are the procedure implemented in the 1996 Statistical Parametric Mapping software (SPM), a popular software package for neuroimagers, and a multi-grid technique based on Navier-Lamé continuum mechanics theory (FMG). The details of the SPM and FMG methods do not concern us. What matters is that they excelled in different respects: the FMG method was the best of the four methods at aligning anatomical landmarks, while the SPM method was best at CSF alignment (*ibid.*, 237). None of the methods dominated the others.

Although the authors do not discuss the implications of this result (indeed, it is not even highlighted as an important finding), it lends support to registration pluralism. It suggests that if a researcher is interested in using the cartographic approach to answer a question about CSF, she ought to use SPM; if the alignment of macroanatomy is more important given her project, she ought to use FMG. Surprisingly, despite the obvious way in which their findings support purpose-relative registration, Crivello et al. do not endorse registration pluralism. At the end of their paper, they argue that FMG is the "normalization procedure providing the highest degree of accuracy" and recommend that researchers adopt it (*ibid.*, 248). They seem to think that different registration techniques are simply better or worse at providing "the most accurate brain," ignoring the fact that accuracy is purpose-relative (*ibid.*, 248).

Papers like this one also show that the errors introduced by failing to register one's data in a context-appropriate way can be large enough to scuttle a statistical analysis. One might have thought that neuroimaging is so noisy that minor failures of alignment produced by use of a general-purpose registration method would not make a difference to one's ability to obtain statistically significant results. But the literature on the validation of registration methods suggests otherwise. It is not uncommon for researchers to compare different methods using statistical tests on data that have been registered in different ways. For instance, Nenning et al. (2017) perform a group-level activation analysis of task-based fMRI data that have been registered with two different methods. They find that one method results in central regions with higher *t*-values, meaning that it allows more sensitive region detection than the other. Such results show that the use of context-appropriate registration methods can increase statistical power and thus appreciably improve experimental outcomes, despite the noisiness of neuroimaging.

There are some neuroscientists who, unlike Crivello et al. (2002), embrace registration pluralism by acknowledging that different projects demand different mappings (Friston et al. 1995, Evans et al. 2012). Hellier et al. (2003) claim that "the 'ideal' transformation surely depends on the application" (1120). Crum et al. (2003) agree that "the kind of correspondence, the manner of achieving it, and the acceptable accuracy are application dependent…the scientific question defines the kinds of correspondence that *should* be sought" (1434). And Dubois and Adolphs (2016) recommend "that investigators try more than one approach to alignment, and report all of them, so we can see which might work best for which kinds of questions" (427). As we will see in Section 5.7, however, even researchers who endorse registration pluralism in the abstract may not have fully considered its potential methodological implications.

5.6 The Scope of Registration Pluralism

I have argued that the goal of registration in the cartographic approach is to align homologous locations across brains, but that individual differences make it impossible to achieve this goal with a single registration method. When I offered empirical evidence of individual differences, I focused on macroanatomical, cytoarchitectural, and functional regions because they provide paradigmatic examples of the kinds of homologous locations to be aligned in different research contexts. Registration pluralism would be of limited interest, however, if it held only across these broad categories. A critic might grant that researchers interested in macroanatomy need to use different registration methods than those interested in function, but insist that all macroanatomists should use the same method. The same could be said of all functional researchers and all cytoarchitectural researchers. None of the evidence discussed above entails that registration has to be purpose-sensitive even within broad research areas.

This objection hints at a broader question about the scope of my view. I have so far spoken of registration pluralism as an all-or-nothing thesis asserting that, given the nature of the human brain, there cannot be single registration method appropriate for all neuroscientific projects. So conceived, registration pluralism is true. However, such a thesis is of limited utility by itself. What we really want to know is which projects require different registration methods and which do not. After all, the claim that no method will suffice for all purposes doesn't imply that there are no two purposes for which a single method will suffice. So, just how pluralist should we be about registration? A pithy answer is: as pluralist as the evidence requires. The more variation there is in the positioning of different types of brain locations relative to one another, the greater the variety of registration methods that are needed.

There is in fact empirical evidence of organizational variability even within the three major brain modalities, implying that the scope of registration pluralism is wider than the critic above suggests. Much of the research cited in Section 5.5 in support of the idea that cytoarchitectural and macroanatomical regions do not stand in constant spatial relationships also shows that different cytoarchitectural regions are variably positioned relative to one another. In their microstructural study, Scheperjans et al. (2008) find that which cytoarchitectural regions border one another is different across brains. For example, area hIP3 only borders hIP1 in half of the hemispheres they examine. They conclude that "a considerable number of cytoarchitectonic borders are not present in every brain" (*ibid.*, 2152). It is also well-known that the three-dimensional size and shape of cytoarchitectural regions differ between people. This suggests that a scientist needing to align, say, Brodmann's area (BA) 17 across subjects may require a different registration method than a scientist needing to align BA44. Moreover, there are sub-regions within cytoarchitectural areas whose relative positioning varies. Amunts et al. (1999) describe substantial variability within BA44 and BA45, including lamina whose distribution patterns differ substantially between people. This evidence suggests that different registration methods are needed for research concerning different cytoarchitectural features.

A similar conclusion is arguably true of functional research as well: different kinds of locations count as functionally homologous, and the relative positioning of different functional homologues is such that they cannot all be simultaneously aligned. The reasons for this are somewhat more theoretical. First, given the assumedly tight relationship between cytoarchitecture and function, if different types of cytoarchitectural regions cannot be simultaneously aligned, as suggested above, the same is likely true of different functional regions. Second, recent theorizing about neural reuse supports the idea that the functional organization of the brain precludes the use of a single registration method for all functional purposes. The

neural reuse hypothesis states that "individual neural elements (at multiple spatial scales) are used and reused for multiple cognitive and behavioral ends" (Anderson 2016, 1). Anderson (2014, 2016) argues that each brain region has a functional "fingerprint" or "profile" but can be recruited for a diverse array of tasks. The brain is constantly re-organizing and functional partnerships between brain regions are not fixed. Which regions are recruited for a particular task depends on current activation patterns and other functional demands. Anderson's dynamic, fragmented view of the brain's functional organization supports the idea that locations that are functionally "the same" may be highly variable, both across time and across people. Such neural flexibility and complexity make it impossible to align all functionally homologous points at once.

All of this suggests a wide scope for registration pluralism. The critic imagined above conceded that different neuroscientific contexts require different registration methods, but claimed that contexts should be individuated quite coarsely: we need only differentiate between a "cytoarchitectural context," a "macroanatomical context," and a "functional context," because within each of these, a single registration method suffices. The evidential and theoretical considerations raised here, however, suggest that contexts need to be considerably more fine-grained. I have not specified the level of grain precisely, partly because to do so would be premature. We understand relatively little about individual differences in the spatial positioning of the brain's different features. Our ideas about which projects require different registration methods and which don't must therefore be continually refined in light of new empirical findings.[61]

[61] This section has focused on questions about the scope of registration pluralism with respect to the granularity of neuroscientific contexts. But scope questions are also spatial: it is possible that registration pluralism holds with respect to some brain regions but not others (namely, those that are more variable across individuals).

5.7 Potential Methodological Implications of Registration Pluralism

Defending registration pluralism in principle is easier than figuring out what it means for neuroscientists in practice. The challenge is not just to delimit its scope, but also to determine the broader impacts of the methodology it seems to recommend. The use of different registration procedures in different scientific contexts might well increase alignment accuracy, but it could also have adverse effects on other scientific desiderata. Deciding whether such trade-offs are worth making requires data- and simulation-driven assessments of various methodological approaches along multiple dimensions. In this section, I'll gesture at three potential methodological consequences of registration pluralism that are deserving of further study.

5.7.1 Purpose-Sensitive Selection of Registration Methods

First, registration pluralism seems to imply that researchers who adopt the cartographic approach should select a registration method in a purpose-sensitive manner. In practice, this is rare in studies with neurotypical individuals.[62] Neuroscientists usually do not justify their choice of registration method in print, and when they do, it is in general terms that do not engage with the specific features of their research question. They may, for example, explain that the registration method was chosen because it came as a default in a software package or because it was not computationally demanding. Registration pluralism suggests that neuroscientists who adopt the cartographic approach should explicitly identify the locations they are hoping to align

[62] There is more discussion of registration in research on aging and non-neurotypical populations (Ganzetti et al. 2018).

and then select a registration method likely to align those homologous locations (Crum et al. 2003). There are several worries one might have about this recommendation: one could object that choosing a registration method in a purpose-sensitive manner is not practically feasible; that it presents an obstacle to comparability across studies; or that it could be a source of bias. While all three are serious concerns, I believe they do not decisively undermine purpose-sensitive registration.

The first objection is that it is too much to ask of researchers that they tailor registration to their research question. Neuroscientists, on this view, have too little information about the performance of different methods to make a purpose-sensitive choice. Luckily, this is not the situation that neuroscientists find themselves in. Researchers can choose a registration method by consulting the rapidly expanding literature on the validation of registration methods, which I discussed briefly in Sections 5.3 and 5.5 (Woods et al. 1998, Crivello et al. 2002, Hellier et al. 2003, Crum et al. 2004, Ng et al. 2009, Klein et al. 2009, Conroy et al. 2013, Robinson et al. 2014). As many authors have noted, it is usually impossible to directly assess how well a registration method aligns homologous regions because there is no "gold standard" against which to compare (Woods et al. 1998, Brett et al. 2002, Gholipour et al. 2007). It is, however, possible to use indirect measures of evaluation. Many of these metrics measure some dimension of accurate alignment (Gholipour et al. 2007). Klein et al. (2009), in one of the most comprehensive validation efforts, compare fourteen registration methods along eight different dimensions. Since they are interested in anatomical alignment, they use manually labeled structural images as a "silver standard" for comparison. For each of the fourteen methods, they compare the registered images of individual subjects with the manually labeled images. They measure volume overlap agreement (three different measures), volume overlap error (two measures), surface overlap agreement (one measure), volume similarity (one measure), and distance error (one measure)

between the source and target images. The data show that the fourteen registration methods perform differently on these eight metrics.

We ought to think of such papers as providing information for choosing registration methods in a purpose-sensitive way rather than identifying the all-around best method. The variety of evaluation metrics on offer permit an individual neuroscientist to select the method that does the best on the metrics that are most relevant to his project. Which locations he needs to align will determine which metrics are important. At least some of the authors working on validation do seem to think about their findings in this way. Conroy et al. (2013), for example, who propose a new registration algorithm and compare it to two alternatives, claim that their method is especially good at aligning prefrontal regions (which are particularly difficult to co-register), and should therefore be used in studies of social cognition.

Second, one might worry that the use of different registration methods by different researchers presents an obstacle to comparing results across studies. This concern is present in the literature. Van Essen and Dierker (2007) claim that "unintended biases may be introduced when comparing datasets registered by different algorithms to different templates" (1052). They argue that, to provide "apples-to-apples comparisons," researchers should use simple linear registration methods (*ibid.*, 1052). Brett et al. (2002) similarly claim that, "if we have used a different template or a different normalization method, then…meta-analysis might have low spatial resolution and power" (248). These are reasonable worries, especially concerning templates. There is a trade-off between selecting a template that fits the population at hand and choosing one that permits easy generalization and cross-study comparison (Evans et al. 2012). However, the problem is less acute in the case of registration. When one conducts a meta-analysis, the data one is compiling typically concern the same phenomenon. Since the phenomenon of

interest determines which registration method is appropriate to apply, the data being aggregated for meta-analysis usually should not have been registered in wildly different ways.

Third, one might worry that, given how little we know about the neural basis of many cognitive processes, adopting purpose-sensitive registration will introduce bias into research. On this line of thinking, making registration decisions based on a partial or incorrect understanding of the phenomenon being studied will cause systematic errors of alignment, biasing the results. Using a one-size-fits-all registration method also leads to alignment errors, so the objection goes, but at least they are theory-free or random. Better to introduce random noise than theory-driven bias. I believe this concern, too, is overblown. We are not as ignorant of the brain's functioning as is suggested. To select a purpose-sensitive registration method, we only need some idea of the areas that are homologous given the phenomenon under investigation; we do not need to know exactly how cognitive capacities are realized in the brain. Moreover, it is not the case that the misalignments that occur under a system of purpose-sensitive registration are directional while the misalignments resulting from the use of a single, general-purpose method are random. When everyone uses the same registration method, researchers make similar alignment errors, leading to systematic biases. The status quo of general-purpose registration fares no better than purpose-sensitive registration from the perspective of bias.[63]

Hence, there is reason to think that registration methods should be chosen in a way that is sensitive to the phenomenon under investigation, despite legitimate concerns about feasibility,

[63] An anonymous reviewer raised a related worry: when we apply the cartographic approach in neuroimaging, we want to be able to identify activity we were not expecting to find. Purpose-sensitive registration, with its emphasis on aligning locations we already know to be involved in the function under study, might make this less likely. This is a reasonable concern. However, it is an open empirical question whether the use of a general-purpose registration method will make it more likely that unexpected activation will be uncovered than a method known to excel at aligning at least some of the implicated areas.

comparability, and bias. Entertaining this implication of registration pluralism opens the door to several other possible methodological consequences.

