A NEW APPROACH TO ALGORITHM-ORIENTED VISUAL PSYCHOPHYSICS

David Nathan White

A DISSERTATION

in

Neuroscience

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2024

Supervisor of Dissertation

Johannes Burge, Associate Professor of Psychology

Graduate Group Chairperson

David Brainard, RRL Professor of Psychologyy

Dissertation Committee

Nicole Rust, Professor of Psychology
Diego Contreras, Professor of Neuroscience
Josh Gold, Professor of Psychology

A NEW APPROACH TO ALGORITHM-ORIENTED VISUAL PSYCHOPHYSICS

ABSTRACT

A NEW APPROACH TO ALGORITHM-ORIENTED VISUAL PSYCHOPHYSICS

David Nathan White

Johannes Burge

The goal of visual neuroscience is to understand how visual systems are able to reconstruct scenes, infer scene properties, and make inferences about the natural environment from a pair of retinal images. Some approaches use psychophysical methods, which allow for controlled sensory presentation and perceptual measurement. Theory holds that through principled application of psychophysics and perceptual modelling, the mechanisms of the visual system can be uncovered. The most common approach has been to measure and model behavioral responses to simple stimuli. However, contemporary neuroscience continues to reveal a more holistic and complex visual system than initially expected. In this work, I argue that these findings suggest behavioral responses to simple stimuli may not as directly correspond to specific visual processing mechanisms, nor provide as powerful an assessment of mechanistic models as originally assumed. Algorithmic theory provides mathematical support for why this is the case and provides a prescription for how psychophysics and modelling can more efficiently explore the space of visual mechanisms. Here, I propose the use of natural or naturalistic images and repeated-measure experimental design to meet this prescription. Not only do natural images provide a diversity of stimuli—an effective substrate for exploring the space of visual mechanisms—but they are also relevant to the overarching goals of vision science—to understand how vision works in the real world. Last, I present the empirically-based portion of this work that acts as as a proof of concept for this framework. In this study I investigate human stereo-depth discrimination performance using naturalistic images and repeated-measures design. I develop broadly applicable methods that can be used to make powerful model assessments based upon the diversity of stimuli. Additionally, the methods and procedures developed provide highly interpretable results: these results show how natural image variability in luminance-patterns and depth-profiles limit human stereo-depth discrimination.

TABLE OF CONTENTS

ABSTRACT iii

LIST OF ILLUSTRATIONS vii

CHAPTER 1 : BACKGROUND 1

1.1 Introduction 1

1.2 Psychophysical approaches 2

1.2.1 The Marr-Poggio approach 2

1.2.2 A note on terminology 3

1.2.3 Modern approaches 4

1.3 Motivating thesis 5

1.4 Visual Processing 6

1.4.1 Stimulus-driven processing 6

1.4.2 Task-driven processing 8

1.5 Conclusion 10

CHAPTER 2 : THEORY 13

2.1 Limitations of psychophysics 13

2.1.1 Algorithmic theory 15

2.1.2 Model testing 16

2.2 Simple stimuli 18

2.3 Alternative Approach 18

CHAPTER 3 : NATURAL IMAGES 20

3.0.1 Formation 20

3.0.2 Content 21

3.0.3 Space 21

3.0.4 Variability 23

3.1 Psychophsyical Approaches . . . 24
3.2 Procuring natural images . . . 26
3.3 Presenting Natural Images . . . 29
3.4 Concluding remarks . . . 30

CHAPTER 4 : HOW DISTINCT SOURCES OF NUISANCE VARIABILITY IN NATURAL IMAGES AND SCENES LIMIT HUMAN STEREOPSIS . . . 31
4.1 Abstract . . . 31
4.2 Introduction . . . 32
4.3 Materials and Methods . . . 34
4.3.1 Human Observers . . . 34
4.3.2 Data and Software . . . 34
4.3.3 Apparatus . . . 35
4.3.4 Stimuli . . . 35
4.3.5 Procedure . . . 39
4.3.6 Psychometric fitting . . . 41
4.3.7 Modeling the decision variable . . . 43
4.3.8 Decision-variable correlation . . . 44
4.3.9 Partitioning the externally-driven component of the decision variable . . . 48
4.3.10 Between-observers decision-variable correlation . . . 53
4.3.11 Spatial integration . . . 55
4.4 Results . . . 57
4.4.1 Decision-variable correlation . . . 59
4.4.2 Experiment 1: Natural stimuli with natural depth profiles . . . 62
4.4.3 Experiment 2: Natural stimuli with flattened depth profiles . . . 68
4.4.4 Partitioning sources of variability in natural stimuli . . . 72
4.4.5 Shared stimulus drive between observers . . . 78
4.5 Discussion . . . 81
4.5.1 Progress and limitations . . . 82

4.5.2 Performance variation and prediction 83
4.5.3 Noise and its impact on performance 85
4.5.4 External limits to human performance 86

CHAPTER 5 : SUMMARY OF CONTRIBUTIONS 88
5.1 Contribution of Chapter 1 88
5.2 Contribution of Chapter 2 88
5.3 Contribution of Chapter 3 89
5.4 Contribution of Chapter 4 90
5.4.1 Contribution to behavioral visual science 90
5.4.2 Contribution to algorithmic perceptual science 91

BIBLIOGRAPHY 93

LIST OF ILLUSTRATIONS

FIGURE 2.1 Simple illustration of how different functions can describe the same data . . . 14

FIGURE 4.1 Sources of uncertainty in stereo-depth perception, stereo-image database, and experimental stimuli . . . 58
FIGURE 4.2 Double-pass experimental design . . . 60
FIGURE 4.3 Discrimination thresholds, response agreement, and estimates of decision-variable correlation results for one observer . . . 63
FIGURE 4.4 Experiment 1 discrimination thresholds and decision-variable correlations . . . 64
FIGURE 4.5 External stimulus-driven and internal noise-driven contributions to thresholds in Experiment 1 . . . 67
FIGURE 4.6 Between-observer variability is primarily attributable to differences in internal noise . . . 69
FIGURE 4.7 Experiment 2 disparity discrimination thresholds and decision-variable correlation . . . 70
FIGURE 4.8 External stimulus-driven and internal noise-driven contributions to thresholds in Experiment 2 . . . 71
FIGURE 4.9 Robustness of fitting methods . . . 76
FIGURE 4.10 Contributions of distinct stimulus-specific factors to thresholds, as revealed by the quasi-quadruple-pass analysis . . . 77
FIGURE 4.11 Between-observer correlation in the stimulus-driven component of the decision variable, as revealed by the quasi-quadruple-pass analysis . . . 80

CHAPTER 1

BACKGROUND

1.1. Introduction

Natural scenes—characterized by their vast diversity—produce a multitude of visual stimuli, consisting of intricate textures, varying lighting conditions, and dynamic elements. The variability and complexity of stimuli present challenges for visual systems in sensing and interpreting this information in real time. For human observers, the visual percepts are sufficiently convincing and useful, so much so that observers are rarely made aware of their indirect and illusory nature. The automatic, continuous, and seeming ease of visual perception conceals the vast complexity of its necessary underlying visual processing, much of which remains to be understood.

One approach to uncovering details about how the visual system functions is by measuring and modelling psychophysical performance. Psychophysics provides a unique insight into visual processing, relating inputs to outputs of the system—what is sensed to what is perceived. However, the overarching goals of vision science include not only a detailed understanding of *what* is perceived, but also *how* this is accomplished. While many current psychophysical approaches are undoubtedly invaluable in their ability to uncover perceptual outcomes, their effectiveness in revealing perceptual mechanisms remains unclear. In this dissertation, I reason for and introduce a new approach towards uncovering the computational mechanisms, or algorithms, of perception.

In this chapter, traditional psychophysical approaches that investigate the algorithmic mechanisms of vision are outlined. Then, I argue that the assumptions that support these approaches are at odds with a modern understanding of visual processing. In Chapter 2, the limitations of traditional approaches are made more explicit mathematically through the application of an algorithmic-theory-based analysis. Importantly, this analysis provides a general prescription for how algorithmic processing should be investigated and how models of processing should be tested. In light of this analysis, the new approach is then proposed. Central to this approach is the use of natural images, which not only meets that prescription but are relevant to an understanding of how visual systems

function in the real world. In Chapter 3, natural images are discussed in terms of their structure and their use in psychophysical experiments. In Chapter 4, a psychophysics study conducted by myself is presented. This study follows the new approach developed in Chapter 2, thus acting as proof-of-concept for the tractability of the experimental portion of the approach. This study also develops the necessary methods for subsequent model testing and provides valuable insight into how stereo-depth perception is limited by natural image variability.

1.2. Psychophysical approaches

1.2.1. The Marr-Poggio approach

Psychophysics has long been a staple of the perceptual sciences. Throughout perceptual science's history, its psychophysical approach has undergone several transformations. Some changes were methodological, such as the adoption of signal detection theory, while others were more philosophical, shifting the field's investigative focus, as seen in the computational theory proposed by Marr and Poggio.

Although psychophysics has a longer history of efforts aimed at uncovering perceptual mechanisms, the theories proposed by Marr and Poggio have largely influenced contemporary views of what psychophysics can uncover mechanistically, consequently shaping various psychophysical approaches taken. The philosophical theory of analysis proposed by Marr and Poggio, known now informally as "Marrs' three levels of analysis," is key to understanding these approaches. According to this Marr-Poggio theory of analysis (Marr, 1982; Marr & Poggio, 1976), the three levels at which a complex system, such as the brain, can be understood are:

1. computational—the objectives that the system is trying to accomplish
2. algorithmic—the algorithms the system employs to meet the objectives, and
3. mechanistic—the physical implementation of the algorithmic.

Moreover, the various sub-disciplines of neuroscience are concerned with and engaged in a specific level(s) of analysis. For example, neuroanatomy is concerned with the mechanistic level, neu-

rophysiology engages with the mechanistic and algorithmic, and psychophysics engages with the computational and algorithmic level.

While the three levels of analysis themselves are the most well-known aspect of Marr and Poggio's theory, these levels were in fact axioms of a larger thesis. Namely, that "in the eventual understanding of perceptual processing [...] although the [computational] level is the most neglected, it is also the most important." (Marr & Poggio, 1976). In their words, "the structure of computations that underlie perception depends more upon the computational problems that have to be solved than on the particular hardware in which solutions are implemented." Their theory provided no alternative means of applying computational-level theory, suggesting that psychophysical methods were the most crucial in understanding the structure of computation—that is, algorithms.

Importantly, Marr and Poggio also provided a framework whereby psychophysics could be applied towards algorithmic-level investigations. This framework can be roughly sketched as follows:

1. form an ecologically informed computational-level theory of perception,
2. develop a model of how the perceptual system behaves under this theory,
3. measure and compare performance between the model and a real observer.

Then, when models and humans are tested at their limits, they will produce distinct performance outcomes which will either align in support of the theory, or misalign in support of an alternative.

1.2.2. A note on terminology

What is meant today by "computation," "algorithm," and "mechanism," can be hard to decipher and depends on context. Before continuing, their use is made clear.

The word "computation" is synonymous with the word "calculation." It is the object of analysis withing the the levels of analysis, hence the levels are called a "theory of computation." Descriptions existing at any of the three levels describe the overall computation, albeit at different levels of abstraction.

A "computational model" has at least two meanings. First, the most general usage of a "computational model" refers to models that are built programmatically using computers. This is its most common usage. Second, models can be "computational" in that they describe said system according to the respective Marr-Poggio level of analysis. Because models at the computational level are abstract, they are not "computational" in the sense of being built programmatically. Thus, when the term "computational model" is used it should be implied that the model is at least at the algorithmic level, but it can also include mechanistic-level aspects.

An algorithm describes the process of performing a computation. There are many levels of detail in which one could describe an algorithm. At one level, an algorithm can be viewed as a function—a mapping that takes a set of inputs and maps them to outputs. At a finer level of detail (separate from the levels of analysis), an algorithm can be described by specific sub-processes, or a sequence of mappings or functions that compose the more general function-level description; depending on the algorithm, there could be various levels of description based on its structure. In this work, the use of algorithm will refer to the highest level of description—a function.

1.2.3. Modern approaches

The Marr-Poggio levels of analysis are, and continue to be, a useful analogy for analyzing complex systems. Nevertheless, most of the stronger assumptions regarding these levels have received justifiable criticism and are now considered outdated by many criticisms such as:

1. the computational level is the most important from the information-processing point-of-view and reveals more about an algorithm than the mechanistic-level—mechanistic-level investigations have revealed many important details about the algorithmic-level (Peebles & Cooper, 2015);

2. the goals of the visual system are clear cut and can be made *a priori*—the goals of the visual system are contextual and task driven (Warren, 2012), and can be confounded evolutionary by-products (Anderson, 2013);

3. the levels of analysis are only loosely connected—there is strong evidence that neurological

(mechanistic level) constraints limit what types of algorithms are practical or even possible (Pillow, 2024).

Modern adaptations of the Marr-Poggio framework—modern algorithm-oriented psychophysical approaches—have taken these criticisms into consideration. Some of these approaches have motivations less concerned with computational-level theory and are more data driven. Others have incorporated mechanistic-level constraints into their computational models—for example, ideal observer models are designed to solve specific tasks optimally under biological constraints (Burge, 2020; W. S. Geisler, 2003). Nevertheless, the underlying principle of the Marr-Poggio framework remains—that through psychophysical measurements and computational modelling, the algorithms underlying perception processing can be uncovered.

1.3. Motivating thesis

Critiques of the Marr-Poggio levels of analysis have emerged from extensive research conducted across various sub-disciplines of neuroscience since the theory's inception. Broadly, this research has increasing shown that the visual system possesses a substantially more intricate and diverse architecture than assumed during Marr and Poggio's time (Kreiman & Serre, 2020). While contemporary algorithm-oriented psychophysical investigations have adapted to these criticisms, there remains a subtle tradition in modern approaches that has largely avoided criticism. Specifically, many contemporary algorithm-oriented approaches operate under the assumption that various stages of processing—namely, components of the underlying algorithms—can be effectively isolated through the presentation of idealized stimuli. From psychophysical outcomes measured in response to such stimuli, it is assumed that fundamental insights into the algorithmic structure of visual perception can be reliably uncovered. While these assumptions are well suited for the older, simpler view of visual processing, the more complex view is likely in conflict with these assumptions, suggesting a need for an alternative approach.

Here a review of visual processing is presented where the various factors that challenge the extent to which algorithmic mechanisms of visual perception can be isolated through psychophysical methods. Firstly, a review of bottom-up, stimulus-driven processing reveals its holistic, context-

dependent nature, suggesting highly distributed visual processing *across* visual cortex. Secondly, an examination of top-down, goal-driven processing reveals the task-dependent nature of psychophysical tasks, indicating a necessarily task-dependent component of visual processing *outside* of visual cortex. Together, these observations suggest that traditional algorithm-oriented psychophysical approaches may not effectively isolate various stages of visual processing in a coherent manner.

1.4. Visual Processing

1.4.1. Stimulus-driven processing

Vision is useful because it informs action within the environment (Nakayama et al., 2022). A robust visual system enhances survivability by enabling organisms to accurately perceive and interpret their surroundings, which is crucial for making informed decisions and taking appropriate actions. While the overall goal of the human visual system is to facilitate action, one intermediary goal is a rapid and precise reconstruction of the immediate scene—the region of the environment that is most actionable.

Scene perception starts with sensation at the retina, after light has passed through the optics of the eye. The image projected onto the retina is a representation of the region of the scene that produced it—it is not the region of the scene itself, but a photon-based description or encoding of it. Because the retinal image is an incomplete description of the the scene, constrained by factors such as the eyes' limited field of view, the visual system is tasked with inferring many of its ambiguous details. The most optimal approach that the visual system could take is by following a chain of inference tracing backwards from the retinal image to the scene (W. S. Geisler, 2003, 2011; Maloney & Mamassian, 2009). In this process of inference, the retinal image is termed the proximal stimulus, while the physical regions of the scene that give rise to this proximal stimulus are referred to as the distal stimulus. Further inferences about the scene can extend beyond the visible region—the distal stimulus—but are best achieved when some aspects of the distal stimulus are inferred from the proximal stimulus.

Depth structure is one distal property that must be inferred from the proximal stimulus. When the

spatially three-dimensional scene is projected onto the retina, it transforms into a two-dimensional image, losing its depth dimension. Generally, inference relies on various visual cues—regularities in the proximal stimulus patterns that relate to properties of the scene, such as depth. Depth cues are either binocular, which rely on differences between stereo-halves of the retinal image, such as binocular disparity, or monocular cues such as texture gradients, lighting, and shading.

Beyond the retina, the brain detects and interprets visual cues, mapping them into estimates of various properties of the distal scene. In visual cortex, the visual system builds a hierarchy of these estimates—or property maps—that contribute to forming percepts of the distal scene and can aid in constructing other property maps (Seilheimer et al., 2014). As processing continues, there is a distinct progression in scene representation through these property maps, advancing from points to contours, to shapes, and eventually to objects.

At the retina, processing units are at their smallest and most spatially localized. As processing ascends the visual hierarchy, units become larger, integrating more information across the visual field and becoming less localized. However, the processing of property maps is not solely feedforward. To fully utilize available information, the visual system contextualizes property maps by integrating inputs from the same and higher levels of processing, facilitated by lateral connections and feedback/recurrent connections respectively. As processing moves up the visual hierarchy, processing units become larger, integrating more information across the visual field and becoming less localized as it does so. However, the processing of property maps are not entirely feedforward. In order to take full advantage of the information available, the visual system contextualizes property maps. This is done by integrating information at the same and higher levels of processing, that is, from lateral and feedback/recurrent connections respectively (Angelucci et al., 2017; Siu et al., 2021).

Implications

At the time of Marr and Poggio, the visual system was seen primarily as a feedforward process (Kreiman & Serre, 2020). From this perspective, individual processing steps were predominantly distinct and localized, making it suitable to attribute simple, idealized property maps to specific processing stages. In line with this perspective, a traditional approach involved identifying and

describing these property maps or "units of perception" through measurements of psychophysical responses to simple stimuli—stimuli designed to embody the idealized nature of these units. Modernly, this approach is still taken, but some degree of contextual effects are argued to be controlled for by presenting small, simple, and/or briefly presented stimuli. However, this approach assumes that context does not fundamentally alter these apparent base, localized units of perception—that context is ancillary and can be optionally perceived as separate from these units.

Research conducted since Marr and Poggio has highlighted the importance of global scene context over localized details. Visual scenes have been shown to be perceived in a "coarse-to-fine" manner, where more localized details become discernible only after several iterations through the full feedfowrward pathway of processing (Hegde, 2008; Johnson & Olshausen, 2003). This finding may seem paradoxical given that steps in the visual feedforward hierarchy progress from smaller to larger computations. However, this can be explained by recognizing that local features may have diverse interpretations dependent upon their larger contextual framework—localized computations might not be particularly informative in isolation, serving primarily as intermediaries for larger context in earlier processing stages. For instance, flat 2D retinal image content is not particularly useful until it has been appropriately integrated into the 4D spatio-temporal structure of the scene. Moreover, environmental regularities are more effectively applied to objects rather than their individual features—scene regions corresponding to the same objects are constrained to move in unison and likely to be of similar physical composition. Overall, the visual system's global-to-local processing suggests that basic units of perception are global not local and that localized perception is highly dependent upon context. Together, these imply that psychophysical outcomes are not easily localized within visual cortex.

1.4.2. Task-driven processing

Vision is useful because it informs action within the environment, which ultimately facilitates the evolutionary fitness of the observer. Given an awareness of the estimated distal stimulus—a representation of the scene—the observer is able to interface with the environment.

Action is both stimulus-driven (bottom-up) and executive goal-driven (top-down) (Turner et al.,

2019)—the observer acts according to what is perceived to exist and according to their goals. In order to accomplish a goal, the observer follows some strategy or set of strategies. For a given strategy, the observer may need to accomplish a number of perceptual, cognitive, and behavioral tasks. For psychophysical tasks, the primary goal is to choose the correct response, which can be obtained by extracting the relevant information—the latent variable–from the distal stimulus.

Visual attention (in part) is a set of processes that mediates the extraction of relevant stimulus information by selectively directing visual information—e.g. property maps—from early cortical processing into working memory (Pashler et al., 2000). Because the computational capacity of working memory is limited, the visual system must prioritize what perceptual information is pulled into working memory. Priority is determined based upon properties of the stimulus and executive goals (MacLean et al., 2009). Together, both stimulus-driven and goal-driven processing compete and induce biases in what ultimately gets encoded into working memory (Deco & Rolls, 2005; Soto et al., 2012).

Once the observer has the relevant stimulus information within working memory, the latent variable can be estimated. The latent variable may not directly relate to the extracted visual information. Under such circumstances, the observer is able to use other mechanisms—such as those related to decision making—to compute the latent variable from available information. In any case, the observer arrives at some conclusion of what was observed which is stored in working memory.

Implications

One traditional view of visual attention is that its capacity is the only behaviorally limiting factor of significance; the degree to which a property map is actionable—that is, perceptible in any behaviorally, measurably relevant way—is only limited to the degree attention is able to select it. However, the degree to which a property map is actionable is also dependent upon which visual attention pathways exist and the extent to which attention modifies the selected map.

Because not all processing outputs—such as intermediary processing steps—have *direct* use in goal-oriented tasks, and because pathways have energetic cost to generate and maintain, not all maps

should be equally perceptible. Indeed, attentional selection should only have access to a limited, curated set of pathways that provide use in natural tasks and grounded within the reconstructed scene and its objects. While physiology might suggest that certain property maps exist, they may not be directly accessible by working memory. Thus latent variables corresponding to such property maps may in fact be computed from other property maps via executive level processing.

Further, studies have shown that attentional processes not only select property maps, but may also exaggerate them, making them appear different than they would otherwise (Carrasco, 2018). It is speculated that this exaggeration acts as contrast gain or feature enhancement to increase discrimination sensitivity in a task dependent fashion. Indeed, different discrimination tasks would benefit from different latent-variable-dependent types of gain, and estimation tasks may not benefit from any type of gain at all.

Collectively, these findings suggest that psychophysical outcomes may not be as straightforwardly attributed to visual cortex as previously assumed. The influences and constraints of goal-driven processing are likely to confound the effects of stimulus-driven processing, thereby preventing their isolation through psychophysical means.

1.5. Conclusion

The estimated latent variable is of particular interest to psychophysics based vision research. It is what is perceived. Thus it is what psychophysics attempts to recover even if it is not directly measurable. Naturally, statements regarding visual perception are often directed towards the decision variable, or in terms of the estimated latent variable, or its decision correlate—the decision variable. The degree to which psychophysics can make claims about visual processing ultimately relies on what the estimated latent variable represents.

In a psychophysical experiment, the estimated latent variable is goal-driven and specific to the psychophysical task. Note however that scene properties estimated for distal-stimulus estimation can also be construed as a latent variable, but one that is stimulus driven. These two latent variables are distinct in their utility and their order of processing. The stimulus-driven latent variable is

general purpose and estimated early in the visual processing hierarchy, whereas the goal-driven latent variable is task specific and estimated later. A traditional view of psychophysics makes little distinction between these two latent variables. While distinct estimates of both latent variables might be approximately equivalent, one should be careful in assuming that they are. In order for equivalence between both latent variables to be likely, a number of assumptions—as discussed above—must hold:

1. the executive latent variable has correspondence with a single property map or has a linear combination rule with all of its constituent maps;

2. the property map is persistent or is the final product—is not simply a temporal intermediary before contextualization and recursion;

3. attentional processing has direct access to the property map, is able to sufficiently isolate the map, and does not significantly modify it.

While these assumptions don't necessarily have to hold for psychophysical results to be relevant, they do apply when making statements about specific stimulus-driven processing. If correspondence between the two different latent variables hold true experimentally, then results can be made attributable to processing mechanisms that are isolated to a particular stage of the processing hierarchy. If not, then visual processing is much less straightforward.

This view that the (executive) latent variable has a clear relationship with early cortical processing is an old assumption that is still somewhat common today. While it may not be explicitly stated, it directs a fair amount of algorithm-oriented psychophysics-based work. Again, this view is that visual cues are able to be distilled into some fundamental unit of computation which can be studied in isolation by the use of simple stimuli—simple stimuli which more or less represent that unit of computation. However, given the complex, distributed, and holistic nature of visual processing, where scene and task contextualization are paramount, this idealized view of processing seems increasingly less likely. While behavioral work from this approach is likely valuable, its value may be overstated in its ability to inform mechanisms underlying visual processing.

Together, these implications suggest that an alternative approach to algorithm-oriented psychophysics be taken into consideration, one which is indifferent to idealized notions of representation and views visual processing in context of its complex holistic nature. In the next chapter, the limitations of the traditional approach are made more explicit, from which a new approach is developed. The practice of psychophysics is straightforward—measure changes in behavioral responses to physical changes in stimuli—but its implications are profound—a precise quantification of perception.

CHAPTER 2

THEORY

While it is clear that psychophysics itself is likely the best tool available to probe one aspect of this level—its inputs and outputs—it is unclear to what extent this postulate holds for algorithms. In this chapter, we discuss the weaknesses of traditional algorithm-oriented psychophysical approaches, then present a theoretical basis for an alternative. A general approach for application of this theory to psychophysics is then prescribed.