5.7.2 Functional Registration

Functional registration is a type of feature-based registration that uses functional rather than structural features to select a mapping between an individual brain image and a template (Sabuncu et al. 2010, Conroy et al. 2013, Nenning et al. 2017). Most functional registration methods use fMRI data for alignment. In those that rely on task-based fMRI, subjects are presented with a specific stimulus or task while functional data are collected. Researchers then find a transformation that aligns the functional signals from different subjects. Functional registration is a relatively new tool that has generated substantial interest among neuroscientists. Researchers investigating brain function no longer need to register brains to a template using structural information and hope that structure and function correlate. Instead, functional data can directly drive registration. (Though the data driving registration must not be the data to be analysed, on pain of circularity [Sabuncu et al. 2010, 139]).[64]

Registration pluralism warns us to watch out for a potential complication: it may be a mistake to think that a single task-based functional registration method can serve all functional purposes. I suggested above that there might be different kinds of functional regions that cannot be simultaneously aligned. If so, different functional registration methods could be needed in

[64] Some functional registration methods begin by aligning images with a simple anatomical registration procedure before fine-tuning the alignment with functional information (Sabuncu et al. 2010, Conroy et al. 2013). Others are explicitly "multi-modal": they use functional and structural information simultaneously (Robinson et al. 2014). It is interesting to note that the registration methods called "functional" typically rely on other types of information as well.

different scientific contexts. However, most current methods are intended to be generally applicable. In pioneering work on functional registration, researchers asked subjects to watch a full-length action movie, *Raiders of the Lost Ark*, while they collected fMRI data to be used for alignment (Sabuncu et al. 2010, Haxby et al. 2011, Conroy et al. 2013). The researchers explained that they chose a movie, a "complex and dynamic natural stimulus," because it sampled a "diverse variety of representational states" (Haxby et al. 2011, 411) and because "neural activity during a movie viewing is synchronized across subjects in a large percentage of the cerebral cortex" (Sabuncu et al. 2010, 131).

The discussion in Section 5.6 suggests that trying to find a generic stimulus to be used in a universal task-based functional registration method may be misguided. Given the variable and dynamic functional organization of the brain, a registration algorithm that aligns brain regions involved in passively watching a movie may not align the neural substrates of, say, social cognition. If a researcher is ultimately interested in analyzing brain activity during social interaction, then, she may be best off not using functional data collected during a movie-watching task as the basis for registration. Instead, registration pluralism seems to suggest that functional registration should be purpose-sensitive. Researchers ought to consider aligning brains using fMRI data collected while subjects are performing a task that is related to the domain or phenomenon of interest. This would generate functional data that could bring the areas that matter into alignment. The researcher interested in social cognition, for example, could base her registration method on functional data collected while subjects engage in a social task to ensure that areas essential to social interaction are aligned.[65] There is at least one example in the literature

[65] Some of the researchers who use a movie stimulus for functional registration recognize that it may not be appropriate for all scientific contexts. Sabuncu et al. (2010) explain that "it is possible that function-based normalization based on neural activity evoked by more controlled experiments could be more effective for

of (something resembling[66]) purpose-sensitive functional registration. Langs et al. (2010) aim to identify and ultimately align brain regions involved in language-processing across tumor patients. They propose a registration method based on fMRI data collected while subjects are engaged in antonym generation, a language task. Hence, they ask subjects to perform a task that produces activation in the areas they need to align and then use the data collected for registration. It may be fruitful for other researchers to follow their example.

Alternatively, one might argue we should turn away from task-based functional registration altogether. Some authors have claimed that task-based methods rely on the implausible assumption that different people's brains are performing the same functions at precisely the same times (Jiang et al. 2013). They argue that we ought instead to use resting-state fMRI to build a functional connectivity profile for each subject and then find transformations that bring the individual connectivity patterns into alignment (Jiang et al. 2013, Zhou et al. 2017, Nenning et al. 2017, Chen et al. 2017). I am somewhat skeptical of these methods, partly because I share others' doubts about what resting-state fMRI really tells us (Buckner et al. 2013, McCaffrey and Danks forthcoming), and partly because functional connectivity-based methods are meant to be general-purpose. However, such qualitative considerations are far from decisive. A resolution awaits systematic quantitative comparison of connectivity-based methods with purpose-sensitive task-based methods of the kind I have proposed.

---

a specific functional region… The key question is whether a single, optimal warp exists for the cerebral cortex or for sectors of the cerebral cortex – or will overlapping topographic maps for different functions be aligned optimally by different warps" (138). They are betting on the former and I am betting on the latter.

[66] The example is, strictly speaking, not an instance of the cartographic approach because Langs et al.'s (2010) technique does not use a standard spatial reference frame.

5.7.3 Standardized Preprocessing Pipelines

Finally, registration pluralism may undermine projects aimed at standardizing the preprocessing of neuroscientific data. There is currently concern among neuroimagers about the lack of uniformity in early data processing. They worry that preprocessing is usually ad hoc, making neuroimaging less replicable (Esteban et al. 2019) or generalizable (Gabard-Durnam et al. 2018), and forcing individual labs to "repeatedly reinvent the wheel" (Freeman 2015, 156). Such concerns have led to efforts to construct standardized preprocessing protocols for different imaging modalities (Bigdely-Shalmo et al. 2015, Gabard-Durnam et al. 2018). One example is fMRIPrep, a preprocessing workflow recently introduced by Esteban et al. (2019) to provide "robust and reproducible preprocessing" for data from task-based or resting state fMRI (111). fMRIPrep involves a spatial normalization step in which individual data are registered to an ICBM template using the ANTs registration procedure. Esteban et al. claim that fMRIPrep is "analysis-agnostic" in the sense that it supports a wide range of analysis and works across many kinds of input datasets (*ibid.*, 112).

Registration pluralism suggests that any preprocessing pipeline that includes spatial registration is not analysis-agnostic. Which registration method is appropriate depends on the phenomenon of interest, and hence on the analysis to be performed. Uniformity of preprocessing therefore comes at a cost: sometimes different projects really do call for particularized preprocessing steps. Some researchers recognize this. Freeman (2015), for instance, stresses the importance of balancing "standardization and scalability with…flexibility and interactivity" (157). Registration pluralism seems to imply that registration should be highlighted as a locus of deliberate choice in standardized workflows. Researchers using tools like fMRIPrep ought to be encouraged to think about which registration method suits their purposes best.

5.8 Registration Pluralism and the Study of the Brain

The previous section suggested that registration pluralism may have implications for how scientists select registration procedures, what tasks are used in functional registration, and which parts of preprocessing can be standardized. These issues, I claimed, are deserving of more systematic analysis by neuroscientific methodologists. Zooming out from such concrete methodological questions, we can also inquire into the broader significance of registration pluralism and the empirical evidence of variability on which it is based. What does registration pluralism mean for neuroscience in general?

First, registration pluralism represents an obstacle to the transfer of evidence across contexts. Data that have been collected and registered to a template in one way, with one purpose in mind, may need to be re-registered in order to apply to a new research question. This is because the initial registration may have aligned a different set of features than those requiring alignment for the new project. This is resonant with current philosophical work discussing the difficulty of sharing scientific data across epistemic contexts (Leonelli 2016, Boyd 2018). To use Boyd's (2018) terminology, empirical results may be maladapted to a scientific theory, meaning that they cannot serve as a constraint on that theory because of how data were collected or processed. However, sometimes data can be repurposed and become well adapted to theories to which they were initially maladapted. According to Boyd, this repurposing is accomplished (in part) using "workflow metadata," that is, auxiliary information about the data processing that had been performed. Information about the registration procedure initially used to aggregate neuroscientific data is a kind of workflow metadata that is important to preserve and deploy when we want to repurpose neuroscientific results. Just like in the sciences that Leonelli (2016)

and Boyd (2018) discuss, there is danger in transferring neuroscientific data from one context to another without taking account of how they were produced.

The argument that supports registration pluralism also suggests that there is a type of theory-ladenness in data aggregation that is often overlooked by philosophers of science. One might have thought that gathering all of one's data in the same place once it has been collected is relatively straightforward. But in neuroscience, this apparently simple task can rely on prior information about localization, information that counts as "theory" on a minimal understanding of that term. When we deploy the cartographic approach, best practice requires us to know something about which locations matter given our explanatory target, or so I have argued. Since our goal is accurate alignment, the conditions of success for aggregation depend on the target of investigation. Aggregation succeeds when the locations we care about are aligned, but to accomplish this, it is helpful to have background neuroscientific theory about which locations those are. As I pointed out in Section 5.7.1, this doesn't require an extensive understanding of the brain's workings, but it does make data aggregation using the cartographic approach somewhat theory-laden.

The theory-ladenness of the cartographic approach strengthens an existing philosophical argument about the epistemic status of neuroimages. Pushing back against the hype surrounding neuroimaging, several authors have claimed that neuroimages are quite unlike standard photographs. Klein (2010) highlights two primary dissimilarities: unlike ordinary pictures, neuroimages are laden with theoretical assumptions and present the results of statistical tests rather than raw data. Klein explains that neuroimages cannot be interpreted without understanding the experimental design that produced the data and the specific tasks that the subjects performed. Roskies (2007) calls this property "belief-opaqueness": neuroimages are belief-opaque because "the information needed for the[ir] interpretation is not present in the

resultant image" (868). Registration pluralism reinforces these claims. Neuroimages depicting aggregate results are theory-laden in an additional way not discussed by Klein (2010) or Roskies (2007), since aggregation itself is another substantive and theory-laden data processing step involved in visualizing neuroscientific data. Moreover, one cannot tell from looking at an aggregate neuroimage how the data have been registered to the template. This is yet another way in which neuroimages are belief-opaque.

An interesting philosophical question is whether similar conclusions hold in other domains. Aggregation of data would seem to require additional input whenever it isn't obvious which data points count as "the same." One (but by no means the only) source of such ambiguity, as I have shown, is individual differences. Variation complicates the task of figuring out which observations or measurements are directly comparable across subjects. Data aggregation may therefore be purpose-driven and theory-laden in interesting ways throughout the psychological and social sciences.

5.9 Conclusion

This chapter has aimed to characterize, defend, and explore the implications of a pluralist view about the cartographic approach to aggregation across brains. Because registration aims to align homologous regions, but not all homologous regions can be simultaneously aligned, there can be no single spatial registration method that suffices for all neuroscientific purposes. A few neuroscientists endorse the thesis of registration pluralism outright (Hellier et al. 2003), some acknowledge it implicitly (Dubois and Adolphs 2016), but at least a few seem to reject it (Robinson et al. 2014). Even those who recognize the truth of registration pluralism may not have fully

grappled with its methodological and philosophical implications, which I began to explore. While Chapters 2 and 3 showed that modeling must be approached differently if one hopes to explain individual differences, registration pluralism demonstrates that variation needs to be taken into account even when it is not the investigative target. The methodological consequences of variation are felt far outside the narrow bounds of individual differences research.

Neuroscience has undoubtedly benefited from its increasing reliance on multi-subject studies in addition to single, pathological cases: it has been able to study normal functioning in non-clinical populations, enjoyed an increase in statistical power, and produced findings that are more generalizable. But the field is still wrestling with the additional challenges posed by the aggregation of data across brains. As I hope to have shown here, reflection on the strategies used for aggregation is needed to appropriately deploy neuroscientific tools in multi-subject research and fully understand the results they produce.

6.0 Concluding Remarks

6.1 An Existential Threat to Cognitive Science?

It is tempting to see individual differences as deeply threatening to the project of cognitive science. The core scientific aims of explanation, prediction, and manipulation all require an ability to generalize. One might worry that variation, an obstacle to generalization, prevents us from uncovering general principles about how our minds work or predicting and influencing the behavior of people we haven't yet encountered. Moreover, as we saw in Chapter 1, individual differences have a track record of getting in the way of scientific progress. Variation has led to theoretical missteps in the study of skill and memory (Estes 2002), and today it jeopardizes ambitious scientific initiatives like the Human Connectome Project (Sporns 2013). Thus, individual differences can lead to pessimism about the prospects of cognitive science: if we are so different from one another, perhaps a scientific understanding of the mind and brain is unattainable.[67]

This dissertation has shown that such worries are misplaced. Although individual differences do throw a wrench in data aggregation (Chapters 1 and 5) and standard modeling approaches (Chapters 3 and 4), cognitive scientists can adapt to the challenges that variation poses by modifying existing methods and developing new ones. I have tried to constructively

[67] For example, an anonymous reviewer for Chapter 5 suggested that individual differences in the brain might "make it practically impossible for us to fully understand the working of the mind."

discuss what a few of these modifications (e.g., to rational analysis, the cartographic approach) might involve, though there is much work left to be done.