2.1. Limitations of psychophysics

Neuroscience has undoubtedly detailed many important findings about the brain, yet continues to grapple with understanding the algorithmic workings of neural systems, even in the simplest organisms such as *C. elegans* (Antonopoulos et al., 2016). Although the internal workings of neural systems can be examined and many specific details are known, it remains unclear which of these details are relevant to algorithmic investigations. A problem with the traditional visual psychophysics approach is that one has to assume which idealized stimuli are most relevant, with no clear way of verifying whether the assumption is valid, even though such stimuli may be highly non-representative of stimuli typically encountered in the real world. Given the complexity of the systems being investigated, this approach risks focusing on incidental phenomena or misinterpreting highly nuanced detail. While these concerns may not necessarily materialize, a different approach is valuable for verifying whether they will. A new approach should aim to minimize assumptions, that is, to investigate these systems as black boxes where the inner workings of the system are treated as unknown.

Unfortunately, it is likely that typical black-box systems identification methods are insufficient, as they seek to fit models that fit expected values. For most scientific models, this is not so much an issue—it is less important that the specifics of a model are correct, but that the model does well to characterize the data. This is expressed in one respect by the familiar adage: "all models are wrong, but some are useful (Box, 1976)." However, for algorithmic modelling, the specifics of the model

are not just meant to capture trends in the data, but computational or algorithmic underpinnings. Thus, while statistical significance may indicate a model is well representative of outcomes—that the data can be well captured by the model—it is not a good indication that the underlying algorithm can be captured by the model. In other words, model fitting is not algorithm fitting.

As a simple illustration, consider the data generated by functions $y = x + \epsilon$, $y = \frac{4}{3}\log(x+1) + \epsilon$, $y = \frac{4}{1+\exp(-x)} - 2 + \epsilon$ between the interval $0 < x < 1$, where ϵ represents a noise term (see Fig 2.1) All of these functions, can easily describe each others data, but have fundamentally underlying mathematical representation. Indeed, many other functions can potentially describe the same data just as well.

Some have attributed this general issue to simple stimuli—that performance measured by simple stimuli are too easily characterized mathematically (Sebastian et al., 2015). While this is true to some extent (and described in more detail later), the issue is much more systemic. Indeed, in the machine-learning field, it is well known that different models can accomplish the same task equally as well (Geirhos et al., n.d.). However, this is an understatement—the space of potential algorithms that can solve a given problem from a set of inputs and outputs is infinite (Gödel, 1931). And while details about the specific internal workings of the systems can limit the number of potential algorithms, unless the system is fully specified, that number will always be infinite. While this sounds daunting, there exists clear theory can be used to narrow down the set of potential algorithms and perform perfectly reasonable model testing. However, for this to have any real impact, analyses

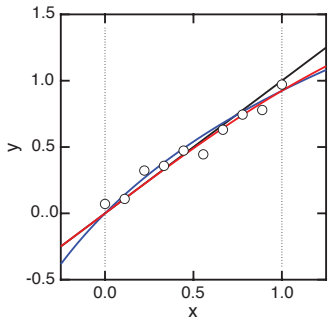

Figure 2.1. Simple illustration of how different functions can describe the same data. Shown are functions $y = x$ (black), $y = \frac{4}{3}\log(x+1)$ (blue), and $y = \frac{4}{1+\exp(-x)-2}$ (red).

need to move away from performance analysis and move toward algorithm analysis.

2.1.1. Algorithmic theory

In order to fully asses a model of an algorithm, what is needed is a comparison between the structure of the model and the true underlying algorithm of the visual system, not a succinct simplified comparison of how they behave, such as performance thresholds. Thus, algorithms themselves need to be considered as data and compared. Algorithmic probability theory provides the mathematical framework for this type of analysis. (Rissanen, 1978; Rissanen, 2007).

Both the model algorithm the visual system algorithm can take on the characteristics of data by being written as code—a string of characters. However, this is not immediately straightforward. There are infinite ways to encode the model algorithm and no way of accessing the true algorithm. What can be done nevertheless, is to use data generated by the true algorithm—the observer data—as a metamodel or description of the true algorithm, and the data generated by the model as a description of the model. This is possible because at its core an algorithm (according to the definition given in Chapter 1) is a mapping that can be fully represented as the complete set of its inputs and outputs. When the data fully covers the domain and its image, it represents a complete description of the mapping. Because the data for both are limited, the observer data is an incomplete description of the visual system algorithm and the data generated by the model is an incomplete description of the model algorithm.

In order for these to be descriptions to be complete, their lengths—the amount of data gathered—must be infinite. This must occur in order to have complete coverage of each mapping's domain and image (irrespective of noise). Incomplete descriptions can be descriptive of other models, and are thus model classes or sets of models/descriptions that fit certain characteristics. This is to say, they contain certain elements. Because there is an inherent limitation to how much data can be collected, one cannot with perfect certainty determine a single underlying algorithm via data collection. However, one can narrow in on a specific set of model classes, whose differences may have no meaningful value.

2.1.2. Model testing

The set of infinitely possible models that can represent an algorithm of visual processing can be further narrowed by asserting criteria for model admissibility based upon similarities between observer and model responses. By imposing such criteria, one rules out certain classes of models distant from the true underlying algorithm. The more unique criteria and the more informative those criteria are, the nearer the model is likely to be to the true algorithm (and the more representative each incomplete description is of the true model which they describe).

The number of unique trials (approximately) specifies the number of unique criteria. For every unique trial, the observer will respond in a very particular way, driven by their underlying visual processing. Each response therefore represents a nuanced output of the particular algorithm used, and thus potential criteria for (in-)admissibility of the model.

The number of unique criteria does not imply those criteria are tested, which is why it is simply not enough to measure performance to complex stimuli. For experiments that only differ in terms of their independent variables, testing a model's goodness-of-fit ensures all unique criteria are tested. However, if said experiments had trials with differences, the number of unique criteria tested is less clear. Here, the number of unique criteria is equivalent to the number of unique trials,but they are not tested directly, but as a group. This second case has more (and potentially substantially more) applied admissibility power than the first, but not as much as it could. If the model was used to predict responses on a trial-by-trial basis, then this would be a maximal use of its available criteria.

Some criteria will test more model classes than others, thus be more informative than others. While it is not clear which criteria will be most informative *a priori*, in general, how different a trial is from previous tested trials tends to reflect informativeness. If trials are similar, then such criteria are likely to test only a narrow range of algorithm classes and/or do so weakly. However, similar or identical trials are informative in a different way. If there is noise internal to the system—which there is for biological vision—it is possible that noise will impact the observer's response, thus confounding the underlying structure of the algorithm. Therefore, if there is noise in the system,

there will be some level of uncertainty as to whether or not criteria were met by nature of the algorithm or as the result of the noise. Similar trials can inform as to whether that was the case by acting as samples in the estimation of the true response, and thus reduces the degree of uncertainty regarding the tested admissibility of that particular condition. This implies a trade-off in the type of uncertainty reduction afforded by trial similarity—high-diversity affords exploration, similarity affords certainty.

Traditional psychophysics experiments, whose trials differ only in terms of the independent variables, represent a near extreme in (non-)diversity. They have particularly strong power for testing the admissibility of individual conditions, but they are particularly weak in narrowing down the set of possible algorithm classes. At the other extreme are experiments with a completely diverse set of stimuli that could potentially maximize the number of algorithm classes tested, were it possible to ascertain their noise structure. Both extremes provide substandard utility in narrowing the set of possible algorithms. A better alternative is to be found somewhere between these extremes where both noise structure and algorithm space can be well ascertained.

This framework also describes why complex-stimuli are more informative than simple-stimuli. In terms of algorithmic constructions, information is not quantified probabilistically (although it can be), but in terms of shortest encoding length, quantified as algorithmic complexity. Thus, the general informativeness of criteria (a type of information gain) is naturally quantified in terms of complexity as well. Simple stimuli are also simple in terms of their algorithmic complexity. Because simple stimuli have short descriptions they are likely to be less informative than complex stimuli.

While these abstractions may seem very specific to algorithms, it also applies to psychophysical findings that make inferences about internal processing and not simply their outputs. Here, testing narrows in on a larger set of algorithms, rather than attempting to narrow in on a specific one. Further, one might consider descriptive based work also providing criteria.

2.2. Simple stimuli

One conclusion to draw from this discussion is the potential pitfall of relying too heavily on a small, non-diverse set of stimuli to explore a complex algorithmic space. With that appropriately considered, simple stimuli and their results can be potentially powerful. First, criteria based on complex stimuli do not necessarily cover the same model classes as those based on simple stimuli. They ought to be used as well. Second, criteria based upon simple stimuli offer special insight into complex models. Estimators lose some of their potency to estimate very particular phenomena as they become more generalized, as characterized by the bias-variance tradeoff. When an observer/model poorly estimates a property in a simple stimulus, this may say something particularly insightful about the algorithms being implemented that may not have been revealed otherwise. Thus simple stimuli have particular, and perhaps necessary utility.

One aversion a experimenter might have to a highly diverse set of stimuli, is the desire to have highly interpretable results. The advantage that simple stimuli have is that they afford the experimenter control to do so. As will be discussed in the next chapter, stimuli can have their complexity reduced while maintaining a high degree of variability, allowing the space to be explored while maintaining moderate interpretability. This might offer addition insight into how system dynamics and perceptual outcomes change as the stimulus space becomes more complex—in particular, how results track as the stimulus space increases in complexity.

2.3. Alternative Approach

In summary an analysis of algorithms suggests that one should adopt several refinements to psychophysical approaches that seek to uncover underlying algorithms: frst, a more diverse, sufficiently complex set of stimuli, and; second a procedure that allows the exposure of internal noise structure. Naturalistic stimuli are a strong choice of stimuli because they are relevant, are less likely to produce cognitive confounds, and because they are diverse and vary in range of complexity. Natural images, their relevance to perception and their use in psychophysics is the topic of Chapter 3. Multiple pass procedures, coupled with correlation methods (discussed thoroughly in Chapter 4) are a balanced way of inferring noise impact while exploring the algorithmic space. In Chapter 4, I present an

empirical study that utilizes these methods in the investigation of stereo-depth perception. There, I also develop methods that are able draw out more specific relationships between inputs and outputs, which can potentially be used to make more fine-grained criteria for admissibility.

CHAPTER 3

NATURAL IMAGES

Natural images are the substrate of natural visual processing. They are not just the stimuli that are encountered, but the stimuli that the visual system evolved to process. If one is to understand how vision works in the real world, then an understanding of how the visual system processes natural images is required. Whether psychophysics is applied to algorithmic or purely behavioral interests, natural images should play a key role.

3.0.1. Formation

Natural images are a limited perspective of a natural scene. The process by which they are produced begins by following the paths of light and their interaction with surfaces in the scene. In natural scenes, sensible light interacts with or is emitted from surfaces within the scene. Every point in the scene scatters light uniquely in all directions. The light travelling in the same direction trace out a light ray, described as an arrow that juts outwards from the surface towards infinity or until it intersects with another surface. Light travelling along each ray may be described by its general intensity or radiance, wavelength, and polarity.

The set of all light rays at every location in the scene forms a field. Surfaces placed in the field will receive light from other surfaces whose rays intersect with it (the surface). The light rays at the points of intersection across the surface form a projected image. Only a small subset of the field rays within the field of view of the projection surface form the projected image. If these particular rays are traced backwards to the surfaces from which they last interacted or were generated from, this set of surface points where they intersect form the preimage of the projection surface—representing the portion of the scene that was visible to the projection surface.

If the surface has sensory machinery, then the projection image can be captured to produce a sensory image. For photographic images, the projection surface is a plane, and light is recorded at individual sensors which correspond to pixels in the projected image. For retinal images—the basis for human

vision—the sensory surface is curved, ellipsoidal surface and photoreceptor cells act as sensors.

The image produced at the sensory surface has (at least) two representations. There is the image represented by photons the moment before they interact with the sensors, and there is the image represented by the sensor responses. In this photonic representation, the image is encoded by the quality of the light it receives. In the sensory representation, the particular form of the representation depends on the sensor. For for digital photography, this is an electronic code. For human vision—a neural code.

3.0.2. Content

The content of a natural images is best characterized in terms of the content of scenes they represent. Scenes are composed of space, physical objects and light. Objects include anything that has a physical surface, including the ground, backdrops, and particulates, such as mist or dust. The composition of these objects—particularly their surfaces—determine how they interact with light or whether they emit light. Object composition, position and orientation in the scene and relative to each other objects determine the light field which uniquely characterizes the scene. The content of the natural image is also based upon the particular perspective from which the scene is viewed. The perspective depends on the position and orientation of the projection surface and the distribution and properties of its light sensors.

Here, the reader might object (no pun intended) with this definition of natural scenes—that they are comprised of objects. This might be raised because the term "object" can be used synonymously with a classification or as a specific instantiation of a classification, which are perceptual abstractions. Using the language of Kant, these objects that compromise scenes are noumenal—they exist outside perception (Kant, 2009). They are real, and are not fully represented by any object other than themselves.

3.0.3. Space

The term "natural" in natural images can be used to denote images produced ecologically—i e. under normal viewing conditions—or images of content produced by nature. While the latter is the

most common definition (Carlsson et al., 2008), these types of images describe a limited portion of images relevant to humans. Images containing artificial content—like human-made objects and structures—are not natural by this definition of "natural," yet encountered regularly by human observers. In fact, for human observers in highly populated cities, real-world encounters with natural-world content can be quite rare. Because this work investigates human vision, the term "natural images" will take on the ecological definition rather than the traditional one, and the latter will be simply referred to as "images of nature."

Images of nature are often described as a small subset of all possible images, meanying they occupy only a limited portion of (full) image space. Additionally, they have binary classification: either they are images of nature or not. This definition is inherently problematic on its own account. If such images contain some degree of human made artificial content, a binary distinction becomes more difficult. To remedy this, one might dispense the binary classification, as will be done here with the ecological definition of natural images.

Here, natural images are defined to include all of image space, which is necessary, but not problematic given its non-binary nature. By including artificial objects in the space, digital displays and their graphical content are inadvertently included. Because digital displays can more or less present any image, natural image space must be inclusive of all images. Therefore, natural images span all of image space. However, natural image space is made distinct from the full image space by the inclusion of ecological relevancy by means of added statistical structure. Namely, image space can be defined as a probability space, whose probabilities describe how probable or likely images and/or their properties are to be encountered. This can be used to place images, their properties, and their viewing conditions on a spectrum of artifical to natural, rather than assigning them a binary category. Although, one may map the space into a binary categorization based on whether an image is more natural than artificial.

All natural images are produced by a perspective projection mapping from scenes to surfaces, or that natural image space is a projection of the set of all natural scenes—the natural scene space—onto the set of all possible perspectives—the perspective space. The details of these constructions

are less important than the fact that these spaces are much larger than natural image space—which may seem counter-intuitive. One reason for this is that perspectives are constrained within a natural scene, but mostly because because perspective projection results in an extreme loss of information (i.e. perspective projection is surjective whose codomain has a much smaller dimensionality than its domain). This latter reason can be intuited by considering that natural images are four-dimensional (x and y positions and luminance, color), whereas the light-field is at least six-dimensional (x, y, z, azimuth and inclination heading, radiance, and wavelength) *per ray of light*. Indeed, there is an enormous loss of information between natural scene space and natural image space to say the least, even without other losses such as limited image resolution. This highlights the immense under-determinism in trying to reconstruct the preimage from its corresponding projected image that visual systems face.

3.0.4. Variability

What is distinctively powerful about the human visual system is its ability to make useful and consistent inferences about the scene from cues despite the high degree of variability in the retinal image. For example, there are many regions of an natural image or across many images that correspond to scenes regions with the same underlying depth value. The variability in content across these images can be extreme, yet the visual system rarely seems fooled.

The variability of natural images is often denoted as nuisance, in that it detracts from the objective of specific tasks. For example, in depth estimation of a specific region in a natural image, anything that is not a depth cue provides no relevant information and can only detract from the task. However, natural image variability is different from noise in that it may be useful elsewhere, or integral to a larger task at hand. An observer making a depth estimate of an object in a scene needs also to identify the object itself, for which non-depth cues are entirely relevant.

Note the circular nature of this example. Here depth-estimates themselves—being part of spatial reconstruction of the scene—are also useful in object identification, as stated earlier. For this particular task, it may seem that the visual system would have to either ignore depth-information at the cost of performance, or do something fairly more complex. The case is in fact the latter.

Indeed, this provides a normative explanation for why recursive and lateral processing exists.

The ability of the visual system to effectively handle natural image variability and the complexity it imposes, should be considered its primary feature. The complexity nature of natural images and scene perception has accompanied the evolution of eye the over the past hundreds of millions of years. It is perception of the natural world, with its intricate detail and variability that the visual system has evolved to process.

3.1. Psychophsyical Approaches

Natural stimuli are fairly complex. They can be both difficult to procure and present in a way that is natural. Depending on several factors, the experimenter, at some point, will need to concede what is more natural for what is practical. In some cases there is some flexibility in choosing what aspects will be unnatural and which are natural.

Here we consider and two general approaches—one that uses fully natural stimuli, the other, a hybrid between artificial and natural—naturalistic.

Natural

The primary objective of a fully natural psychophysical approach is to evaluate how observers perform under natural stimuli and conditions, with the most direct method being to measure vision in the natural environment. Most investigations that have attempted fully natural approaches have sought to gauge perceptual experience (subjective psychophysics) rather than performance (objective psychophysics), and hence the scope of this work. Fully natural approaches of the latter type are quite difficult, and rare. For those studies that exist, the number of trials conducted are relatively few.

One challenge with this approach is its logistical demands. Stimuli either need to be physically set up or identified within the environment. Measurements of stimuli ground truth is likely also a requirement, and may require that physical measurement within the environment be performed. The lack of environment control means that the dynamics of the environment, such as lighting, may alter the stimulus before response can be measured.

Adhering to the strict natural requirement implies that tasks should also be natural. Traditional psychophysical procedures, such as the two-alternative forced choice (2IFC) method, would not be suitable. While it is clear that tasks should resemble everyday activities, integrating such tasks into an experiment poses difficulties and restricts the types of analyses that can be performed.

Under natural viewing, observers are not physically constrained—observers have the freedom to position and orientation of their head, body, and eyes. A fully natural approach should lack control of how an observer views the scene. Although this was previously a significant limitation, advances in motor and eye-tracking technology now allow for the tracking of observer movements and visual focus (Carlsson et al., 2008).

Robust psychophysical measurement within the environment is a lofty goal. Given the difficulty of performing psychophysical investigations within the natural environment, it is likely that new methods should be developed. The extreme variability of the environment demands a much higher number of trials in comparison to traditional studies, but its physical constraints require data to be gathered at much slower pace. While this goal may be distant, it provides a benchmark from which other naturalistic experiments should be gauged.

Naturalistic

Given the definition of natural stimuli presented earlier this chapter, one might say that all psychophysical experiments are at least weakly natural, some potentially more naturalistic elements than others. Nevertheless, classic psychophysical experiments are among the most unnatural.

The desire to make experiments more natural is not a new idea (Felsen & Dan, 2005). Many naturalistic experiments have been done with this aim in mind and others have been done simply out of relevancy to the questions being asked. There are many approaches.

The general approach of Burge and Geisier has been to use natural images using classic psychophysics. Their work includes investigations of defocus blur, speed, slant and tilt, all with accompanying ideal observer analyses (Burge & Geisler, 2012; Chin, n.d.; Kim & Burge, 2018; Sebastian et al., 2015). This approach attempts to sit at the cusp of what is known about human

perception from classic experiments, and what is not known—although not entirely unique to them (Brainard et al., 2018). It attempts to manage the challenges of using natural stimuli just enough such that there is room for clearly interpretable results, yet exploration. As such, performance data generated using this approach are good substrate for algorithm assessment or model validaton.

Of particular note is work done by Chin and Burge in their 2020 study. This study also included multiple pass procedures, making it a particularly good fit for algorithm assessment as discussed in Chapter 2. There, they included an assessment of their ideal observer model, but not to its fullest extent, given the data they had available. Methods developed in Chapter 4 could be used to perform a more stringent and more informative assessment there.

The work presented in Chapter 4 also contributes an investigation of binocular disparity to the stream of work by Burge and Geisler, and accompanies their binocular disparity ideal observer (Burge & Geisler, 2014).

3.2. Procuring natural images

There are different ways to procure natural or naturalistic images. Methods generally fall into two categories: photographic and computer generated. Neither is generally better than the other, but there may be clear choices depending on experimental needs.

Photography

The most straightforward way is through photography. Photographs are a strong choice because they are natural images from the real world. Generally they have a few limitations, namely resolution and range of luminance and color levels. Because they will be presented to observers with their own optics, its ideal that images have minimal lens distortion or blurring.

Another disadvantage of photographs is their static nature, and scenes usually have some dynamic elements, even if minor. For monocular movement, there are ways to generate stimuli with movement by simply moving and/or scaling the stimulus images within the presentation environment. (Burge & Geisler, 2014). With move complex scene movement, this might be done with by morphing stimulus images, but this typically requires more knowledge about the 3D structure of the scene

(this is discussed in more detail in Chapter 4). The best option here would be to use video of natural scenes rather than individual images. However, these options only account for movements when an observer is completely still. Even when an observer holds as still as possible, an observers uncontrolled eye and head movements generate small optical and parallax changes in the scene. There are ways to track both coarse and fine movements with high eye tracking, and potentially ways to morph images to account for these effects—but also requiring more information about the 3D structure of the scene—however this is something that is rarely, if ever done.

The easiest way to obtain natural images is through existing photography. Many datasets exist of natural images, but their quality may have varying quality. Some datasets exist that have images gathered from the web, thus might have inconsistent qualities. Generally, natural images that are not generated for natural image study are likely to be prone to photography biases, which may or may not be useful or irrelevant.

Alternatively, there are many advantages of using specialized datasets for natural image study. Foremost, their creators are likely to have considered natural image space and natural image presentation, therefore likely to have capture exceptionally high quality images and calibration for such work (W. S. Geisler & Perry, 2011; W. Geisler et al., 2001; Tkačik et al., 2011). They are also likely to have additional features. For example, the dataset used for the study in Chapter 4 is in stereo with co-registered range data (A. V. Iyer & Burge, 2018). Depending on the questions being asked, this additional information is invaluable in that it can represent ground truth information, which is difficult, or even impossible to get through photography otherwise.

The lack of ground-truth information is perhaps the biggest motivation to procure a new dataset. This however can be quite costly, time consuming, photography expertise, and careful deliberation in order to do well. In order to produce a highly diverse image set, there is a lot of travel that may be required. In order to obtain accurate measurements (or even measurements at all), specialized equipment is required. Additional effort is also required to co-register data with the photographic images, which itself is a laborious process.

One difficulty with co-registered data is that it can be prone to error. Some is at least expected to some degree (A. V. Iyer & Burge, 2018). As best as possible, photographs and additional data need to be captured at the same time and place. If photographs and data do not aligned, then there are likely to be small deviations between data and images. For example, small movement or scene lighting changes can occur within temporal lag. Spatial lags are expected to cause data or images to have more or less vantage over certain scene locations than the other. Small enough deviations may be small enough to be ignored. Larger, yet still small deviations may be ignored through down sampling. If deviations are significantly large, they may render certain scene regions or even entire images as unusable. Some technology exists to minimize lag between captures, but they do not completely overcome this limitation.

Computer generated

Computer generated images can fall into two categories, 3D-rendered and AI-generated. One major advantages of 3D rendering is that is its extremely flexibility. It can allow for the rendering of any 3D scene whether it exists in reality or not. With the flexibility of 3D rendering, it is easy to avoid the unnatural image distortion of the scene. Once a 3D scene is rendered, 2D images can be generated by projection at any vantage point in the scene. 3D rendering also adds the advantage of online-rending, which can be flexible to observers movements given eye tracking.

Flexibility is also the downside of 3D render/mansing, in that it is much easier to generate scenes and images that are ecologically less probable, than it is to generate ones that are natural. Detailed intricacies of scenes become increasingly more demanding and subtle with the naturality of the scene. There often is both a substantial skill and time requirement to make a scene. However, there is also the option to use free or for-profit pre-generated scenes and assets. Procedural generation can also be taken advantage of, but this is not a straightforward technique. There also exists specialized 3D-rendered datasets with particular research interests in mind (Radonjic et al., 2015).

The second major advantage of 3D scene rending is its accessibility to ground truth information. Thanks to the precision of computers and the power of rending software, ground-truth precision is likely only be limited by presentation hardware.

AI Generated

AI image generation technology is quickly finding its way into many applications. In the context of natural image generation, it is quickly becoming more viable. It is quite lucrative in its ability to generate images within seconds that appear to be fairly natural. Although some amount of skill is required, there is far less skill required than with 3D graphics. There is also the advantage that these images can be generated with intricate details that are harder to achieve using 3D graphics.

However, AI methods appear to be trained on the statistics of image content rather than scene content, thus it has a hard time extrapolating content beyond whats been included in their dataset. This type of training can also lead to the regular occurrence of artifacts and impossible scene geometry. Additional AI techniques can be used to fix artifacts or introduce additional content, but these have their limitations.

The major disadvantages of AI generation compared to computer generation is its inability to generate information about three-dimensional scene geometry and other ground truth information, or perform dynamic rendering. AI generated video is also becoming available. However, at the time of this publication, the object motion that AI video generates is currently in a fairly unnatural state.