The idea that there is a deep tension between variation and scientific generalization does contain a grain of truth, captured by what I have called the uniformity/uniqueness dilemma. It is true that cognitive science cannot treat every individual as unique, or else it will lose projectability, overfit the data, and fail to formulate explanatory generalizations. However, we are not thereby forced to assume that all individuals are the same. Chapters 3 and 4 discussed two approaches to navigating the uniformity/uniqueness dilemma. Hierarchical Bayesian modelers use stochastic parameter models to capture variability in individual parameter values while maintaining a degree of uniformity by drawing those values from overarching distributions. The rational analysis of variation holds everyone to the same rational skeleton while allowing that different individuals might be facing subtly different tasks. The availability of these and other strategies for navigating the dilemma shows that variation is not incompatible with generalization.

Furthermore, a theme running through the preceding chapters has been the continuity between the study of psychological regularities and the study of psychological variation. As a sociological matter, it is undeniable that investigating individual differences has shed light on regularities, and investigating regularities has generated insights into individual differences. This was emphasized by several papers cited in Chapter 1 (e.g. Kosslyn et al. 2002, Sauce and Matzel 2013). It is also supported by research on #TheDress. As mentioned in Section 2.4.2, studies of variation in perceptions of the dress have been informed by research showing that, in general, human color discrimination is poor along the daylight locus (Pearce et al. 2014). Moreover, the color constancy hypothesis owes a huge debt to previous work on color constancy in the visual system, most of which aimed to elucidate general mechanisms, not variation (Olkkonen & Ekroll

2016). Investigation of variation in interpretations of the dress has provided insights into visual system regularities as well. For instance, one factor that has been found to correlate with one's perception of the dress (but which I did not discuss in Chapter 1) is pupil size: subjects who see the dress as white and gold have a lower pupil size when they look at the photograph than subjects who see it as blue and black (Vemuri et al. 2016). Vemuri et al. (2018) show that these size differences are *caused* by differences in subjects' interpretations of the illumination in the image, and hence, that there is "top-down modulation of the pupil size rather than a direct reflectance triggered pupillary light reflex" (B354). If they are right, individual differences research on the dress has revealed an intriguing general feature of the visual system: the presence of top-down influences on pupil dilation.

Chapter 2 further refutes the idea that individual differences fundamentally upend cognitive science in at least two ways. First, it shows that variation can be assimilated to the standard scientific practice of causal explanation. And second, it implicitly establishes overlap between the explanation of regularity and the explanation of variation. SHRINK implies that all of the causes of variation in Y are also type-level causes of Y. This is because any variable that can be intervened on to reduce Var(Y) (while holding fixed the population-level distributions of off-path variable values) can also be intervened on to change the value of Y (under some assignment of values to off-path variables). The fact that causes of variation in Y are a proper subset of the type-level causes of Y means that the explanation of variation in Y is conceptually linked to the explanation of regularities about Y. It also helps to explain why research on variation and research on general mechanisms can be mutually illuminating. Finding a causal explanation of variation in Y can help explain regularities in Y, and vice versa.

One upshot is that a significant proportion of the research that sheds light on individual differences does not go by the name of individual differences research. A researcher may identify

variables that are causes of variation in *Y*, and thus possess the resources to causally explain variation in *Y*, without having set out to study variation. The distinction between research on individual differences and research on psychological and neuroscientific regularities is blurred. Some cognitive scientists recognize this. In their review of individual differences in fear conditioning, Lonsdorf & Merz (2017) explain that, "much evidence on inter-individual differences in fear conditioning processes are [sic] derived from studies whose primary aim was a different research question" (721). A majority of the papers they cite focus on the general mechanisms underlying fear conditioning rather than on variation *per se*. Hence, the relative scarcity of scientific research that explicitly concerns individual differences should not be taken as evidence that variation is incompatible with the aims of cognitive science.

6.2 The Pervasiveness of Variation

Optimism about the prospects for cognitive science is warranted despite the fact that individual differences are truly ubiquitous. Occasionally one encounters the idea that, although there is extensive variability at "higher levels" of the mind or brain, variability is minimal at "lower levels." Sometimes this claim is expressed in terms of a scientific hierarchy: there is more variation in "higher" branches of cognitive science like psycholinguistics and social psychology than in "lower" branches like neurophysiology and perceptual psychology. At other times, the claim is articulated by appeal to a mental hierarchy: we find more individual differences in "higher" human faculties than in "lower" faculties. On this view, there is a significant degree of uniformity among neurotypical individuals in the mind's "building blocks," which include things like our sensory apparatus, basic neural resources, and perceptual processing. Variation

arises when the building blocks are put together in different ways (and perhaps and with more and more environmental input), leading to greater variability in our "higher" faculties.

I am skeptical of both formulations of the idea that individual differences are limited to "higher levels" of the mind or brain. Even granting that there is an intuitive hierarchy of the cognitive sciences or mental faculties (which is doubtful), this dissertation has shown that variation is no less important at lower "levels" of such hierarchies. Chapters 2 and 5 pointed to the presence of variability in the putatively low-level domains of color perception and neural organization. #TheDress is a vivid illustration of the variation that can arise from color constancy mechanisms. A number of recent papers have called attention to widespread individual differences in color vision and their many sources, from physiological variability in the retina to variability in visual processing in the brain (Webster 2015, Olkkonen and Ekroll 2016, Emery and Webster 2019). Likewise, Chapter 5 cited many studies showing substantial variability in basic brain organization. There are individual differences in both the size and positioning of macroanatomical, cytoarchitectural, and functional brain regions. Much of the variation discussed in the dissertation is thus found in "lower" branches of cognitive sciences.

It might seem intuitive that variation compounds as one ascends the scientific or mental hierarchy: slight variations in psychological and neural building blocks are amplified as they interact and combine. In fact, however, variability at one level sometimes yields uniformity at the next level up. For instance, after surveying the many sources of individual differences in color vision, Emery and Webster (2019) emphasize that, "despite enormous variations both within and across observers, the visual system often maintains a highly stable perceptual experience of color" (28). For example, the ratio of L:M cones in the eye varies tremendously – it can be between 1:1 and 16:1 in individuals with normal color vision (Hofer et al. 2005) – and yet sensory systems usually compensate, producing similar color experiences in people with wildly different cone

ratios. Hence, variability can give rise to uniformity at successive stages of processing or levels of spatial organization. (This idea has been emphasized by biologists and philosophers of science interested in robustness: thanks to stabilizing mechanisms, low-level diversity can give way to stable or uniform system-wide behavior.) The compounding of variation through the levels of the mind or brain is not an inevitability.

A final reason to reject the idea that variation is more widespread at "higher" levels of the scientific or mental hierarchies, or at least to be agnostic about it, is that we remain largely ignorant of the extent of individual differences in most areas of psychology and neuroscience. Not only have cognitive scientists not focused on measuring or documenting so-called normal variation, many of their methods are not good at uncovering it. As Cronbach (1957) emphasizes, experimental psychologists often deliberately minimize inter-subject variation, including by adopting methods to correct for luminance sensitivity in color experiments and developing cognitive tasks that produce largely uniform behavior (see Section 1.1). Hedge et al. (2017) make a compelling argument that the latter has limited our ability to study, and presumably also assess the overall magnitude of, individual differences in cognition (see Section 1.2). Furthermore, there is reason to suspect that variation has been systematically underestimated in the lower levels of the cognitive sciences, since it is precisely in domains like perceptual psychology and neurophysiology where heterogeneity has been under the strictest control. It is therefore highly premature to assume that the prevalence of individual differences increases through successive "levels" of the mind or brain.

6.3 Future Directions

This dissertation has discussed just a handful of the philosophical issues raised by individual differences in cognitive science. In closing, I'll describe three additional topics that merit further study.

6.3.1 When Can We Ignore Variation?

In light of the ubiquity of variation, a question that looms over all of cognitive science is when individual differences can be safely ignored and when they must be reckoned with. This is a question that is partly philosophical, partly scientific, and partly mathematical. No single answer is available, as it depends not only on the researcher's goals and object of study, but also the investigative context: the conditions under which variation can be ignored in experimental design, data analysis, modeling, and explanation are different.

The account of explanation proposed in Chapter 2 makes some inroads with the question of when we can discount individual differences in the context of causal explanation. SHRINK suggests that one should not seek to explain variation that is noise, i.e., heterogeneity that cannot be reduced by intervening on any of the variables in one's variable set. Recall that one's variable set is constituted by all of the potential variables that one is willing to countenance. Under SHRINK, non-reducible heterogeneity cannot be causally explained (given the variables one is considering), and so is best set aside for the purpose of causal explanation.

Questions about when variation can be ignored during data analysis are considerably more complicated. As discussed in Section 1.2, some kinds of variation threaten the use of averaging. Even if a pattern of variability constitutes noise (in the sense just discussed), it cannot

always be discounted because averaging over it can lead to mistaken inferences (Estes 1956). There are many papers in the scientific literature that provide guidance about when it is safe to draw conclusions about individuals from averaged data, a largely mathematical issue (Estes and Maddox 2005). Myung et al. (2000), for example, describe three conditions that are mathematically necessary for the power law artifact to occur. They also provide positive recommendations as to how the artifact can be avoided. Of course, Myung and colleagues' work applies only to a particular kind of data and a particular kind of distortion. A great deal more research is required to delineate the contexts in which data analysis can proceed without regard to variation.

The issue of when to incorporate variation in psychological or neuroscientific models is also a topic of ongoing discussion among researchers. Interestingly, hierarchical modeling (discussed in Chapter 3) serves as the foundation for one recent set of guidelines. Bolger et al. (2019) are concerned with causal effect heterogeneity, or variation "in the size and/or direction of cause-effect links" (601). To determine whether causal effect heterogeneity in an experiment "matters," they propose that researchers build a stochastic parameter model in which the experimental effect is represented by a first-order parameter whose values are drawn from an overarching normal distribution (*ibid.*, 608). The "noteworthiness" of the causal effect heterogeneity can then be assessed by examining the absolute and relative magnitudes of the hyper-parameters, and by comparing the fit of the stochastic parameter model to a model with no variation (*ibid.*, 606).[68] Bolger and colleagues' concrete guidance about when variation matters is a valuable contribution to the literature and a creative use of hierarchical modeling.

[68] Consider experimental effect β, which has different magnitudes for different individuals. In a stochastic parameter model, individual values of β are drawn from an overarching normal distribution with mean μ_β and standard deviation σ_β. Bolger et al. propose three criteria for assessing variation in β: (i) *uncertainty*

Articulating the conditions under which variation can reasonably be ignored falls primarily to scientists. Methodologists in psychology and neuroscience have their work cut out for them, since guidelines for discounting individual differences will often need to be tailored to local research contexts. Nevertheless, philosophers too may have a role to play, particularly in drawing out the implications of philosophical accounts of science (as in the SHRINK example above); distinguishing investigative contexts in which different guidelines apply; and calling attention to the ways in which those guidelines are shaped by goals and values.

6.3.2 Kinds as Clusters? The Role of Variation in Classification

The second set of issues concerns the connection between individual differences and scientific classification. Variability is often seen as an essential guide to genuine kinds. Perhaps the most visible example of this is Richard Lewontin's (1972) essay, "The Apportionment of Human Diversity." Lewontin argues that races are not biologically real because there is more genetic variation within racial groups than between them (cf. Edwards 2003). Lewontin's paper illustrates a kind of argument that is very common in science. When considering the validity of a classificatory scheme, scientists often compare the amount of within-group variability to between-group variability. If the latter is greater than the former, it is taken as evidence that the

---

interval: does the CI_{95} for σ_β include zero? (ii) *comparative model fit*: does a non-hierarchical model with no heterogeneity provide as good a fit as the stochastic parameter model? and (iii) *relative size*: is σ_β less than ¼ of μ_β? Bolger and colleagues claim that an answer of "yes" to these questions suggests that heterogeneity is "absent" or "ignorable" (*ibid.*, 616, 608). (Unlike the modelers discussed in Chapter 3, they use non-Bayesian estimation methods, hence the appeal to confidence intervals.) Criterion (iii) captures the idea that the variability in the experimental effect (σ_β) must be reasonably large relative to the average value of the effect (μ_β) in order to be noteworthy. The ratio of σ_β / μ_β = ¼ is a "rule of thumb" (*ibid.*, 609). As they explain, "there may be cases in which, based on the goals of the research, researchers may decide to apply stricter or more liberal cutoffs" (*ibid.*, 609). In this respect, (iii) is analogous to Cohen's (1988) standard intervals for large, medium, and large effect sizes.

proposed groups are "real." If within-group variability exceeds between-group variability, the groups are taken to be undermined. In addition to race, arguments of this form have been offered to vindicate or undermine categories related to personality, culture, species, and much else.