3.3. Presenting Natural Images

By choosing the artificial approach, the experimenter opting to present stimuli in an illusory manner, usually by means of a digital display. A large aspect display presentation seems particularly contrived, given how inflexible and clearly non-illusory they are in a critical way: they do a poor job at replicating directional qualities of light fields characteristic of the scene being presented. Digital displays are typically flat surfaces set at a fixed distance away from the observer, and generate images by emitting light in a parallel light field. The resulting retinal image generated from the light field will have a accommodation, chromatic aberration, defocus blur, motion parallax cues, and binocular cues (unless using a stereoscope) suggestive of the display and not the presented scene. Additionally, displays are limited in their dynamic range (luminance and color levels), resolution,

and peripheral extent. Nevertheless, they do offer an immense amount of control over the content being displayed, which makes them particularly powerful, given their limitations. Furthermore, these limitations should be considered as discussed in Chapter 2.

Geometry relating to how the original scene was sourced and how it is presented are an additional point of consideration, especially in regards to stereo-presentation. This is discussed in Chapter 4.

3.4. Concluding remarks

Motivated by theory for a robust algorithmic-oriented psychophysical approach, this chapter discussed natural images in terms of human ecology and psychophysical experimentation.

Chapter 4 is the manuscript of an empirical psychophysical study that I, the author, designed and conducted with my advisor Johannes Burge. It is an application of the considerations discussed in this and previous chapters, and an application of the approach prescribed in Chapter 2.

CHAPTER 4

HOW DISTINCT SOURCES OF NUISANCE VARIABILITY IN NATURAL IMAGES AND SCENES LIMIT HUMAN STEREOPSIS

4.1. Abstract

Stimulus variability—a form of nuisance variability—is a primary source of perceptual uncertainty in everyday natural tasks. How do different properties of natural images and scenes contribute to this uncertainty? Using binocular disparity as a model system, we report a systematic investigation of how various forms of natural stimulus variability impact performance in a stereo-depth discrimination task. With stimuli sampled from a stereo-image database of real-world scenes having pixel-by-pixel ground-truth distance data, three human observers completed two closely related double-pass psychophysical experiments. In the two experiments, each human observer responded twice to ten thousand unique trials, in which twenty thousand unique stimuli were presented. New analytical methods reveal, from this data, the specific and nearly dissociable effects of two distinct sources of natural stimulus variability—variation in luminance-contrast patterns and variation in local-depth structure—on discrimination performance, as well as the relative importance of stimulus-driven-variability and internal-noise in determining performance limits. Between-observer analyses show that both stimulus-driven sources of uncertainty are responsible for a large proportion of total variance, have strikingly similar effects on different people, and—surprisingly—make stimulus-by-stimulus responses more predictable (not less). The consistency across observers raises the intriguing prospect that image-computable models can make reasonably accurate performance predictions in natural viewing. Overall, the findings provide a rich picture of stimulus factors that contribute to human perceptual performance in natural scenes. The approach should have broad application to other animal models and other sensory-perceptual tasks with natural or naturalistic stimuli.

4.2. Introduction

An ultimate goal for perception science is to understand and predict how perceptual systems work in the real world. One approach to achieving this goal is to probe the system with naturalistic stimuli—stimuli that are derived from the natural environment, or bear substantial similarities to them. By examining how stimulus variation characteristic of real-world scenes affects stereo-depth discrimination, we show that performance patterns are similar across different humans, and we partition the effects of distinct stimulus and scene factors on performance—with some surprising results. Further, natural-stimulus variation causes a high degree of stimulus-by-stimulus consistency across observers, consistency that, in principle, could be used to develop and constrain future image-computable models of human perceptual performance.

There is a long tradition of investigating visual performance in human and animal models using simple stimuli and simple tasks. Recent years have been marked by the realization that simple stimuli and tasks may be insufficiently complex to understand how vision works in the real world. A number of recent efforts have taken steps to make the tasks during which psychophysical and neurophysiological data are collected more ecologically valid, while using traditional stimuli (e.g. gratings, Gabors). Some such efforts have, for example, removed the requirement that animals maintain fixation, allowing them fixate freely on stimuli presented on a monitor (Yates et al., 2023). Here, we use a traditional forced-choice task, and focus effort on probing perceptual performance with stimuli that are more similar to those encountered in real-world viewing situations (see Discussion).

The use of natural or naturalistic stimuli, however, poses challenges. With such stimuli, it is difficult to maintain the rigor and interpretability that has characterized classic research. One important source of difficulty is the sheer number of factors that inject variability into natural retinal images. Some of these factors depend on the environment: the textural patterns on surfaces, the 3D structure of those surfaces, and how the objects that own those surfaces are arranged in 3D space. Other factors are due to the organism and its relationship to the environment, including the optical state of the eyes and the posture and movements of the eyes, head, and body relative to objects in the scene. All of these factors combine to generate many different retinal images, all of which

are associated with a particular value of a distal property (e.g. depth) of interest. Such natural-stimulus variability—a form of "nuisance stimulus variability"—impacts neural response (Baddeley, 1997; Felsen & Dan, 2005; A. Iyer & Burge, 2019), and is an important reason that estimation and discrimination of behaviorally-relevant latent variables (e.g. depth, size, 3D orientation) is difficult. In order to perform well, perceptual systems must select for proximal stimulus features that provide information about the latent variable of interest, while generalizing across (i.e. maintaining invariance to) stimulus variation that is not useful to the task. In natural viewing, the computations run by the vision system should minimize, to the maximum possible extent, the degree to which natural-stimulus variability causes variability in human estimation and discrimination in each critical task (Burge & Geisler, 2011, 2014; Burge, 2020; Burge & Geisler, 2012, 2015; Chin & Burge, 2020; W. S. Geisler, 1989; Sebastian et al., 2015, 2017).

Using binocular disparity as a model system, we report a systematic investigation of how various forms of natural-stimulus variability impact performance in a depth discrimination task. To approximate natural-stimulus variation, thousands of stimuli were sourced from a natural stereo-image database with co-registered laser-based range data at each pixel using constrained sampling techniques. The sampled stimuli were used to probe human depth-from-disparity discrimination and to determine distinct properties of natural scenes that place limits on human performance. With appropriate experimental designs and data analysis methods, the natural (random) variation across the uncontrolled aspects of the stimuli in each condition provides one with the ability to determine the limits that distinct types of nuisance stimulus variability place on performance.

Two experiments were conducted using the double-pass psychophysical paradigm (Burgess & Colborne, 1988; Chin & Burge, 2020; J. Gold et al., 1999; Neri & Levi, 2006). In contrast to typical 2AFC forced-designs, in which hundreds of responses are collected for each unique stimulus (or trial), double-pass experiments collect two responses for each of two presentations of hundreds of unique stimuli (or trials) in each condition. The conditions of the experiments were defined by different fixation disparities and levels of local-depth variation. These aspects of the stimuli were parametrically manipulated and tightly controlled. Other aspects of the stimuli—luminance-contrast patterns and

local-depth structure—were allowed to vary randomly (as they do in natural viewing). We develop new analytical methods that allow us to infer, from the double-pass data, i) the relative importance of natural-stimulus variability and internal noise in limiting performance, and ii) the specific impact that distinct sources of natural-stimulus variability—luminance pattern variability and local depth variability—have on performance.

Several key findings emerge. First, we replicate a performance pattern from the classic literature: discrimination thresholds increase exponentially as targets move farther in depth from fixation. Second, we show that performance limits are increasingly attributable to stimulus variability (rather than internal noise) as the stimuli used to probe performance have more local-depth variability. Third, we show that two distinct types of natural-stimulus variability—luminance-pattern variation and local-depth variation—have distinct and largely separable effects on human performance. Fourth, we find that as stimulus variation becomes more severe, the absolute impact of that stimulus-by-stimulus variation on performance becomes more severe and also becomes more uniform across human observers.

4.3. Materials and Methods

4.3.1. Human Observers

All observers had normal or corrected-to-normal acuity. Two of the observers were authors, and the third was naive to the purpose of the study. All observers provided informed written consent in accordance with the declaration of Helsinki. The Institutional Review Board at the University of Pennsylvania approved all protocols and experiments.

4.3.2. Data and Software

Psychophysical experiments were performed in MATLAB 2017a using Psychtoolbox version 3.0.12. Stimulus sampling and data-analyses were also performed in MATLAB 2017a. Data from this study is available upon reasonable request.

4.3.3. Apparatus

Stimuli were presented on a custom-built four-mirror haploscope. The haploscope displays were two identical VPixx ViewPixx 23.9 inch LED monitors. Displays were 53.3×30.0 cm in size, with 1920×1080 pixel resolution and a native 120 Hz refresh rate. The maximum luminance of each display was 106 cd/m^2. After light loss due to mirror reflections, the effective luminance was 94 cd/m^2. The mean background gray level of the displays was set to 40 cd/m^2. The gamma function was linearized over 8 bits of gray level.

All mirrors in the haploscope were front-surface mirrors, to eliminate secondary reflections. The mirrors most proximal to the observer were housed in mirror cubes with 2.5 cm circular viewports. The viewports were positioned 65 mm apart, a typical human interpupillary distance. The openings of the cubes limited the field of view to approximately 16° of visual angle.

The optical and vergence distances of the displays were set to 1.0 m. This distance was verified both by standard binocular sighting techniques and via laser distance measurement. At this distance, each pixel subtended 1.07 arcmin. A chin and forehead rest stabilized the head of each observer.

4.3.4. Stimuli

Stereo-image patches (32×32 pixels each for the left- and right-eye patches) were sampled from 98 large stereo-images (1920×1080 pixels) of the natural environment with co-registered laser range data at each pixel (Burge et al., 2016). Sampling procedures are described below. Stimuli subtended 1° of visual angle, were spatially windowed by a raised cosine function, and were presented dichoptically. When viewed monocularly, the windowing caused the stimulus to fade into the mean luminance surround. When viewed dichoptically—assuming the patch had uncrossed disparity, which it always did in these experiments—the windowing caused the stimulus to appear behind a fuzzy aperture. Uncrossed fixation disparities (i.e. uncrossed disparity pedestals) of appropriate size were introduced at the stereo-patch sampling stage by cropping the patch from its source image, assuming that a virtual pair of eyes was fixating a point along the cyclopean line of sight in front of the sampled scene location (A. V. Iyer & Burge, 2018). This created stereo-pairs that are

geometrically identical to the retinal images that would have formed on the eyes of an observer standing in the original scene. The size of the virtual fixation error was set such that the uncrossed disparity would have the desired value when the stereo-patch was viewed in the haploscope rig.

Each stereo-image patch in the dataset was labeled by the amount of local-depth variation in the imaged scene region, as quantified by disparity-contrast. Disparity-contrast is given by the root-mean-squared difference between the vergence demand of the central corresponding point and the vergence demands of the points in the local surround

$$c_\delta = \sqrt{\frac{\sum\limits_{\mathbf{x}} (\mathbf{v}(\mathbf{x}) - \mathbf{v}_0)^2 \ \mathbf{w}(\mathbf{x})}{\sum\limits_{\mathbf{x}} \mathbf{w}(\mathbf{x})}}, \tag{4.1}$$

where $\mathbf{w}$ is the raised cosine weighting function that windows the image, $\mathbf{x} = \{x, y\}$ is the spatial location of each pixel, $\mathbf{v}_0$ is the vergence angle that is required to fixate the 3D-scene point specified by the center pixels of the left- and right-eye image patches, and $\mathbf{v}(\mathbf{x})$ is the vergence angle required to fixate the scene points corresponding to the other pixels in the patch. Note that the difference in vergence demand $\mathbf{v}(\mathbf{x})$ - $\mathbf{v}_0$ is simply equal to the relative disparity between the center pixel and the other pixels in the patch. The vergence demand at each point in the patch was computed for an observer viewing the stimulus at the viewing distance and direction set by the experimental rig (i.e. 1 meter away, straight-ahead) across patches that were 32x32 pixels in size.

Each stereo-image patch was contrast fixed to the median root-mean-squared (RMS) contrast (i.e. c_{rms}=0.3) in the natural-stimulus dataset. RMS contrast is given by

$$c_{\text{rms}} = \sqrt{\frac{\sum\limits_{\mathbf{x}} \mathbf{c}^2(\mathbf{x}) \ \mathbf{w}(\mathbf{x})}{\sum\limits_{\mathbf{x}} \mathbf{w}(\mathbf{x})}}, \tag{4.2}$$

where $\mathbf{c}$ is a Weber contrast image, $\mathbf{w}$ is the raised cosine weighting function that windows the image, and $\mathbf{x} = \{x, y\}$ is the location of a given image pixel.

Stimulus sampling

Left- and right-eye image patches from a natural-scene database were sampled (i.e. centered) on corresponding points (Burge et al., 2016). Because the stereo-photographs were of natural scenes, each local patch was characterized by a different luminance pattern and by some amount of local-depth variability (see Fig. 4.1B). Corresponding points in the image were determined directly from the range data (see A. V. Iyer and Burge, 2018).