Such arguments are sometimes framed in terms of clustering: a proposed set of kinds are real if and only if individuals fall into distinct clusters. (Clusters occur when between-group variation is high and within-group variation is low.) The philosophical association of natural kinds with clustering goes back at least as far as Locke (1690), who held that there must be "chasms" between species for them to qualify as real kinds. Because he believed nature contains smooth gradations between kinds, Locke maintained that species were merely conventional types. Broad (1920) expressed a similar view of kinds, imagining a many-dimensional space with axes representing different properties of the states of material things, such as "colour, sound, taste, smell, temperature, shape, [and] size" (24). He argued that, "antecedently there seems no reason why any one of the possible sorts of states should be represented in nature by more instances than any other…But our actual experience of the world has been utterly and flagrantly contrary to this expectation. What we have found is not a regular distribution of all the states…but a 'bunching together' of instances" (*ibid.*, 25). These bunches are natural kinds.

Today clustering and kinds remain closely associated in both philosophy and science. Boyd's (1991, 1999) influential homeostatic property cluster theory claims roughly that individuals belong to a natural kind when their properties contingently cluster together in way that is maintained by an underlying homeostatic mechanism (see also Slater 2015). Many contemporary statistical techniques are committed to a clustering requirement on kinds as well. Cluster analysis is used to identify kinds *de novo* in a variety of fields, from psychology (Clatworthy et al. 2005) to astronomy (Buccheri et al. 1986) to chemistry (Drab and Daszykowski 2014). Biologists, for example, deploy analyses of molecular variance to determine whether a

population contains distinct species (Roca et al. 2001). When there is more genetic variation between groups of organisms than within groups, it is taken to be evidence that the groups are separate species.

There are many open philosophical questions about why and how we associate clusters with kinds. What is the significance of the ratio of within- to between-group variability? Is Lewontin's argument against the reality of racial categories a good one? Should clustering should be considered a necessary condition on kinds or merely a desirable feature? Answering these questions would provide us with a more complete account of the epistemic and metaphysical significance of variation.

6.3.3 Ethics of Individual Differences Research

The final set of issues arises out of the ugly history of research on human variation. For centuries, research on variation has been tightly linked with scientific racism and sexism (Gould 1981/1996, Dennis 1985, Subramaniam 2014). Many fundamental statistical tools were developed to investigate group differences by researchers pursuing deeply racist projects (Louça 2009). One of the most infamous of these statistical pioneers was Francis Galton, father of both the concept of statistical correlation and of eugenics. Galton's studies of variation in "talent and character" were motivated by and fueled his support for selective breeding (Gilham 2001). More recently, research on variation has been institutionalized in the field of "differential psychology," which traces its roots back (in a surprisingly uncritical way) to the "amazing contributions" of Galton (Revelle et al. 2011, 8). Differential psychology has been largely concerned with theories of intelligence and personality, which have been the subject of considerable controversy and moral

criticism (Block and Dworkin 1976, Gould 1981/1996, Pervin 1985, Herrnstein and Murray 1994). It has also tended to analyze differences between groups defined by race and gender (Tyler 1947), a research strategy that may sometimes be inappropriate (Cho 2006, Joel et al. 2015). Hence, although contemporary differential psychology disavows "elitist, racist, or exclusionary" ends, its subject matter is morally controversial and its past is morally dubious (Revelle et al. 2011, 21).

The ethical dimension of individual differences research deserves greater philosophical attention. Most philosophical work in this vicinity has focused on the investigation of group differences, that is, research on variation between sexes, genders, races, cultures, and so on (e.g. Kitcher 2003, Fine 2010, Bluhm 2013). There are less-explored ethical issues, however, that arise outside the context of group differences research. Do moral considerations have a legitimate role to play in scientific choices about how individual differences are characterized or explained? How do arguments against the study of group differences apply to individual differences research (narrowly construed)? Ethical scrutiny can help shape the direction and reception of scientific understandings of human variation.

Appendix A Problems with Actual Difference Making

Chapter 2 argued that an account of the explanation of variation based on Waters' (2007) concept of actual difference making is untenable because there are causes of variation that are uniform in the target population. Even putting this aside, ACTUAL inherits several technical issues with Waters' characterization of actual difference making that have yet to be discussed in the philosophical literature.

Recall that Waters' definition of an actual difference maker is as follows: *X* is *an actual difference maker* (ADM) with respect to *Y* in population *p* if and only if:

(v) *X* causes *Y*.
(vi) The value of *Y* actually varies among individuals in *p*.
(vii) The relationship *X* causes *Y* is invariant over at least parts of the space(s) of values that other variables actually take in *p*.
(viii) Actual variation in the value of *X* partially accounts for the actual variation of *Y* values in population *p* (via the relationship *X* causes *Y*).

Imagine that *Y* represents rate of metabolism. Adults have different metabolisms, variability which is caused by a number of factors. ADMs for metabolism include diet, exercise, and the possession of specific genes. Each of these factors is a cause of metabolism (satisfying [i]) that exerts an influence on metabolism under some of the values that other variables take in the population *p* of human adults (satisfying [iii]). Moreover, there is variation in metabolism (satisfying [ii]) that is partially accounted for by each of the factors just listed (satisfying [iv]). Hence, each of these factors qualifies as an ADM. As we saw in Section 2.4.2, Waters explicates condition (iv) as follows: *X* partially accounts for variation in *Y* when conditions (i), (ii), and (iii) are met, and "An intervention on *X* with respect to *Y* that changed the *X* values in one or more individuals in *p* to the *X* value that one of the individuals had sans intervention would change *Y*

values in *p*" (*ibid.*, 571). I decided to call this kind of intervention a "swapping intervention," since it involves swapping one individual's *X* value for the *X* value of a different individual in *p*. Condition (iv) is satisfied, according to Waters, iff there is a swapping intervention on *X* that would change *Y*.

This quasi-formal characterization of actual difference making is unclear and infelicitous in a number of ways. First, Waters does not specify the conditions under which it applies. For instance, we are not told whether the definition assumes that causation among the variables is deterministic, indeterministic, or pseudo-indeterministic. (Perhaps Waters intends to defer to Woodward [2003], who restricts his account to deterministic causal contexts.) Nor does Waters make explicit his assumptions about the causal structure in population *p*. It seems that he is assuming that there is causal homogeneity in *p,* meaning that all individuals in *p* conform to the same causal model. A related issue is whether condition (i) requires that *X* be a cause of *Y* for *all* individuals in *p*, or only some. Moreover, under certain assumptions, conditions (i) and (iv) are redundant. If there is a swapping intervention on *X* that would change *Y*, then *X* has to be a cause of *Y*.

In addition, Waters sends mixed signals about whether actual difference making is a type- or token-causal notion (or both, or neither). On the one hand, Waters presents Woodward's account of type causation as the foundation for his own view. The definition above also incorporates (something like) Woodward's notion of invariance, which is primarily a property of type-causal claims. This suggests that Waters conceives of ADMs as type-level causes. On the other hand, most of Waters' examples involve token causation. He argues that a particular gene was the ADM in Thomas Morgan's classic experiments on the eye color of fruit flies. In a toy case in which Mary lights a match, he claims that the striking of the match is the ADM. Furthermore, to motivate his view, he briefly argues that Woodward's account of token causation is inadequate,

a move which implies that he is providing an alternative analysis of token-causal claims (Waters 2007, 566). His frequent use of the term "actual" (as in "actual population" and "actual differences") also supports the idea that he is proposing an actual (i.e. singular or token) causal concept. As discussed in Chapter 2, type- and token-causal claims are typically analyzed differently within the interventionist framework. Thus, understanding what conditions (i), (iii), and (iv) require demands greater clarity about the causal concepts they contain.

A final set of issues concern Waters' explication of (iv) and the phrase "partially account for." Condition (iv) requires that variation in *X* partially account for the variation in *Y*, by which Waters means that there must be a swapping intervention on *X* that changes *Y*. This elaboration fails to capture the intuitive meaning of one variable "accounting for" variation in another. By Waters' own acknowledgment, the elaborated formulation of (iv) is satisfied even if *all possible* swapping interventions on *X* with respect to *Y* would increase the total variation in *Y*. As he explains, "eliminating the variation in the value of an actual difference maker might actually increase, rather than decrease, variation in the value of *Y* in *p*" (*ibid.*, 571-2). An implication is that *X* partially accounts for the variation in *Y* even if *X*'s taking the values it does in fact *minimizes* the actual variation in *Y*. In my view, this is out of step with the intuitive meaning of "partially account for."

Bibliography

Abouheif, E., M. Akam, W. J. Dickinson, P. W. Holland, A. Meyer, N. H. Patel, R. A. Raff, V. L. Roth, and G. A. Wray. 1997. "Homology and Developmental Genes." *Trends in Genetics: TIG* 13 (11): 432–33.

Allan, Lorraine G. 1980. "A Note on Measurement of Contingency between Two Binary Variables in Judgment Tasks." *Bulletin of the Psychonomic Society* 15 (3): 147–49.

Amunts, K., M. J. Hawrylycz, D. C. Van Essen, J. D. Van Horn, N. Harel, J.-B. Poline, F. De Martino, et al. 2014. "Interoperable Atlases of the Human Brain." *NeuroImage* 99 (October): 525–32.

Amunts, K., A. Schleicher, U. Bürgel, H. Mohlberg, H. B. Uylings, and K. Zilles. 1999. "Broca's Region Revisited: Cytoarchitecture and Intersubject Variability." *The Journal of Comparative Neurology* 412 (2): 319–41.

Amunts, K., A. Schleicher, and K. Zilles. 2007. "Cytoarchitecture of the Cerebral Cortex--More than Localization." *NeuroImage* 37 (4): 1061-1065-1068.

Anderson, J. R., and C. F. Sheu. 1995. "Causal Inferences as Perceptual Judgements." *Memory & Cognition* 23 (4): 510–24.

Anderson, John R. 1990. *The Adaptive Character of Thought*. Hillsdale, NJ: Psychology Press.

———. 1991a. "Is Human Cognition Adaptive?" *Behavioral and Brain Sciences* 14 (3): 471–85.

———. 1991b. "More on Rational Analysis." *Behavioral and Brain Sciences* 14 (3): 508–17.

Anderson, John R., and Robert Milson. 1989. "Human Memory: An Adaptive Perspective." *Psychological Review* 96 (4): 703–19.

Anderson, Michael L. 2014. *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: A Bradford Book.

———. 2016. "Précis of After Phrenology: Neural Reuse and the Interactive Brain." *The Behavioral and Brain Sciences* 39 (January): e120.

Anderson, Richard B., and Ryan D. Tweney. 1997. "Artifactual Power Curves in Forgetting." *Memory & Cognition* 25 (5): 724–30.

Arslan, Salim, Sofia Ira Ktena, Antonios Makropoulos, Emma C. Robinson, Daniel Rueckert, and Sarah Parisot. 2017. "Human Brain Mapping: A Systematic Comparison of Parcellation Methods for the Human Cerebral Cortex." *NeuroImage*, April.

Ashburner, J., P. Neelin, D. L. Collins, A. Evans, and K. Friston. 1997. "Incorporating Prior Knowledge into Image Registration." *NeuroImage* 6 (4): 344–52.

Ashburner, John. 2012. "SPM: A History." *NeuroImage* 62 (2): 791–800.

Aston, Stacey, and Anya Hurlbert. 2017. "What #theDress Reveals about the Role of Illumination Priors in Color Perception and Color Constancy." *Journal of Vision* 17 (9): 4–4.

Averell, Lee, and Andrew Heathcote. 2011. "The Form of the Forgetting Curve and the Fate of Memories." *Journal of Mathematical Psychology*, Special Issue on Hierarchical Bayesian Models, 55 (1): 25–35.

Azevedo, Bruna Fernandes, Lorena Barros Furieri, Franck Maciel Peçanha, Giulia Alessandra Wiggers, Paula Frizera Vassallo, Maylla Ronacher Simões, Jonaina Fiorim, et al. 2012. "Toxic Effects of Mercury on the Cardiovascular and Central Nervous Systems." *Journal of Biomedicine and Biotechnology* 2012.

Bakan, David. 1954. "A Generalization of Sidman's Results on Group and Individual Functions, and a Criterion." *Psychological Bulletin* 51 (1): 63–64.

Bartlema, Annelies, Michael Lee, Ruud Wetzels, and Wolf Vanpaemel. 2014. "A Bayesian Hierarchical Mixture Approach to Individual Differences: Case Studies in Selective Attention and Representation in Category Learning." *Journal of Mathematical Psychology* 59 (April): 132–50.

Baxter, Janella. 2019. "How Biological Technology Should Inform the Causal Selection Debate." *Philosophy, Theory, and Practice in Biology* 11 (2).

Beatty, John. 1980. "Optimal-Design Models and the Strategy of Model Building in Evolutionary Biology." *Philosophy of Science* 47 (4): 532–61.

Becker, Gordon M. 1991. "The Nonoptimality of Anderson's Memory Fits." *Behavioral and Brain Sciences* 14 (3): 487–88.

Bigdely-Shamlo, Nima, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A. Robbins. 2015. "The PREP Pipeline: Standardized Preprocessing for Large-Scale EEG Analysis." *Frontiers in Neuroinformatics* 9 (June).

Block, N. J., and Gerald Dworkin, eds. 1976. *The IQ Controversy*. New York, NY: Pantheon.

Bluhm, Robyn. 2013. "New Research, Old Problems: Methodological and Ethical Issues in fMRI Research Examining Sex/Gender Differences in Emotion Processing." *Neuroethics* 6 (2): 319–30.

Bogen, James, and James Woodward. 1988. "Saving the Phenomena." *Philosophical Review* 97 (3): 303–352.

Bohland, Jason W., Hemant Bokil, Cara B. Allen, and Partha P. Mitra. 2009. "The Brain Atlas Concordance Problem: Quantitative Comparison of Anatomical Parcellations." *PLOS ONE* 4 (9): e7200.

Bolger, Niall, Katherine S. Zee, Maya Rossignac-Milon, and Ran R. Hassin. 2019. "Causal Processes in Psychology Are Heterogeneous." *Journal of Experimental Psychology. General* 148 (4): 601–18.

Bowers, Jeffrey S., and Colin J. Davis. 2012. "Bayesian Just-so Stories in Psychology and Neuroscience." *Psychological Bulletin* 138 (3): 389–414.

Boyd, Nora Mills. 2018. "Evidence Enriched." *Philosophy of Science* 85 (3): 403–421.

Boyd, Richard. 1991. "Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 61 (1/2): 127–48.

———. 1999. "Homeostasis, Species, and Higher Taxa." In *Species: New Interdisciplinary Essays*, edited by R. A. Wilson, 141–85. Cambridge, MA: MIT Press.

Brainard, David H., and Anya C. Hurlbert. 2015. "Colour Vision: Understanding #TheDress." *Current Biology: CB* 25 (13): R551-554.

Brett, Matthew, Ingrid S. Johnsrude, and Adrian M. Owen. 2002. "The Problem of Functional Localization in the Human Brain." *Nature Reviews. Neuroscience* 3 (3): 243–49.

Brigandt, Ingo. 2002. "Homology and the Origin of Correspondence." *Biology and Philosophy* 17 (3): 389–407.

Brighton, Henry, and Henrik Olsson. 2009. "Identifying the Optimal Response Is Not a Necessary Step toward Explaining Function." *Behavioral and Brain Sciences* 32 (1): 85–86.

Broad, C. D. 1920. "The Relation between Induction and Probability (Part II)." *Mind* 29 (113): 11–45.

Buccheri, R., V. di Gesù, M. C. Maccarone, and B. Sacco. 1986. "Application of Cluster Analysis to Astronomical Data." *Bulletin d'Information Du Centre de Donnees Stellaires* 31 (November): 205.

Buckner, Cameron. 2016. "Transitional Gradation in the Mind: Rethinking Psychological Kindhood." *British Journal for the Philosophy of Science* 67 (4): 1091–1115.

Buckner, Randy L., Fenna M. Krienen, and B. T. Thomas Yeo. 2013. "Opportunities and Limitations of Intrinsic Functional Connectivity MRI." *Nature Neuroscience* 16 (7): 832–37.

Buehner, Marc J., Patricia W. Cheng, and Deborah Clifford. 2003. "From Covariation to Causation: A Test of the Assumption of Causal Power." *Journal of Experimental Psychology. Learning, Memory, and Cognition* 29 (6): 1119–40.

Chater, Nick, and Mike Oaksford. 2000. "The Rational Analysis Of Mind And Behavior." *Synthese* 122 (1): 93–131.

———. 2008a. "The Probabilistic Mind: Prospects for a Bayesian Cognitive Science." In *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, edited by Nick Chater and Mike Oaksford, 3–31. Oxford, UK: Oxford University Press.

———. 2008b. "The Probabilistic Mind: Where Next?" In *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, edited by Nick Chater and Oaksford, 501–14. Oxford, UK: Oxford University Press.

Chater, Nick, Joshua B. Tenenbaum, and Alan Yuille. 2006. "Probabilistic Models of Cognition: Conceptual Foundations." *Trends in Cognitive Sciences* 10 (7): 287–91.

Chau, Wilkin, and Anthony R. McIntosh. 2005. "The Talairach Coordinate of a Point in the MNI Space: How to Interpret It." *NeuroImage* 25 (2): 408–16.

Chen, Bing, Ting Xu, Changle Zhou, Luoyu Wang, Ning Yang, Ze Wang, Hao-Ming Dong, et al. 2015. "Individual Variability and Test-Retest Reliability Revealed by Ten Repeated Resting-State Brain Scans over One Month." *PLOS ONE* 10 (12): e0144963.

Chen, H., Y. Zhao, Y. Li, J. Lv, and T. Liu. 2017. "Inter-Subject fMRI Registration Based on Functional Networks." In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 863–67.

Cheng, Patricia W. 1997. "From Covariation to Causation: A Causal Power Theory." *Psychological Review* 104 (2): 367–405.

Cheng, Patricia W., and Keith J. Holyoak. 1995. "Complex Adaptive Systems as Intuitive Statisticians: Causality, Contingency, and Prediction." In *Comparative Approaches to Cognitive Science*, 271–302. Complex Adaptive Systems. Cambridge, MA: The MIT Press.

Chetverikov, Andrey, and Ivan Ivanchei. 2016. "Seeing 'the Dress' in the Right Light: Perceived Colors and Inferred Light Sources." *Perception* 45 (8): 910–30.

Cho, Mildred K. 2006. "Racial and Ethnic Categories in Biomedical Research: There Is No Baby in the Bathwater." *The Journal of Law, Medicine & Ethics : A Journal of the American Society of Law, Medicine & Ethics* 34 (3): 497–479.

Clatworthy, Jane, Deanna Buick, Matthew Hankins, John Weinman, and Robert Horne. 2005. "The Use and Reporting of Cluster Analysis in Health Psychology: A Review." *British Journal of Health Psychology* 10 (Pt 3): 329–58.

Cohen, Andrew L., Adam N. Sanborn, and Richard M. Shiffrin. 2008. "Model Evaluation Using Grouped or Individual Data." *Psychonomic Bulletin & Review* 15 (4): 692–712.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Routledge.

Collins, Darrell J., and David R. Shanks. 2006. "Conformity to the Power PC Theory of Causal Induction Depends on the Type of Probe Question." *Quarterly Journal of Experimental Psychology (2006)* 59 (2): 225–32.

Conroy, Bryan R., Benjamin D. Singer, J. Swaroop Guntupalli, Peter J. Ramadge, and James V. Haxby. 2013. "Inter-Subject Alignment of Human Cortical Anatomy Using Functional Connectivity." *NeuroImage* 81 (November): 400–411.

Crivello, Fabrice, Thorsten Schormann, Nathalie Tzourio-Mazoyer, Per E. Roland, Karl Zilles, and Bernard M. Mazoyer. 2002. "Comparison of Spatial Normalization Procedures and Their Impact on Functional Maps." *Human Brain Mapping* 16 (4): 228–50.

Cronbach, Lee J. 1957. "The Two Disciplines of Scientific Psychology." *American Psychologist* 12 (11): 671–84.

Crum, W. R., L. D. Griffin, D. L. G. Hill, and D. J. Hawkes. 2003. "Zen and the Art of Medical Image Registration: Correspondence, Homology, and Quality." *NeuroImage* 20 (3): 1425–37.

Crum, W R, T Hartkens, and D L G Hill. 2004. "Non-Rigid Image Registration: Theory and Practice." *The British Journal of Radiology* 77 (supplement 2): S140–53.

Cummins, Robert. 1975. "Functional Analysis." *Journal of Philosophy* 72 (November): 741–64.

Currie, Adrian. 2018. *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences*. Cambridge, MA: MIT Press.

Danks, David. 2007. "Causal Learning from Observations and Manipulations." In *Thinking with Data*, 359–88. Carnegie Mellon Symposia on Cognition. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

———. 2008. "Rational Analyses, Instrumentalism, and Implementations." In *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, edited by Nick Chater and Mike Oaksford, 59–75. Oxford, UK: Oxford University Press.

Danks, David, and Frederick Eberhardt. 2009. "Conceptual Problems in Statistics, Testing and Experimentation." In *Routledge Companion to the Philosophy of Psychology*, 214–30. New York, NY: Oxford University Press.

Dickie, David Alexander, Susan D. Shenkin, Devasuda Anblagan, Juyoung Lee, Manuel Blesa Cabez, David Rodriguez, James P. Boardman, Adam Waldman, Dominic E. Job, and Joanna M. Wardlaw. 2017. "Whole Brain Magnetic Resonance Image Atlases: A Systematic Review of Existing Atlases and Caveats for Use in Population Imaging." *Frontiers in Neuroinformatics* 11: 1.

Drab, Klaudia, and Michal Daszykowski. 2014. "Clustering in Analytical Chemistry." *Journal of AOAC International* 97 (1): 29–38.

Drissi-Daoudi, Leila, Adrien Doerig, Khatuna Parkosadze, Marina Kunchulia, and Michael H. Herzog. 2020. "How Stable Is Perception in #TheDress and #TheShoe?" *Vision Research* 169 (April): 1–5.

Dubois, Julien, and Ralph Adolphs. 2016. "Building a Science of Individual Differences from fMRI." *Trends in Cognitive Sciences* 20 (6): 425–43.

Dupré, John. 2007. "Fact and Value." In *Value-Free Science? Ideals and Illusions*, edited by Harold Kincaid and Alison Wylie. Oxford, UK: Oxford University Press.

Ebbinghaus, Hermann. 1913. *Memory: A Contribution to Experimental Psychology*. Translated by Henry A. Ruger and Clara E. Bussenius. New York, NY: Teachers College, Columbia University.

Eberhardt, Frederick, and David Danks. 2011. "Confirmation in the Cognitive Sciences: The Problematic Case of Bayesian Models." *Minds and Machines* 21 (3): 389–410.

Edwards, A. W. F. 2003. "Human Genetic Diversity: Lewontin's Fallacy." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 25 (8): 798–801.

Ellermeier, Wolfgang, and Karin Zimmer. 1997. "Individual Differences in Susceptibility to The 'irrelevant Speech Effect.'" *Journal of the Acoustical Society of America* 102 (4): 2191–99.

Emery, Kara J, and Michael A Webster. 2019. "Individual Differences and Their Implications for Color Perception." *Current Opinion in Behavioral Sciences*, Visual perception, 30 (December): 28–33.

Esteban, Oscar, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, et al. 2019. "fMRIPrep: A Robust Preprocessing Pipeline for Functional MRI." *Nature Methods* 16 (1): 111–16.

Estes, W. K. 1956. "The Problem of Inference from Curves Based on Group Data." *Psychological Bulletin* 53 (2): 134–40.

———. 2002. "Traps in the Route to Models of Memory and Decision." *Psychonomic Bulletin & Review* 9 (1): 3–25.

Estes, W. K., and W. Todd Maddox. 2005. "Risks of Drawing Inferences about Cognitive Processes from Model Fits to Individual versus Average Performance." *Psychonomic Bulletin & Review* 12 (3): 403–8.

Evans, Alan C., Andrew L. Janke, D. Louis Collins, and Sylvain Baillet. 2012. "Brain Templates and Atlases." *NeuroImage* 62 (2): 911–22.

Fahrmeir, Ludwig, Thomas Kneib, and Stefan Lang. 2013. "Bayesian Multilevel Models." In *The SAGE Handbook of Multilevel Modeling*, edited by Marc Scott, Jeffrey Simonoff, and Brian Marx, 53–72. London, UK: SAGE Publications Ltd.

Faisal, A. Aldo, Luc P. J. Selen, and Daniel M. Wolpert. 2008. "Noise in the Nervous System." *Nature Reviews. Neuroscience* 9 (4): 292–303.

Farrell, Simon, and Casimir J. H. Ludwig. 2008. "Bayesian and Maximum Likelihood Estimation of Hierarchical Response Time Models." *Psychonomic Bulletin & Review* 15 (6): 1209–17.

Fedorenko, Evelina, and Nancy Kanwisher. 2009. "Neuroimaging of Language: Why Hasn't a Clearer Picture Emerged?" *Language and Linguistics Compass* 3 (4): 839–65.

Fine, Cordelia. 2010. *Delusions of Gender: How Our Minds, Society, and Neurosexism Create Difference*. New York, NY: W. W. Norton & Company.

Finn, Emily S., Dustin Scheinost, Daniel M. Finn, Xilin Shen, Xenophon Papademetris, and R. Todd Constable. 2017. "Can Brain State Be Manipulated to Emphasize Individual Differences in Functional Connectivity?" *NeuroImage* 160: 140–51.

Foulkes, Lucy, and Sarah-Jayne Blakemore. 2018. "Studying Individual Differences in Human Adolescent Brain Development." *Nature Neuroscience* 21 (3): 315–23.

Freeman, Jeremy. 2015. "Open Source Tools for Large-Scale Neuroscience." *Current Opinion in Neurobiology*, Large-Scale Recording Technology (32), 32 (June): 156–63.