Stimuli were sampled with known amounts of fixation disparity (i.e. pedestal disparities), relative to the corresponding points, up to a maximum of 5 arcmin of uncrossed disparity and known amounts of disparity-contrast. Patches were screened to ensure that the disparity variability within the central region of each patch equaled the nominal fixation disparity within a tight tolerance (see below). Note that because depth varies naturally across any given patch, this central region was the only region of the patch that was guaranteed to equal the nominal fixation disparity. Disparity-contrasts were constrained to be either "high" (0.025-0.117 arcmin) or "low" (0.393-1.375 arcmin). To ensure that each stereo-image patch was unique, patches were not allowed to overlap radially in their source images by more than 10 pixels; this level of overlap was rare.

If the viewing geometry (i.e. distance and direction) of stimulus presentation in an experimental rig does not match the viewing geometry during stereo-image patch sampling, the stereo-specified 3D structure of presented stimulus will be distorted relative to the geometry of the original scene (Held & Banks, n.d.). Stereo-image patches were sampled from all distances and directions, but presented patches at a fixed distance and direction (i.e. one meter away, straight-ahead). Hence, the stereo-specified depth structure during presentation was distorted from that in the original 3D scene. It is possible to prevent these distortions, but only at the cost of distorting the left- and right-eye luminance images. We opted to preserve luminance structure rather than the details of the stereo-specified 3D geometry of the original natural scene. Throughout the article, the disparity-contrast values that are used to characterize the stereo-specified depth variation in each stereo-image patch were set by each patch as it was viewed by the participants in the experimental rig.

Stimulus vetting

Before being included in the experimental stimulus set, stereo-image patches underwent a vetting procedure. The vetting procedure had two primary aims.

The first, most fundamental aim was to ensure accurate co-registration between the luminance and range information in the half-images of each patch. Accurate co-registration was critical for all aspects of the experiment, because the values of the independent variables (i.e. disparity and disparity-contrast) are determined directly from the range data. Although inaccurate co-registration was rare, it was present in a non-negligible proportion of patches. In such cases, the luminance data that observers would have used to estimate disparity would have been inconsistent with the range data used by the experimenters to compute the nominal ground truth disparity. Hence, failing to identify and exclude poorly co-registered patches would mar the accuracy of the results. Potential stereo-image patches were manually vetted by viewing each patch in the experimental rig with onscreen disparities that were nominally uncrossed, zero, or crossed with respect to the screen. Patches that did not pass scrutiny (i.e. that had the wrong depth relationship relative to the screen) were discarded from the pool. The manual vetting procedure was conducted until thousands of unique stimulus patches without co-registration problems were obtained.

The second aim of the vetting procedure—which was enforced programmatically—was to ensure that the center of each stereo-image patch was a coherent target for depth estimation (see above). We required that the most central ($1/8^\circ$ of visual angle $\approx$ 4x4 pixel) region of each patch contained neither a substantial change in disparity (i.e. a disparity-contrast greater than 20 arcsec), or a half-occluded region. Pixels containing half-occluded regions were allowed outside of the most central region. Because regions that are half-occluded have undefined disparity, stimuli including a half-occluded region have undefined disparity-contrast. For patches containing half-occlusions, disparity-contrast was computed by excluding pixels corresponding to half-occluded regions of the scene from the calculation. We did not exclude stimuli with half-occlusions from the dataset because they occur commonly in natural viewing (A. V. Iyer & Burge, 2018).

Stimulus flattening

From the sampled set of natural stereo-image patches—which contain both natural- luminance-pattern variation and natural-depth variation—we also created a "flattened"—but otherwise matched—dataset of stereo-image patches. To convert patches with natural-depth structure into patches with flat depth structure, either the left- or right-eye half-image patch (chosen by random) was replaced by a duplicate of the remaining right- or left-eye half-image patch. This procedure ensured that there is essentially zero-disparity variation across the patch, such that the disparity pattern specifies a fronto-parallel plane.

4.3.5. Procedure

Stimuli were presented at the center of a fixation cross-hairs reticle. The reticle was positioned in the center of a circular, 4° diameter, mean-luminance gray area. The circular area was surrounded by a mean-luminance 1/f noise field. The reticle itself consisted of a 2° diameter circle punctuated by hairs jutting outwards at the cardinal and ordinal directions. Hairs were 1° in length and 4.2 arcmin in thickness.

Stimuli were presented using a two-interval forced choice (2IFC) procedure. Each interval had a duration of 250 ms. The inter-stimulus interval was also 250 ms. In one interval of each trial, a stimulus with a standard disparity was presented. In the other interval, a stimulus with a comparison disparity was presented. The order in which the standard or comparison stimulus was presented was randomized.

The task was to report, with a key press, whether the stimulus in the second interval appeared to be nearer or farther than the stimulus in the first interval. Feedback was provided after each response: a high frequency tone indicated a correct response; a low frequency tone indicated an incorrect response.

Psychometric data was collected in a fully-crossed design with disparity pedestal and disparity-contrast as the independent variables. For each combination of disparity pedestal and disparity-contrast, the method of constant stimuli was used for stimulus presentation. Disparity pedestals

were defined by one of five standard disparities: $\delta_{std} = [-11.25, -9.38, -7.5, -5.63, -3.75]$ arcmin. Five equally spaced comparison disparities were paired with each standard. Disparity-contrast levels were defined as $\delta_C = [0.025—0.117, 0.393—1.375]$ arcmin, which were labeled "low" and "high" disparity-contrasts respectively. Stimuli in the low disparity-contrast conditions were just-noticeably non-flat to observers. Stimuli in the high disparity-contrast conditions appeared quite noticeably non-uniform in depth. The high disparity-contrast condition contained stimuli that were easily fusible in most cases.

The comparison disparity pedestals and disparity-contrast levels were chosen based on pilot data. Comparison disparities were chosen with the aim that proportion comparison chosen would be approximately 10% on the low-end and 90% on the high-end across the low disparity-contrast condition. Data with 0% and 100% comparison chosen provides no useful information for estimating decision-variable correlation (see subsection "Partitioning the variability of the decision variable" below). Before collecting the data, each observer completed practice sessions to ensure that discrimination performance was stable.

To simulate the stimulus variability that occurs in natural-viewing conditions, a unique natural stereo-image patch was presented on each interval of each trial. This feature of the experimental design represents a departure from more standard experimental designs, in which either the same stimulus is presented many times each or stimulus differences (e.g. different random dot stereograms) are considered unimportant and not analyzed.

Experiments were conducted using a double-pass experimental paradigm. In double-pass experiments, observers respond to the exact same set of unique trials two times each. Double-pass experiments enable one to determine the relative importance of factors that are repeatable across trials (e.g. external stimulus variation), and factors that vary randomly across trials (e.g. internal noise).

Two double-pass experiments were conducted. In one, all stimuli had natural-depth variation. In the other, all stimuli were "flattened" (see sub-subsection "Stimulus flattening" above). Im-

portantly, both double-pass experiments used the same scene-locations (and hence, near-identical luminance contrast patterns). This design feature allowed us to examine the relative importance of luminance-pattern-driven variability and disparity-contrast-driven variability in the decision variable (see subsection "Partitioning the externally-driven component of the decision variable" below).

Over the course of each double-pass experiment, 10,000 unique stimuli were presented in 5000 unique trials of each double-pass experiment. Five hundred trials were collected in each of ten conditions (5 standard disparities $\times 5$ comparison disparities $\times 2$ disparity-contrasts). Data was collected in 100-trial blocks (i.e. twenty repeats per comparison disparity level per block). The order in which the blocks were run was randomized and counterbalanced across conditions. Two double-pass experiments were conducted, for a total of 20,000 trials per observer.

4.3.6. Psychometric fitting

Cumulative Gaussian functions were fit to the psychometric data in each condition using maximum likelihood methods. Discrimination thresholds were calculated from the fitted functions. The relationship between the sensitivity index d' (i.e. d-prime) and percent the comparison chosen PC in a two-interval two-alternative forced-choice experiment is given by

$$\mathrm{PC} = \Phi\left(\frac{d'}{\sqrt{2}}\right), \tag{4.3}$$

where Φ is the cumulative normal function, with d' given by

$$d' = \frac{\Delta\delta}{\sigma_T^2}, \tag{4.4}$$

where $\Delta\delta = \delta_{\mathrm{cmp}} - \delta_{\mathrm{std}}$ is the difference between the comparison and standard disparities (i.e. the mean value of the decision variable), and σ_T^2 is the variance of the underlying decision variable. (In accordance with standard practices, we assume that decision variable variance is constant for all comparison-disparity levels at a given standard-disparity level—that is, pedestal disparity. The psychometric data is consistent with this assumption.)

The discrimination threshold T is set by choosing a criterion d-prime that defines the just-noticeable

difference. In a two-interval, two-alternative forced-choice (2AFC) experiment, threshold is given by

$$T = \sqrt{\sigma_T^2 \; d'_{\text{crit}}}, \tag{4.5}$$

where d'_{crit} is the criterion d-prime. For computational simplicity, we assume a criterion d-prime of 1.0 such that threshold level performance corresponds to the 76% point on the psychometric function. Thresholds are thus given by the change in the disparity required to go from the 50% to the 76% points on the psychometric function.

Discrimination thresholds were computed from data across both passes of the experiment. When fitting psychometric data across one or both double-pass experiments (see below), thresholds were constrained to change log-linearly across disparity pedestals. Under this constraint, discrimination thresholds in the conditions of a double-pass experiment associated with a given disparity-contrast are specified by

$$T = \sigma_T = \exp(m\delta_{std} + b), \tag{4.6}$$

where δ_{std} is the standard pedestal disparity, m and b are the slope and y-intercept of the line characterizing the log-thresholds. This constraint is consistent the predictions of normative models of disparity discrimination with natural stimuli, previously reported patterns in psychophysical data (Blakemore, 1970), and the log-linear patterns in the current threshold data (see Figs. 4.4 and 4.7). The maximum-likelihood estimates of the parameters defining threshold under the constraint were fit across all conditions having a given disparity-contrast. They are given by

$$\hat{m}, \hat{b} = \arg\max_{m,\; b} \sum_s L_s([\exp(m\delta_{\text{std}}^{(s)} + b)]^2), \tag{4.7}$$

where L_s is the likelihood of the raw response data in the s^{th} condition, under the assumption that percent correct is governed by a cumulative normal function with mean parameter equal to the s^{th} disparity pedestal $\delta_{\text{std}}^{(s)}$ and variance parameter equal to $[\exp(m\delta_{\text{std}}^{(s)} + b)]^2$. Finally, the variance of

decision variable at each pedestal disparity was obtained by plugging these estimated parameters into Section 4.3.6.

4.3.7. Modeling the decision variable

The decision variable can be modeled as a difference between disparity estimates from the stimuli presented on each trial

$$D = \hat{\delta}_{cmp} - \hat{\delta}_{std}, \tag{4.8}$$

where $\hat{\delta}_{std}$ is the estimate from the stimulus with the standard disparity and $\hat{\delta}_{cmp}$ is the estimate from the stimulus with the comparison stimulus. In accordance with signal detection theory, if the value of the decision variable is greater than zero (and if the observer sets the criterion at zero), the observer will select the stimulus with the comparison disparity. If the decision variable is less than zero, the observer will select the stimulus with the standard disparity.

The decision variable can be more granularly modeled as the sum of two independent random variables. The first random variable accounts for stimulus-driven variability (i.e. variance that is due to nuisance stimulus variability), and has its value set by the particular stimulus (or stimuli) that are presented on a given trial. The second random variable accounts for internal noise, and has its value set randomly on each trial. In a double-pass experiment, across the two presentations of a particular unique trial in a double-pass experiment (i.e. the presentation in the first pass and the presentation in the second pass), the value of the decision variables are given by

$$\begin{aligned} D_1 &= V + W_1, \\ D_2 &= V + W_2, \end{aligned} \tag{4.9}$$

where V is stimulus-driven contribution to the decision variable, W is a sample of internal noise, and the subscripts index on which pass the trial was presented. Across the two passes of the double-pass experiment, the decision variables can be described as a single two-dimensional random variable $\mathbf{D} = [D_1, D_2]^\intercal$.

The stimulus-driven component of the decision variable on a single pass of the experiment $V \sim \mathcal{N}(\delta_{cmp} - \delta_{std},\ \sigma_E^2)$ is modeled as unbiased and normally distributed with stimulus-driven variance σ_E^2. The noise-driven component of the decision variable $W \sim \mathcal{N}(0, \sigma_I^2)$ is modeled as zero-mean and normally distributed with variance σ_I^2. If the external (i.e. stimulus-driven) and internal (i.e. noise-driven) components of the decision variable are independent, as we assume they are here, the total variance of the decision variable on a given pass is given by the sum of the internal and external components

$$\sigma_T^2 = \sigma_E^2 + \sigma_I^2. \tag{4.10}$$

4.3.8. Decision-variable correlation

The correlation of the decision variable across passes is given by the fraction of the total variance that is accounted for by external (i.e. stimulus-driven) factors, the factors that are repeated across passes. Hence, decision-variable correlation is given by

$$\rho = \frac{\sigma_E^2}{\sigma_T^2} = \frac{\sigma_E^2}{\sigma_E^2 + \sigma_I^2}, \tag{4.11}$$

where σ_E^2 is the component of the decision-variable variance accounted for by external (i.e. stimulus-driven) factors and σ_I^2 is the component of the decision-variable variance accounted for by internal factors (i.e. noise). In order to partition stimulus- and internally-driven sources of variability, we combine estimates of decision-variable correlation and discrimination thresholds (see below). Decision-variable correlation is an integral factor in determining the repeatability of observer responses across passes of a double-pass experiment.