Friston, Karl. J., J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. J. Frackowiak. 1995. "Spatial Registration and Normalization of Images." *Human Brain Mapping* 3 (3): 165–89.

Frost, Martin A., and Rainer Goebel. 2012. "Measuring Structural-Functional Correspondence: Spatial Variability of Specialised Brain Regions after Macro-Anatomical Alignment." *NeuroImage* 59 (2): 1369–81.

Gabard-Durnam, Laurel J., Adriana S. Mendez Leal, Carol L. Wilkinson, and April R. Levin. 2018. "The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): Standardized Processing Software for Developmental and High-Artifact Data." *Frontiers in Neuroscience* 12: 97.

Gangestad, Steve, and Mark Snyder. 1985. "'To Carve Nature at Its Joints'. On the Existence of Discrete Classes in Personality." *Psychological Review* 92 (3): 317–49.

Ganzetti, Marco, Quanying Liu, Dante Mantini, and Alzheimer's Disease Neuroimaging Initiative. 2018. "A Spatial Registration Toolbox for Structural MR Imaging of the Aging Brain." *Neuroinformatics* 16 (2): 167–79.

Gegenfurtner, Karl R., Marina Bloj, and Matteo Toscani. 2015. "The Many Colours of 'the Dress.'" *Current Biology* 25 (13): R543–44.

Gelman, Andrew. 2015. "The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective." *Journal of Management* 41 (2): 632–43.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gholipour, A., N. Kehtarnavaz, R. Briggs, M. Devous, and K. Gopinath. 2007. "Brain Functional Localization: A Survey of Image Registration Techniques." *IEEE Transactions on Medical Imaging* 26 (4): 427–51.

Gigerenzer, Gerd, Peter M. Todd, and ABC Research Group. 1999. *Simple Heuristics That Make Us Smart*. Oxford, UK: Oxford University Press.

Gill, Jeff, and Andrew Womack. 2013. "The Multilevel Model Framework." In *The SAGE Handbook of Multilevel Modeling*, edited by Marc Scott, Jeffrey Simonoff, and Brian Marx, 3–20. London, UK: SAGE Publications Ltd.

Gillham, N. W. 2001. "Sir Francis Galton and the Birth of Eugenics." *Annual Review of Genetics* 35: 83–101.

Glöckner, Andreas, and Thorsten Pachur. 2012. "Cognitive Models of Risky Choice: Parameter Stability and Predictive Accuracy of Prospect Theory." *Cognition* 123 (1): 21–32.

Glymour, Clark, and Madelyn R. Glymour. 2014. "Commentary: Race and Sex Are Causes." *Epidemiology* 25 (4): 488–90.

Godfrey-Smith, Peter. 1991. "Optimality and Psychological Explanation." *Behavioral and Brain Sciences* 14 (3): 496–497.

Goldstein, Harvey. 2013. "Likelihood Estimation in Multilevel Models." In *The SAGE Handbook of Multilevel Modeling*, edited by Marc Scott, Jeffrey Simonoff, and Brian Marx, 39–52. London, UK: SAGE Publications Ltd.

Goodman, Nelson. 1972. "Seven Strictures on Similarity." In *Problems and Projects*. Bobs-Merril.

Gould, Stephen Jay. 1981. *The Mismeasure of Man*. Revised & Expanded Edition. New York: W. W. Norton & Company.

Griffiths, Paul, and Karola Stotz. 2013. *Genetics and Philosophy: An Introduction*. New York, NY: Cambridge University Press.

Griffiths, Thomas, Charles Kemp, and Joshua Tenenbaum. 2008. "Bayesian Models of Cognition." In *The Cambridge Handbook of Computational Psychology*, 59–100. Cambridge, UK: Cambridge University Press.

Griffiths, Thomas L. 2009. "The Strengths of – and Some of the Challenges for – Bayesian Models of Cognition." *Behavioral and Brain Sciences* 32 (1): 89–90.

Hausman, Daniel M. 2018. "Philosophy of Economics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/economics/.

Haxby, James V., J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ramadge. 2011. "A Common, High-

Dimensional Model of the Representational Space in Human Ventral Temporal Cortex." *Neuron* 72 (2): 404–16.

Hayes, K. J. 1953. "The Backward Curve: A Method for the Study of Learning." *Psychological Review* 60 (4): 269–75.

Hedge, Craig, Georgina Powell, and Petroc Sumner. 2017. "The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences." *Behavior Research Methods*, July, 1–21.

Hellier, P., C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D. L. Collins, A. Evans, et al. 2003. "Retrospective Evaluation of Intersubject Brain Registration." *IEEE Transactions on Medical Imaging* 22 (9): 1120–30.

Herrnstein, Richard J., and Charles Murray. 1994. *Bell Curve: Intelligence and Class Structure in American Life*. New York, NY: Free Press.

Hesslinger, Vera M., and Claus-Christian Carbon. 2016. "#TheDress: The Role of Illumination Information and Individual Differences in the Psychophysics of Perceiving White–Blue Ambiguities." *I-Perception* 7 (2): 2041669516645592.

Hitchcock, Christopher. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy* 98 (6): 273–99.

Hitchcock, Christopher, and James Woodward. 2003. "Explanatory Generalizations, Part II: Plumbing Explanatory Depth." *Noûs* 37 (2): 181–99.

Hofer, Heidi, Joseph Carroll, Jay Neitz, Maureen Neitz, and David R. Williams. 2005. "Organization of the Human Trichromatic Cone Mosaic." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 25 (42): 9669–79.

Holderness, Cates. 2015. "What Colors Are This Dress?" BuzzFeed. February 26, 2015. https://www.buzzfeed.com/catesish/help-am-i-going-insane-its-definitely-blue.

Hugrass, Laila, Jana Slavikova, Melissa Horvat, Alaa Al Musawi, and David Crewther. 2017. "Temporal Brightness Illusion Changes Color Perception of 'the Dress.'" *Journal of Vision* 17 (5): 6–6.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8): e124.

Jiang, Di, Yuhui Du, Hewei Cheng, Tianzi Jiang, and Yong Fan. 2013. "Groupwise Spatial Normalization of fMRI Data Based on Multi-Range Functional Connectivity Patterns." *NeuroImage* 82 (November): 355–72.

Joel, Daphna, Zohar Berman, Ido Tavor, Nadav Wexler, Olga Gaber, Yaniv Stein, Nisan Shefi, et al. 2015. "Sex beyond the Genitalia: The Human Brain Mosaic." *Proceedings of the National Academy of Sciences* 112 (50): 15468–73.

Jonauskaite, Domicele, Nele Dael, C. Alejandro Parraga, Laetitia Chèvre, Alejandro García Sánchez, and Christine Mohr. 2018. "Stripping #The Dress: The Importance of Contextual Information on Inter-Individual Differences in Colour Perception." *Psychological Research*, September.

Jones, Matt, and Bradley C. Love. 2011. "Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition." *The Behavioral and Brain Sciences* 34 (4): 169-188-231.

Kanai, Ryota, and Geraint Rees. 2011. "The Structural Basis of Inter-Individual Differences in Human Behaviour and Cognition." *Nature Reviews Neuroscience* 12 (4): 231–42.

Kao, Shu-Fang, and Edward A. Wasserman. 1993. "Assessment of an Information Integration Account of Contingency Judgment with Examination of Subjective Cell Importance and Method of Information Presentation." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19 (6): 1363–86.

Kidd, Evan, Seamus Donnelly, and Morten H. Christiansen. 2017. "Individual Differences in Language Acquisition & Processing." *Trends in Cognitive Sciences* 22 (2): 154–69.

Kitcher, Philip. 1989. "Explanatory Unification and the Causal Structure of the World." In *Scientific Explanation*, 8:410–505. Minneapolis, MN: University of Minnesota Press.

———. 2003. *Science, Truth, and Democracy*. New York, NY: Oxford University Press.

Klein, Arno, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E. Christensen, et al. 2009. "Evaluation of 14 Nonlinear Deformation Algorithms Applied to Human Brain MRI Registration." *NeuroImage* 46 (3): 786–802.

Klein, Colin. 2010. "Philosophical Issues in Neuroimaging." *Philosophy Compass* 5 (2): 186–98.

———. 2012. "Cognitive Ontology and Region- Versus Network-Oriented Analyses." *Philosophy of Science* 79 (5): 952–960.

Kosslyn, Stephen M., John T. Cacioppo, Richard J. Davidson, Kenneth Hugdahl, William R. Lovallo, David Spiegel, and Robert Rose. 2002. "Bridging Psychology and Biology. The Analysis of Individuals in Groups." *The American Psychologist* 57 (5): 341–51.

Krzywinski, Martin, and Naomi Altman. 2013. "Power and Sample Size." *Nature Methods* 10 (12): 1139–40.

Lafer-Sousa, Rosa, and Bevil R. Conway. 2017. "#TheDress: Categorical Perception of an Ambiguous Color Image." *Journal of Vision* 17 (12): 25.

Lange, Floris P. de, Micha Heilbron, and Peter Kok. 2018. "How Do Expectations Shape Perception?" *Trends in Cognitive Sciences* 22 (9): 764–79.

Lange, Marc. 2013a. "Really Statistical Explanations and Genetic Drift." *Philosophy of Science* 80 (2): 169–188.

———. 2013b. "What Makes a Scientific Explanation Distinctively Mathematical?" *The British Journal for the Philosophy of Science* Preprint: 1–27.

Langs, Georg, Polina Golland, Yanmei Tie, Laura Rigolo, and Alexandra J. Golby. 2010. "Functional Geometry Alignment and Localization of Brain Areas." *Advances in Neural Information Processing Systems* 1: 1225–33.

Lee, Michael D. 2008. "Three Case Studies in the Bayesian Analysis of Cognitive Models." *Psychonomic Bulletin & Review* 15 (1): 1–15.

———. 2011. "How Cognitive Modeling Can Benefit from Hierarchical Bayesian Models." *Journal of Mathematical Psychology*, Special Issue on Hierarchical Bayesian Models, 55 (1): 1–7.

———. 2018. "Bayesian Methods in Cognitive Modeling." In *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, edited by J. Wixted and E.J. Wagenmakers, 4th ed.: 1–48. John Wiley & Sons, Ltd.

Lee, Michael D., and Eric-Jan Wagenmakers. 2013. *Bayesian Cognitive Modeling: A Practical Course*. New York, NY: Cambridge University Press.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago, IL: University of Chicago Press.

Levy, Arnon. 2017. "Causal Order and Kinds of Robustness." In *Landscapes of Collectivity in the Life Sciences*, edited by Snait Gissis, Ehud Lamm, and Ayelet Shavit, 269–280. Cambridge, MA: MIT Press.

Lewontin, R. C. 1972. "The Apportionment of Human Diversity." In *Evolutionary Biology*, edited by Theodosius Dobzhansky, Max K. Hecht, and William C. Steere, Vol. 6:381–98. New York, NY: Springer.

Liebeskind, Benjamin J., David M. Hillis, Harold H. Zakon, and Hans A. Hofmann. 2016. "Complex Homology and the Evolution of Nervous Systems." *Trends in Ecology & Evolution* 31 (2): 127–35.

Lieder, Falk, and Thomas L. Griffiths. 2019. "Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources." *The Behavioral and Brain Sciences* 43 (February): 1–16.

Locke, John. 1690. *An Essay Concerning Human Understanding*. Edited by Roger Woolhouse. London: Penguin Classics.

Longino, Helen E. 2004. "How Values Can Be Good for Science." In *Science, Values, and Objectivity*, edited by Peter K. Machamer and Gereon Wolters, 127–142. Pittsburgh, PA: University of Pittsburgh Press.

Lonsdorf, Tina B., and Christian J. Merz. 2017. "More than Just Noise: Inter-Individual Differences in Fear Acquisition, Extinction and Return of Fear in Humans." *Neuroscience and Biobehavioral Reviews* 80 (July): 703–28.

Louçã, Francisco. 2009. "Emancipation through Interaction--How Eugenics and Statistics Converged and Diverged." *Journal of the History of Biology* 42 (4): 649–84.

Lyon, Aidan. 2014. "Why Are Normal Distributions Normal?" *British Journal for the Philosophy of Science* 65 (3): 621–649..

Machery, Edouard. 2017. "Kinds or Tails." In *Extraordinary Science and Psychiatry*, 15–36. Cambridge, MA: MIT Press.

Mackie, J. L. 1974. *Cement Of The Universe*. Oxford, UK: Oxford University Press.

Marr, David. 1982. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Henry Holt and Company.

Martín-Moro, J.G., F.P. Garrido, F.G. Sanz, I.F. Vega, M.C. Rebollo, and P.M. Martín. 2018. "Which Are the Colors of the Dress? Review of an Atypical Optic Illusion." *Archivos De La Sociedad Espanola De Oftalmologia* 93 (4): 186–92.