Estimating decision-variable correlation

Decision-variable correlation was estimated via maximum likelihood from the pattern of observer response agreement between passes. The log-likelihood of n-pass response data, under the model of the decision variable, is

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_j N_j \ \log \mathcal{L}_n^j(\boldsymbol{\theta}), \tag{4.12}$$

where $\boldsymbol{\theta}$ represents the parameter(s) to be estimated, j is a specific pattern of response, N_j represents the number of times a specific pattern of response was measured. For a double-pass experiment ($n = 2$), the set of response patterns are given by the combination of all possible combinations of responses for each pass. The number of patterns of binary responses is $N = 2^n$. For 2IFC experiment, $N = 2^2 = 4$, with patterns of responses $j \in \{[-,-],\ [-,+],\ [+,-],\ [+,+]\}$. Here, we use $+$ to indicate that the comparison was chosen and $-$ indicates the comparison was not chosen.

We model the joint decision variable as a vector drawn from a multivariate normal distribution $\mathbf{D} \sim \mathcal{N}(\mathbf{x}; \mathbf{m}, \Sigma)$ with a mean vector $\mathbf{m}$ and covariance matrix Σ. The likelihood of a particular pattern of response is given by

$$\mathcal{L}_2^j(\boldsymbol{\theta}) = \int_{s^j(c_1,c_2)} \mathcal{N}(\mathbf{x}; \mathbf{m}, \Sigma)\ d\mu(\mathbf{x}), \tag{4.13}$$

where integration is in respect to probability measure μ and s^j is a subset of the support S. Here, s^j defines the integration limits for a specific pattern of response j and is a function of the decision criterion on each pass $c \in \{c_1, c_2\}$. Specifically, the integration limits for each dimension/pass are determined by the values of response pattern. For a response r_i at pass i, where the comparison is not chosen ($r_i = -$), $\mathrm{P}(D_i < c_i)$ and the integration limits are $[-\infty,\ c_i]$. Likewise, for comparison chosen ($r_i = +$), $\mathrm{P}(D_i \geq c_i)$ with integration limits $[c_i, \infty]$.

It is computationally convenient to estimate decision-variable correlation with a normalized joint decision variable $\mathbf{D}^z = [D_1^z, D_2^z]^\intercal$ such that it has unit variance on each pass. Normalizing the joint decision variable sets the normalized means equal to d'. Normalizing the joint decision variable also confers a practical advantage in converting the covariance matrix into a correlation matrix so that it can be fully characterized by decision-variable correlation.

The normalized mean vector and normalized covariance (i.e. correlation) matrix associated with the normalized joint decision variable are given by $\mathbf{m}^z = \mathrm{M}\mathbf{m}$, and $\Sigma^z = \mathrm{M}\Sigma\mathrm{M}$, where the superscript z indicates a normalized parameter, and Σ^z is the correlation matrix (i.e. the covariance matrix of the normalized joint decision variable). The normalizing matrix is given by $\mathrm{M} = \mathrm{diag}(\frac{1}{\boldsymbol{\sigma}_T})$, where

$\boldsymbol{\sigma}_T$ is a vector of the standard deviation of the joint decision variable $\mathbf{D}$ in each pass, and where the diag(.) function converts a vector into a matrix with the vector-values on the diagonal. The correlation matrix is given by

$$\Sigma^z = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{4.14}$$

Substituting parameters associated with the normalized decision variable into equations, yields mathematically equivalent expressions of the likelihoods:

$$\mathcal{L}_2^j(\boldsymbol{\theta}) = \int\limits_{s^j(c_1^z, c_2^z)} \mathcal{N}\left(\mathbf{x}^z; \mathbf{m}^z, \Sigma^z\right)\ d\mu(\mathbf{x}^z). \tag{4.15}$$

We also assume that the criteria associated with the normalized decision variable on all passes equals zero, which is justified by the data and by the two-interval, two-alternative forced choice design. In the general case, when this assumption is not made, the decision criteria should also be normalized—that is, the normalized criteria are given by $\mathbf{c}^z = \mathrm{M}\mathbf{c}$. Thus, when analyzing double-pass experimental data under the indicated assumptions, decision variable correlation $\boldsymbol{\theta} = \rho$ is the only parameter that needs to be estimated. Specifically, the maximum-likelihood estimate of decision variable correlation is given by

$$\hat{\rho} = \arg\max_{\rho} \sum_j N_j\ \log \mathcal{L}_2^j(\rho), \tag{4.16}$$

where $N_p \in \{N_{--},\ N_{-+},\ N_{+-},\ N_{++}\}$ is the number of each type of response agreement or disagreement, and $\mathcal{L}_p \in \{\mathcal{L}_{--},\ \mathcal{L}_{-+},\ \mathcal{L}_{+-},\ \mathcal{L}_{++}\}$ is the likelihood of the data given an underlying decision variable distribution specified by the decision variable correlation. The likelihoods are given

by

$$\begin{aligned}
\mathcal{L}_2^{--} &= \int_{-\infty}^{c_1^z}\int_{-\infty}^{c_2^z} \mathcal{N}\left(\mathbf{x}^z;\mathbf{m}^z,\mathbf{\Sigma}^z\right)\ d\mu(x_1,x_2),\\
\mathcal{L}_2^{-+} &= \int_{-\infty}^{c_1^z}\int_{c_2^z}^{\infty} \mathcal{N}\left(\mathbf{x}^z;\mathbf{m}^z,\mathbf{\Sigma}^z\right)\ d\mu(x_1,x_2),\\
\mathcal{L}_2^{+-} &= \int_{c_1^z}^{\infty}\int_{-\infty}^{c_2^z} \mathcal{N}\left(\mathbf{x}^z;\mathbf{m}^z,\mathbf{\Sigma}^z\right)\ d\mu(x_1,x_2),\\
\mathcal{L}_2^{++} &= \int_{c_1^z}^{\infty}\int_{c_2^z}^{\infty} \mathcal{N}\left(\mathbf{x}^z;\mathbf{m}^z,\mathbf{\Sigma}^z\right)\ d\mu(x_2,x_2).
\end{aligned} \tag{4.17}$$

Determining the variances of the decision-variable components

With an estimate of the total variance of the decision variable and an estimate of decision-variable correlation, one can estimate the variances of the externally- and internally-driven components of the decision variable. Plugging Eq. (4.5) into Eq. (4.11) and rearranging yields an estimate of the variance the externally-driven component of the decision variable

$$\hat{\sigma}_E^2 = \hat{\rho}\hat{\sigma}_T^2. \tag{4.18}$$

Plugging this estimate into Eq. (4.10) and rearranging gives an expression for the internally-driven component of the decision variable

$$\hat{\sigma}_I^2 = \hat{\sigma}_T^2 - \hat{\sigma}_E^2. \tag{4.19}$$

This series of analytical steps was performed for the two double-pass experiments that were conducted: one with natural and one with flattened local-depth variation.

4.3.9. Partitioning the externally-driven component of the decision variable

To estimate the contributions of luminance-pattern- and local-depth-driven (i.e. disparity-contrast-driven) variability to the decision variable, performance was compared across the stimulus sets with natural and flattened local-depth variation. Recall that the flattened stimulus set effectively eliminates local-depth-variability from the natural-stimulus set—because the disparity pattern in each flattened stimulus specifies a fronto-parallel plane—while leaving luminance contrast patterns essentially unaffected. Hence, because the luminance-pattern-driven component should be essentially the same in both stimulus sets, and because the local-depth-driven component is eliminated in one of the two stimulus sets, an appropriate comparison should reveal the impact of each factor.

To compare performance across the flattened and natural-stimulus sets, we simultaneously analyzed all data from both double-pass experiments using a quasi-quadruple-pass analysis (see below).

Expanded decision variables and correlations

Before explaining in detail how to estimate the contribution of two distinct stimulus-driven factors it is necessary to show how the decision variable depends on these factors in each of the two double-pass experiments. The decision variables in the experiments with flattened and natural-stimuli are given, respectively, by

$$V_{\dagger} = L, \tag{4.20}$$

$$V_{*} = L + B, \tag{4.21}$$

where L and B denote the the luminance-pattern- and local-depth-driven components of the decision variable, respectively, and $\dagger$ and $*$ indicate, respectively, whether the decision variable corresponds to stimuli that have been flattened (2nd double-pass experiment) or have natural-depth profiles (1st double-pass experiment). (Note that, for the simplicity of mathematical development, we present the equations here in the Methods section in the opposite order from which the experiments were conducted and presented in the Results section).

Plugging these expanded forms for the externally-driven component of the decision variable into

Eq. (4.9) yields expanded expressions for the decision variables in each of the two double-pass experiments

$$\begin{aligned} D_\dagger &= \overbrace{(L \qquad)}^{V_\dagger} + W_\dagger, \\ D_* &= \underbrace{(L+B)}_{V_*} + W_*. \end{aligned} \tag{4.22}$$

Clearly, the presence or absence of the local-depth-driven component of the decision variable was the only component that differed across the two double-pass experiments.

Decision-variable correlations across passes in the flattened and natural double-pass experiments, in terms of these new variables, are given by

$$\begin{aligned} \rho_{\dagger\dagger} &= \frac{\sigma^2_{E\dagger}}{\sigma^2_{T\dagger}} = \frac{\sigma^2_L}{\sigma^2_L + \sigma^2_{I\dagger}}, \\ \rho_{**} &= \frac{\sigma^2_{E_*}}{\sigma^2_{T_*}} = \frac{\sigma^2_L + \sigma^2_B + 2\text{cov}[L,B]}{\sigma^2_L + \sigma^2_B + 2\text{cov}[L,B] + \sigma^2_{I_*}}, \end{aligned} \tag{4.23}$$

where $\sigma^2_{T\dagger}$ and $\sigma^2_{T_*}$ are variabilities of the decision variable, where σ^2_L and σ^2_B are the luminance-pattern and local-depth-driven contributions to response variability, $\sigma^2_{I_\dagger}$ is the internal noise when only luminance-pattern-driven variability is present, $\sigma^2_{I_*}$ is the internal noise when both luminance-pattern- and local-depth-driven variability is present, $\dagger\dagger$ indicates comparisons across between passes in the double-pass experiment with flattened-depth profiles (i.e. the second double-pass experiment), and $**$ indicates comparisons across passes in the double-pass experiment with natural-depth profiles (i.e. the first double-pass experiment). Clearly, there are five unknowns—σ^2_L, σ^2_B, $\text{cov}[L,B]$, $\sigma^2_{I_\dagger}$, and $\sigma^2_{I_*}$—and, including the threshold equations from each of the two double-pass experiments (see Eqs. (4.5) and (4.10)), only four equations. However, by computing decision-variable correlation between passes across each of the two double-pass experiments, a fifth equation is obtained. Specifically,

$$\rho_{\dagger *} = \frac{\sigma^2_L + \text{cov}[L,B]}{\sigma_{T_\dagger}\sigma_{T_*}} = \frac{\sigma^2_L + \text{cov}[L,B]}{\sqrt{(\sigma^2_L + \sigma^2_{I_\dagger})}\sqrt{(\sigma^2_L + \sigma^2_B + 2\text{cov}[L,B] + \sigma^2_{I_*})}}, \tag{4.24}$$

where †∗ indicates the cross-double-pass-experiment comparisons. Now, with five equations and five unknowns, the equations can be solved.

Estimating decision-variable correlation with expanded decision variables

A novel quasi-quadruple-pass analysis was used to simultaneously estimate $\rho_{\dagger\dagger}$, ρ_{**}, and $\rho_{\dagger *}$, the decision-variable correlations across all four passes of the two double-pass experiments. The quasi-quadruple pass analysis is distinguished from an "ordinary" quadruple-pass analysis because, in an ordinary analysis, all four passes present identical trials. Here, only some of the four passes present trials with identical stimuli (e.g. the flattened stimuli were similar but not identical to the stimuli with natural depth variation). The quasi-quadruple pass analysis allows the three distinct decision variable correlations to take on different values. An ordinary analysis does not allow this flexibility.