Matute, Helena, Francisco Arcediano, and Ralph R. Miller. 1996. "Test Question Modulates Cue Competition between Causes and between Effects." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22 (1): 182–96.

Mayrhofer, Ralf, and Michael R. Waldmann. 2016. "Sufficiency and Necessity Assumptions in Causal Structure Induction." *Cognitive Science* 40 (8): 2137–50.

Mazziotta, J, A Toga, A Evans, P Fox, J Lancaster, K Zilles, R Woods, et al. 2001. "A Probabilistic Atlas and Reference System for the Human Brain: International Consortium for Brain Mapping (ICBM)." *Philosophical Transactions of the Royal Society of London. Series B* 356 (1412): 1293–1322.

McAllister, James W. 1997. "Phenomena and Patterns in Data Sets." *Erkenntnis* 47 (2): 217–28.

McCaffrey, Joseph, and David Danks. Forthcoming. "Mixtures and Psychological Inference with Resting State fMRI." *The British Journal for the Philosophy of Science*.

Metzger, Anna, and Knut Drewing. 2019. "Memory Influences Haptic Perception of Softness." *Scientific Reports* 9 (1): 1–10.

Mollon, John D., Jenny M. Bosten, David H. Peterzell, and Michael A. Webster. 2017. "Individual Differences in Visual Science: What Can Be Learned and What Is Good Experimental Practice?" *Vision Research*, 141 (December): 4–15.

Myung, In Jae, Cheongtag Kim, and Mark A. Pitt. 2000. "Toward an Explanation of the Power Law Artifact: Insights from Response Surface Analysis." *Memory & Cognition* 28 (5): 832–40.

Myung, In Jae, and Mark A. Pitt. 1997. "Applying Occam's Razor in Modeling Cognition: A Bayesian Approach." *Psychonomic Bulletin & Review* 4 (1): 79–95.

Navarro, Daniel J., Matthew J. Dry, and Michael D. Lee. 2012. "Sampling Assumptions in Inductive Generalization." *Cognitive Science* 36 (2): 187–223.

Navarro, Daniel J., Thomas L. Griffiths, Mark Steyvers, and Michael D. Lee. 2006. "Modeling Individual Differences Using Dirichlet Processes." *Journal of Mathematical Psychology*, Special Issue on Model Selection: Theoretical Developments and Applications, 50 (2): 101–22.

Nenning, Karl-Heinz, Hesheng Liu, Satrajit S. Ghosh, Mert R. Sabuncu, Ernst Schwartz, and Georg Langs. 2017. "Diffeomorphic Functional Brain Surface Alignment: Functional Demons." *NeuroImage* 156 (August): 456–65.

Newell, A., and Paul Rosenbloom. 1981. "Mechanisms of Skill Acquisition and the Law of Practice." In *Cognitive Skills and Their Acquisition*, edited by J.R. Anderson, Vol. 1:1–55. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ng, Bernard, Rafeef Abugharbieh, and Martin J. McKeown. 2009. "Adverse Effects of Template-Based Warping on Spatial fMRI Analysis." In *Medical Imaging 2009: Biomedical Applications in Molecular, Structural, and Functional Imaging*, 7262:72621Y. International Society for Optics and Photonics.

Nieto-Castañón, Alfonso, and Evelina Fedorenko. 2012. "Subject-Specific Functional Localizers Increase Sensitivity and Functional Resolution of Multi-Subject Analyses." *NeuroImage* 63 (3): 1646–69.

Nilsson, Håkan, Jörg Rieskamp, and Eric-Jan Wagenmakers. 2011. "Hierarchical Bayesian Parameter Estimation for Cumulative Prospect Theory." *Journal of Mathematical Psychology*, Special Issue on Hierarchical Bayesian Models, 55 (1): 84–93.

Northcott, Robert. 2009. "Is Actual Difference Making Actually Different?" *Journal of Philosophy* 106 (11): 629–633.

Nosofsky, Robert M. 1998. "Optimal Performance and Exemplar Models of Classification." In *The Rational Analysis Of Mind And Behavior*, edited by Mike Oaksford and Nick Chater, 218–47. Oxford, UK: Oxford University Press.

Nunez, Michael D., Ramesh Srinivasan, and Joachim Vandekerckhove. 2015. "Individual Differences in Attention Influence Perceptual Decision Making." *Frontiers in Psychology* 8 (February).

Oaksford, Mike, and Nick Chater, eds. 1998. *Rational Models of Cognition*. Oxford, UK: Oxford University Press.

———. 2009a. "Précis of Bayesian Rationality: The Probabilistic Approach to Human Reasoning." *Behavioral and Brain Sciences* 32 (1): 69-84-120.

———. 2009b. "The Uncertain Reasoner: Bayes, Logic, and Rationality." *Behavioral and Brain Sciences* 32 (1): 105–20.

Olkkonen, Maria, and Vebjørn Ekroll. 2016. "Color Constancy and Contextual Effects on Color Appearance." In *Human Color Vision*, 159–88. Springer Series in Vision Research. New York, NY: Springer.

Orden, G. C. van. 1997. "Functional Neuroimages Fail to Discover Pieces of Mind in the Parts of the Brain." *Philosophy of Science Supplement* 64 (4): 85–94.

Orzack, Steven Hecht, and Elliott Sober. 1994. "Optimality Models and the Test of Adaptationism." *The American Naturalist* 143 (3): 361–80.

Osaka, Mariko, Naoyuki Osaka, Hirohito Kondo, Masanao Morishita, Hidenao Fukuyama, Toshihiko Aso, and Hiroshi Shibasaki. 2003. "The Neural Basis of Individual Differences in Working Memory Capacity: An fMRI Study." *NeuroImage* 18 (3): 789–97.

Osman, Magda, and David R. Shanks. 2005. "Individual Differences in Causal Learning and Decision Making." *Acta Psychologica* 120 (1): 93–112.

Owen, Richard. 1843. *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals*. London: Longman, Brown, Green, and Longmans.

Panchen, Alec L. 1994. "Richard Owen and the Concept of Homology." In *Homology: The Hierarchical Basis of Comparative Biology*, 21–62. San Diego, CA: Academic Press.

Pearce, Bradley, Stuart Crichton, Michal Mackiewicz, Graham D. Finlayson, and Anya Hurlbert. 2014. "Chromatic Illumination Discrimination Ability Reveals That Human Colour Constancy Is Optimised for Blue Daylight Illuminations." *PLOS ONE* 9 (2): e87989.

Perales, José C., and David R. Shanks. 2008. "Driven by Power? Probe Question and Presentation Format Effects on Causal Judgment." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34 (6): 1482–94.

Pervin, L. A. 1985. "Personality: Current Controversies, Issues, and Directions." *Annual Review of Psychology* 36 (1): 83–114.

Poline, Jean-Baptiste, Bertrand Thirion, Alexis Roche, and Sébastien Merlaux. 2010. "Intersubject Variability in fMRI Data: Causes, Consequences, and Related Analysis Strategies." In *Foundational Issues in Human Brain Mapping*, edited by Stephen José Hanson and Martin Bunzl, 173–91. Cambridge, MA: MIT Press.

Porto, Patricia Ribeiro, Leticia Oliveira, Jair Mari, Eliane Volchan, Ivan Figueira, and Paula Ventura. 2009. "Does Cognitive Behavioral Therapy Change the Brain? A Systematic Review of Neuroimaging in Anxiety Disorders." *The Journal of Neuropsychiatry and Clinical Neurosciences* 21 (2): 114–25.

Potochnik, Angela. 2007. "Optimality Modeling and Explanatory Generality." *Philosophy of Science* 74 (5): 680–91.

Rahman, Maryam, Gregory J. A. Murad, and J. Mocco. 2009. "Early History of the Stereotactic Apparatus in Neurosurgery." *Neurosurgical Focus* 27 (3): E12.

Ramsey, Grant, and Anne Peterson. 2012. "Sameness in Biology." *Philosophy of Science* 79 (2): 255–75.

Ratcliff, Roger, Philip L. Smith, Scott D. Brown, and Gail McKoon. 2016. "Diffusion Decision Model: Current Issues and History." *Trends in Cognitive Sciences* 20 (4): 260–81.

Raudenbush, Stephen W., and Anthony S. Bryk. 2001. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA: SAGE Publications, Inc.

Reeder, Reshanne R. 2017. "Individual Differences Shape the Content of Visual Representations." *Vision Research* 141: 266–81.

Rehder, Bob. 2011. "Taking the Rationality out of Probabilistic Models." *Behavioral and Brain Sciences* 34 (4): 210–11.

———. 2014. "Independence and Dependence in Human Causal Reasoning." *Cognitive Psychology* 72 (July): 54–107.

Reijula, Samuli. Manuscript. "How Could a Rational Analysis Model Explain?" https://cogsci.mindmodeling.org/2017/papers/0563/paper0563.pdf.

Rescorla, Michael. 2018. "An Interventionist Approach to Psychological Explanation." *Synthese* 195 (5): 1909–40.

Revelle, William, Joshua A. Wilt, and David M. Condon. 2011. "Individual Differences and Differential Psychology: A Brief History and Prospect." In *Handbook of Individual Differences*, 3–38. Malden, MA: Wiley-Blackwell.

Robinson, Emma C., Saad Jbabdi, Matthew F. Glasser, Jesper Andersson, Gregory C. Burgess, Michael P. Harms, Stephen M. Smith, David C. Van Essen, and Mark Jenkinson. 2014. "MSM: A New Flexible Framework for Multimodal Surface Matching." *NeuroImage* 100 (October): 414–26.

Roca, A. L., N. Georgiadis, J. Pecon-Slattery, and S. J. O'Brien. 2001. "Genetic Evidence for Two Species of Elephant in Africa." *Science* 293 (5534): 1473–77.

Roland, P. E., and K. Zilles. 1994. "Brain Atlases - a New Research Tool." *Trends in Neurosciences* 17 (11): 458–67.

Rooij, Iris van, Cory D. Wright, Johan Kwisthout, and Todd Wareham. 2018. "Rational Analysis, Intractability, and the Prospects of 'as If'-Explanations." *Synthese* 195 (2): 491–510.

Roskies, Adina L. 2007. "Are Neuroimages Like Photographs of the Brain?" *Philosophy of Science* 74 (5): 860–872.

Ross, Lauren N. Forthcoming. "Causal Control: A Rationale for Causal Selection." In *Causal Reasoning in Biology*, edited by B. Hanley, C.K. Waters, and J.F. Woodward. Minnesota Studies in the Philosophy of Science. University of Minnesota Press.

Rouder, Jeffrey N., and Julia M. Haaf. 2019. "A Psychometrics of Individual Differences in Experimental Tasks." *Psychonomic Bulletin & Review* 26 (2): 452–67.

Rouder, Jeffrey N., Jun Lu, Paul Speckman, DongChu Sun, and Yi Jiang. 2005. "A Hierarchical Model for Estimating Response Time Distributions." *Psychonomic Bulletin & Review* 12 (2): 195–223.

Saad, Ziad S., and Richard C. Reynolds. 2012. "SUMA." *NeuroImage* 62 (2): 768–73.

Sabuncu, Mert R., Benjamin D. Singer, Bryan Conroy, Ronald E. Bryan, Peter J. Ramadge, and James V. Haxby. 2010. "Function-Based Intersubject Alignment of Human Cortical Anatomy." *Cerebral Cortex* 20 (1): 130–40.

Sakamoto, Yasuaki, Matt Jones, and Bradley C. Love. 2008. "Putting the Psychology Back into Psychological Models: Mechanistic versus Rational Approaches." *Memory & Cognition* 36 (6): 1057–65.

Samuels, Richard. 2012. "Science and Human Nature." *Royal Institute of Philosophy Supplement* 70: 1–28.

Sauce, Bruno, and Louis D. Matzel. 2013. "The Causes of Variation in Learning and Behavior: Why Individual Differences Matter." *Frontiers in Psychology* 4 (July).

Scheperjans, Filip, Simon B. Eickhoff, Lars Hömke, Hartmut Mohlberg, Klaudia Hermann, Katrin Amunts, and Karl Zilles. 2008. "Probabilistic Maps, Morphometry, and Variability of Cytoarchitectonic Areas in the Human Superior Parietal Cortex." *Cerebral Cortex* 18 (9): 2141–57.

Schustack, Miriam W., and Robert J. Sternberg. 1981. "Evaluation of Evidence in Causal Inference." *Journal of Experimental Psychology* 110 (1): 101–20.

Seghier, Mohamed L., and Cathy J. Price. 2018. "Interpreting and Utilising Intersubject Variability in Brain Function." *Trends in Cognitive Sciences* 22 (6): 517–30.

Sewell, David K., Daniel R. Little, and Stephan Lewandowsky. 2011. "Bayesian Computation and Mechanism: Theoretical Pluralism Drives Scientific Emergence." *Behavioral and Brain Sciences* 34 (4): 212–13.

Shanks, David R. 1995. "Is Human Learning Rational?" *The Quarterly Journal of Experimental Psychology Section A* 48 (2): 257–79.

Shapiro, Lawrence A. 2000. "Multiple Realizations." *Journal of Philosophy* 97 (12): 635–654.

Shattuck, David W., Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L. Narr, Russell A. Poldrack, Robert M. Bilder, and Arthur W. Toga. 2008. "Construction of a 3D Probabilistic Atlas of Human Cortical Structures." *NeuroImage* 39 (3): 1064–80.

Shen, Jianhong, and Thomas J. Palmeri. 2016. "Modelling Individual Difference in Visual Categorization." *Visual Cognition* 24 (3): 260–83.

Shiffrin, Richard M., Michael D. Lee, Woojae Kim, and Eric-Jan Wagenmakers. 2008. "A Survey of Model Evaluation Approaches with a Tutorial on Hierarchical Bayesian Methods." *Cognitive Science* 32 (8): 1248–84.

Sidman, Murray. 1952. "A Note on Functional Relations Obtained from Group Data." *Psychological Bulletin* 49 (June): 263–69.

Siegler, Robert S. 1976. "Three Aspects of Cognitive Development." *Cognitive Psychology* 8 (4): 481–520.

Simonsohn, Uri. 2015. "Small Telescopes: Detectability and the Evaluation of Replication Results." *Psychological Science* 26 (5): 559–69.

Slater, Matthew H. 2015. "Natural Kindness." *The British Journal for the Philosophy of Science* 66 (2): 375–411.

Sloman, Steven, and Philip M. Fernbach. 2008. "The Value of Rational Analysis: An Assessment of Causal Reasoning and Learning." In *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, edited by Nick Chater and Oaksford, 485–500. Oxford, UK: Oxford University Press.

Snow, Paul. 1991. "Rationality and Irrationality: Still Fighting Words." *Behavioral and Brain Sciences* 14 (3): 505–6.

Sporns, Olaf. 2013. "The Human Connectome: Origins and Challenges." *NeuroImage* 80 (October): 53–61.

Stegmann, Ulrich E. 2012. "Varieties of Parity." *Biology and Philosophy* 27 (6): 903–918.

Stephan, Klaas Enno, Will D. Penny, Jean Daunizeau, Rosalyn J. Moran, and Karl J. Friston. 2009. "Bayesian Model Selection for Group Studies." *NeuroImage* 46 (4): 1004–17.

Steyvers, Mark, Joshua B. Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum. 2003. "Inferring Causal Networks from Observations and Interventions." *Cognitive Science* 27 (3): 453–89.

Strevens, Michael. 2007. "Review of Woodward, 'Making Things Happen.'" Edited by James Woodward. *Philosophy and Phenomenological Research* 74 (1): 233–49.

Subramaniam, Banu. 2014. "A Genealogy of Variation: The Enduring Debate on Human Differences." In *Ghost Stories for Darwin: The Science of Variation and the Politics of Diversity*, 45–69. Urbana, IL: University of Illinois Press.

Swets, Benjamin. 2015. "Psycholinguistics and Planning: A Focus on Individual Differences." In *Individual Differences in Speech Production and Perception*, edited by Susanne Fuchs, Daniel Pape, Caterina Petrone, and Pascal Perrier, 89–122. New York, NY: Peter Lang.

Tabery, James. 2009. "Difference Mechanisms: Explaining Variation with Mechanisms." *Biology and Philosophy* 24 (5): 645–664.

Talairach, J., G. Szikla, P. Tournoux, A. Prosalentis, M. Bordas-Ferrier, L. Covello, M. Iacob, and E. Mempel. 1967. *Atlas D'anatomie Stereotaxique Du Telencephale.* Paris: Masson.

Talairach, J., and Pierre Tournoux. 1988. *Co-Planar Stereotaxic Atlas of the Human Brain*. Stuttgart ; New York : New York: G. Thieme ; Thieme Medical Publishers.

Thagard, Paul, and Richard E. Nisbett. 1983. "Rationality and Charity." *Philosophy of Science* 50 (2): 250–67.

Thirion, Bertrand, Philippe Pinel, Sébastien Mériaux, Alexis Roche, Stanislas Dehaene, and Jean-Baptiste Poline. 2007. "Analysis of a Large fMRI Cohort: Statistical and Methodological Issues for Group Analyses." *NeuroImage* 35 (1): 105–20.

Toga, Arthur W. 1998. *Brain Warping*. San Diego, CA: Academic Press.

Toga, Arthur W., and John C. Mazziotta. 2002. *Brain Mapping: The Methods*. 2nd ed. Boston: Academic Press.

Toga, Arthur W., and Paul M. Thompson. 2001. "Maps of the Brain." *The Anatomical Record* 265 (2): 37–53.

———. 2007. "What Is Where and Why It Is Important." *NeuroImage* 37 (4): 1045–68.

Toga, Arthur W., Paul M. Thompson, Susumu Mori, Katrin Amunts, and Karl Zilles. 2006. "Towards Multimodal Atlases of the Human Brain." *Nature Reviews. Neuroscience* 7 (12): 952–66.

Toscani, Matteo, Karl R. Gegenfurtner, and Katja Doerschner. 2017. "Differences in Illumination Estimation in #thedress." *Journal of Vision* 17 (1): 22–22.

Tucholka, Alan, Virgile Fritsch, Jean-Baptiste Poline, and Bertrand Thirion. 2012. "An Empirical Comparison of Surface-Based and Volume-Based Group Studies in Neuroimaging." *NeuroImage* 63 (3): 1443–53.

Tyler, Leona Elizabeth. 1965. *The Psychology of Human Differences*. 3rd ed. New York, NY: Appleton-Century-Crofts.

Underwood, Benton J. 1975. "Individual Differences as a Crucible in Theory Construction." *American Psychologist* 30 (2): 128–34.

Unsworth, Nash, Josef C. Schrock, and Randall W. Engle. 2004. "Working Memory Capacity and the Antisaccade Task: Individual Differences in Voluntary Saccade Control." *Journal of Experimental Psychology. Learning, Memory, and Cognition* 30 (6): 1302–21.

Upton, Graham, and Ian Cook. 2008. *A Dictionary of Statistics*. 2nd ed. Oxford, UK: Oxford University Press.

Uylings, H. B. M., G. Rajkowska, E. Sanz-Arigita, K. Amunts, and K. Zilles. 2005. "Consequences of Large Interindividual Variability for Human Brain Atlases: Converging Macroscopical Imaging and Microscopical Neuroanatomy." *Anatomy and Embryology* 210 (5–6): 423–31.

Van Essen, David C., and Donna Dierker. 2007. "On Navigating the Human Cerebral Cortex Response to 'In Praise of Tedious Anatomy.'" *NeuroImage* 37 (4): 1050–68.

Van Horn, John Darrell, Scott T. Grafton, and Michael B. Miller. 2008. "Individual Variability in Brain Activity: A Nuisance or an Opportunity?" *Brain Imaging and Behavior* 2 (4): 327–34.

Vandekerckhove, Joachim. 2014. "A Cognitive Latent Variable Model for the Simultaneous Analysis of Behavioral and Personality Data." *Journal of Mathematical Psychology* 60 (June): 58–71.

Vemuri, Kavita, Kulvinder Bisla, SaiKrishna Mulpuru, and Srinivasa Varadharajan. 2016. "Do Normal Pupil Diameter Differences in the Population Underlie the Color Selection of #thedress?" *Journal of the Optical Society of America. A, Optics, Image Science, and Vision* 33 (3): A137-142.

Vemuri, Kavita, Akanksha Srivastava, Saksham Agrawal, and Mithra Anand. 2018. "Age, Pupil Size Differences, and Color Choices for The 'dress' and The 'jacket.'" *Journal of the Optical Society of America. A, Optics, Image Science, and Vision* 35 (4): B347–55.

Wagner, Günter P. 1989. "The Biological Homology Concept." *Annual Review of Ecology and Systematics* 20 (1): 51–69.

———. 1994. "Homology and the Mechanisms of Development." In *Homology: The Hierarchial Basis of Comparative Biology*, edited by Brian K. Hall, 273–99. New York, NY: Academic Press.

———. 2014. *Homology, Genes, and Evolutionary Innovation*. Princeton, NJ: Princeton University Press.

Wallisch, Pascal. 2017. "Illumination Assumptions Account for Individual Differences in the Perceptual Interpretation of a Profoundly Ambiguous Stimulus in the Color Domain: 'The Dress.'" *Journal of Vision* 17 (4): 5–5.

Wallisch, Pascal, and Michael Karlovich. 2019. "Disagreeing about Crocs and Socks: Creating Profoundly Ambiguous Color Displays." https://arxiv.org/abs/1908.05736v1.

Waters, C. Kenneth. 2007. "Causes That Make a Difference." *Journal of Philosophy* 104 (11): 551–579.

Watson, J. D. G., R. Myers, R. S. J. Frackowiak, J. V. Hajnal, R. P. Woods, J. C. Mazziotta, S. Shipp, and S. Zeki. 1993. "Area V5 of the Human Brain: Evidence from a Combined Study Using Positron Emission Tomography and Magnetic Resonance Imaging." *Cerebral Cortex* 3 (2): 79–94.

Webster, Michael A. 2015. "Individual Differences in Color Vision." In *Handbook of Color Psychology*, edited by Andrew J. Elliot, Mark D. Fairchild, and Anna Franklin. Cambridge, UK: Cambridge University Press.

Weiner, Kevin S., Michael A. Barnett, Simon Lorenz, Julian Caspers, Anthony Stigliani, Katrin Amunts, Karl Zilles, Bruce Fischl, and Kalanit Grill-Spector. 2017. "The Cytoarchitecture of Domain-Specific Regions in Human High-Level Visual Cortex." *Cerebral Cortex* 27 (1): 146–61.

White, Peter A. 2000. "Causal Judgment from Contingency Information: Relation between Subjective Reports and Individual Tendencies in Judgment." *Memory & Cognition* 28 (3): 415–26.

———. 2009. "Accounting for Occurrences: An Explanation for Some Novel Tendencies in Causal Judgment from Contingency Information." *Memory & Cognition* 37 (4): 500–513.

Wilmer, Jeremy B. 2008. "How to Use Individual Differences to Isolate Functional Organization, Biology, and Utility of Visual Functions; with Illustrative Proposals for Stereopsis." *Spatial Vision* 21 (6): 561–79.

Witzel, Christoph, J. Kevin O'Regan, and Sabrina Hansmann-Roth. 2017a. "The Dress and Individual Differences in the Perception of Surface Properties." *Vision Research* 141: 76–94.

Witzel, Christoph, Chris Racey, and J. Kevin O'Regan. 2017b. "The Most Reasonable Explanation of 'the Dress': Implicit Assumptions about Illumination." *Journal of Vision* 17 (2): 1–1.

Witzel, Christoph, and Matteo Toscani. 2020. "How to Make a #theDress." *JOSA A* 37 (4): A202–11.

Woods, R. P., S. T. Grafton, J. D. Watson, N. L. Sicotte, and J. C. Mazziotta. 1998. "Automated Image Registration: II. Intersubject Validation of Linear and Nonlinear Models." *Journal of Computer Assisted Tomography* 22 (1): 153–65.

Woodward, James. 1995. "Causation and Explanation in Econometrics." In *On the Reliability of Economic Models: Essays in the Philosophy of Economics*, edited by Daniel Little, 9–61. Recent Economic Thought Series. Dordrecht: Springer Netherlands.

———. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford, UK: Oxford University Press.

———. 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology & Philosophy* 25 (3): 287–318.

———. 2016. "Causation and Manipulability." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/causation-mani/.

Woodward, James, and Christopher Hitchcock. 2003. "Explanatory Generalizations, Part I: A Counterfactual Account." *Noûs* 37 (1): 1–24.

Woody, Andrea. 2019. "Smart Search Through Complex Landscapes." presented at the Annual Lecture Series, Center for Philosophy of Science, University of Pittsburgh, March 22.

Wright, Larry. 1976. *Teleological Explanations: An Etiological Analysis of Goals and Functions*. Berkeley, CA: University of California Press.

Ylikoski, Petri. 2013. "Causal and Constitutive Explanation Compared." *Erkenntnis* 78 (2): 277–297.

Zeigenfuse, Matthew D., and Michael D. Lee. 2009. "Bayesian Nonparametric Modeling of Individual Differences: A Case Study Using Decision-Making on Bandit Problems." In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 1412–17. Austin, TX: Cognitive Science Society.

Zhang, Qiong, Jelmer P. Borst, Robert E. Kass, and John R. Anderson. 2017. "Inter-Subject Alignment of MEG Datasets in a Common Representational Space." *Human Brain Mapping* 38 (9): 4287–4301.

Zhou, Yujia, Pew-Thian Yap, Han Zhang, Lichi Zhang, Qianjin Feng, and Dinggang Shen. 2017. "Improving Functional MRI Registration Using Whole-Brain Functional Correlation Tensors." *Medical Image Computing and Computer-Assisted Intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 10433 (September): 416–23.