For a quadruple-pass (whether quasi or not), the likelihood function is obtained takes the form $\mathcal{L}_n(\boldsymbol{\theta}) = \sum_j N_j \ \log \mathcal{L}_n^j(\boldsymbol{\theta})$ from Eq. (4.12) above, but across sixteen response patterns

$$j \in \left\{ \begin{array}{lllllll} [++++], & & & & & & \\ [+++-], & [++-+], & [+-++], & [-+++], & & & \\ [++--], & [+-+-], & [+--+], & [-++-], & [-+-+], & [--++], & \\ [---+], & [--+-], & [-+--], & [+---], & & & \\ [----] & & & & & & \end{array} \right\}.$$

The individual likelihoods for these response patterns are extended from Eq. (4.15), such that

$$\mathcal{L}_4^j(\boldsymbol{\theta}) = \int\limits_{s^j(c_1^z, c_2^z, c_3^z, c_4^z)} \mathcal{N}\left(\mathbf{x}^z; \mathbf{m}^z, \Sigma^z\right) \ d\mu(\mathbf{x}^z), \tag{4.25}$$

with integration limits s^j as described in the text proceeding Eq. (4.13). Here, an example likelihood is the region in the space in which the decision variable is positive on all four passes, given by

$$\mathcal{L}_4^{++++}(\boldsymbol{\theta}) = \int\limits_{c_1^z}^{\infty} \int\limits_{c_2^z}^{\infty} \int\limits_{c_3^z}^{\infty} \int\limits_{c_4^z}^{\infty} \mathcal{N}\left(\mathbf{x}^z; \mathbf{m}^z, \boldsymbol{\Sigma}^z\right) \ d\mu(x_1^z, x_2^z, x_3^z, x_4^z). \tag{4.26}$$

Just as with the double-pass analysis described above, it is convenient to normalize the joint decision variable $\mathbf{D}$ in quadruple-pass analyses via application of a normalization matrix $\mathrm{M} = \mathrm{diag}(\frac{1}{\boldsymbol{\sigma}_T})$. In a quasi-quadruple-pass analysis, the vector $\boldsymbol{\sigma}_T$ of standard deviations is given by

$$\boldsymbol{\sigma}_T = \begin{bmatrix} \sigma_{T\dagger} \\ \sqrt{\sigma_{T\dagger}\sigma_{T*}} \\ \sqrt{\sigma_{T*}\sigma_{T\dagger}} \\ \sigma_{T*} \end{bmatrix}, \tag{4.27}$$

with resulting in correlation matrix

$$\Sigma^z = \begin{bmatrix} 1 & \rho_{\dagger\dagger} & \rho_{\dagger *} & \rho_{\dagger *} \\ \rho_{\dagger\dagger} & 1 & \rho_{\dagger *} & \rho_{\dagger *} \\ \rho_{\dagger *} & \rho_{\dagger *} & 1 & \rho_{**} \\ \rho_{\dagger *} & \rho_{\dagger *} & \rho & 1 \end{bmatrix}. \tag{4.28}$$

With estimates i) of the total variance of the decision variables from the two double-pass experiments (i.e. $\sigma^2_{T_*}$ and $\sigma^2_{T_\dagger}$) which are obtained from the thresholds, and ii) of the three decision-variable correlations between passes within and across the two double-pass experiments (i.e. $\rho_{\dagger\dagger}$, ρ_{**}, and $\rho_{\dagger *}$), the values of the five unknown parameters can be determined.

Estimates of decision-variable correlation are obtained by maximizing the likelihood of the data under the model. Specifically,

$$\hat{\rho}_{\dagger\dagger},\ \hat{\rho}_{\dagger *},\ \hat{\rho}\ = \underset{\rho_{\dagger\dagger},\ \rho_{\dagger *},\ \rho}{\arg\max} \sum_j N_j\ \log \mathcal{L}_4^j(\rho_{\dagger\dagger},\ \rho_{\dagger *},\ \rho\ ). \tag{4.29}$$

We show in the next section how to solve for the contributions of the two distinct natural stimulus-driven factors—i.e. L and B—to the variance of the decision variable.

Determining the variability of the stimulus-driven components

We modeled natural-stimulus variability as being due to two distinct factors: luminance-pattern variability L and local-depth-variability B. To obtain maximum-likelihood estimates of the variance of the luminance-pattern-driven component of the decision variable $\hat{\sigma}_L^2$, the variance of the local-depthnzz-driven component $\hat{\sigma}_B^2$, and the covariance between these two components $\hat{\text{cov}}[L, B]$, from the maximum-likelihood estimates of the three decision-variable correlations (see Eq. (4.29)), we rearranged Eqs. (4.23) and (4.24) for the variables in question. Specifically,

$$\hat{\sigma}_L^2 = \hat{\rho}_{\dagger\dagger}\hat{\sigma}_{T_\dagger}^2, \tag{4.30}$$

$$\hat{\text{cov}}[L, B] = \hat{\rho}_{\dagger *}\hat{\sigma}_{T_\dagger}\hat{\sigma}_{T_*} - \hat{\sigma}_L^2, \tag{4.31}$$

$$\hat{\sigma}_B^2 = \hat{\rho}_{**}\hat{\sigma}_{T_*}^2 - \hat{\sigma}_L^2 - 2\hat{\text{cov}}[L, B]. \tag{4.32}$$

The maximum likelihood estimates indicated in Eqs. (4.30) and (4.32) are plotted in the main text Figure 4.10. The maximum-likelihood estimate of the covariance between the two components (4.31) tended towards zero, and can safely be ignored.

Fitting constraints

Model parameters were fit via the quasi-quadruple-pass analysis under a pair of constraints. The first constraint was that the disparity-discrimination thresholds used in normalization matrix M (see Eq. (4.27)) were set to values obtained from linearly constrained threshold fits (see Eq. (4.7)). The second constraint was that the scaled covariance (i.e. correlation) between the luminance-driven and local-depth-driven components of the decision variable was constrained to lie between -1 and

1. In particular,

$$-1 < \frac{\mathrm{cov}[L, B]}{\sigma_L \sigma_B} < 1, \tag{4.33}$$

where the scaling factor is given by the product the standard deviations of the two stimulus-driven components. Given that most estimates of interaction term were near zero, we re-fit the model with the more stringent constraint that this interaction term equaled zero. Eqs. (4.23) and (4.24) make clear that setting the interaction term equal to zero forces the different decision-variable correlations to have more constrained values with respect to one another than they would be constrained to have otherwise. The log-likelihoods of the models with their best-fit parameters were essentially identical, regardless of whether the interaction term was set equal to zero or not. Non-zero values of the interaction term thus carried little explanatory value.

4.3.10. Between-observers decision-variable correlation

To derive an expression for between-observers decision-variable correlation, the stimulus-driven component of the decision variable is assumed to be the sum of two independent random variables. (Note that this expansion of the stimulus-driven component is not inconsistent the expansion used in Eq. (4.21) above.) One is a stimulus-driven component that is shared across observers; the other is a stimulus-driven component that is private to each observer. Specifically,

$$\begin{aligned} D_1 &= \overbrace{(S_1 + P_1)}^{V_1} + W_1, \\ D_2 &= \underbrace{(S_2 + P_2)}_{V_2} + W_2, \end{aligned} \tag{4.34}$$

where S_1 and S_2 are stimulus-driven components that are identically driven by the stimulus across observers (i.e. the components are proportional $S_1 \propto S_2$, or identical up to a scale factor $S_1 = aS_2$), P_1 and P_2 are the stimulus-driven components that are private to (i.e. uncorrelated between) each observer, and W_1 and W_2 are the respective noise-driven components (see Eq. (4.9)). The total variance of the stimulus-driven component of the decision variable V_i in each subject i is given by $\sigma^2_{Ei} = \sigma^2_{Si} + \sigma^2_{Pi}$, the sum of the variances in the shared and private components. (Note the

overloaded subscript notation. Here, subscripts to denote different subjects. Earlier, subscripts to denoted different passes through the experiment.) Between-subjects decision-variable correlation is given by

$$\rho_{12} = \frac{\text{cov}[S_1, S_2]}{\sqrt{\sigma_{T1}^2 \sigma_{T2}^2}}, \tag{4.35}$$

where σ_{T1}^2 and σ_{T2}^2 are the total variance of the decision variables in each observer. Of course, these variances include the effects of internal noise. To eliminate the impact of internal noise in the two observers, one can divide through by the square-roots of the within-observer decision-variable correlations to obtain the partial correlation

$$\rho_{12 \cdot W} = \frac{\rho_{12}}{\sqrt{\rho_{11}\rho_{22}}} = \frac{\text{cov}[S_1, S_2]}{\sqrt{\sigma_{E1}^2 \sigma_{E2}^2}}, \tag{4.36}$$

where σ_{E1}^2 and σ_{E2}^2 are the variances of the stimulus-driven component of the decision variable for each observer, and ρ_{11} and ρ_{22} are the within-observer decision-variable correlations for each observer. This partial correlation $\rho_{12 \cdot W}$ specifies the degree to which the stimulus-driven components in two different observers are correlated with each other. High levels of this partial correlation indicate that stimulus-driven components of the two observer are highly similar.

Estimating between-observers correlations

Estimation of between-observers decision-variable correlation within a given experiment also utilized the quasi-quadruple pass methodology, with a few small but important differences. The vector of standard deviations that determined the normalizing matrix is given by

$$\boldsymbol{\sigma}_T = \begin{bmatrix} \sigma_{T_1} \\ \sqrt{\sigma_{T_1}\sigma_{T_2}} \\ \sqrt{\sigma_{T_2}\sigma_{T_1}} \\ \sigma_{T_2} \end{bmatrix}, \tag{4.37}$$

where subscripts 1 and 2 indicate observer identity, rather than the experiment number. The resulting correlation matrix is given by

$$\mathbf{\Sigma^z} = \begin{bmatrix} 1 & \rho_{11} & \rho_{12} & \rho_{12} \\ \rho_{11} & 1 & \rho_{12} & \rho_{12} \\ \rho_{12} & \rho_{12} & 1 & \rho_{22} \\ \rho_{12} & \rho_{12} & \rho_{22} & 1 \end{bmatrix}. \tag{4.38}$$

where ρ_{12} is the between-observer decision variable correlation, and ρ_{11} and ρ_{22} are the within-observer decision variable correlations. The maximum likelihood estimates of these parameters was given by

$$\hat{\rho}_{11},\ \hat{\rho}_{12},\ \hat{\rho}_{22} = \underset{\rho_{11},\ \rho_{12},\ \rho_{22}}{\arg\max} \sum_j N_j\ \log \mathcal{L}_4^j(\rho_{11},\ \rho_{12},\ \rho_{21}). \tag{4.39}$$

In each of the two experiments, all three unique pairings of observers per experiment were analyzed, so that three between-observers decision variable correlations were estimated for each condition of each experiment.

Between-observer fitting constraints

Constraints for quasi-quadruple between-observers analysis were similar to those for with-observer analysis. First, disparity-discrimination thresholds used in normalization matrix M (see Eq. (4.37)) were set to values obtained from linearly constrained threshold fits (see Eq. (4.7)). Second, the scaled partial correlation $\rho_{12 \cdot W}$ between the luminance-driven and local-depth-driven components of the decision variable was constrained to lie between -1 and 1. In particular,

$$-1 < \frac{\text{cov[S1,S2]}}{\sigma_{E1}\sigma_{E2}} < 1. \tag{4.40}$$

4.3.11. Spatial integration

We examined whether the decision variable correlations that we observed might be due, at least in part, to observers basing their responses on the disparity averaged over a fixed window size

rather than the disparity at the central pixel. We computed alternative decision variables for different spatial integration areas and tested whether they provide improved ability to account for the observer responses and decision variable correlations.

Throughout the article, we defined the disparity of the patch to be the disparity associated with the central pixel. But there is no guarantee that human observers base their responses upon the disparity of the central pixel alone. It is possible–perhaps, likely–that observers based their responses on the average disparity within some spatial integration region. On this alternative hypothesis about how the task was performed, we computed alternative decision variables as follows

$$D_{\mathrm{a}} = \frac{\sum\limits_{\mathbf{x}} \left(\delta_{\mathrm{cmp}}(\mathbf{x}) - \delta_{\mathrm{std}}(\mathbf{x})\right)\, \mathbf{w}_i(\mathbf{x})}{\sum\limits_{\mathbf{x}} \mathbf{w}_i(\mathbf{x})}, \tag{4.41}$$

where the window $\mathbf{w}$ defines the area of spatial integration. We computed alternative decision variables for pillbox-shaped windows having diameters from a four pixel diameter up to a 32 pixel diameter. For any given alternative decision variable, the binary response predicted by the alternative decision variable value is given directly by its sign. The ability of the alternative decision variable to predict the human responses was then assessed via a logistic regression model.

To setup the logistic regression model, the real-valued decision variables were used as the regressor for the human binary responses. For each window-size, a single random effects model was used, conditioned by disparity pedestal and disparity contrast conditions and observer. The coefficient of determination (R^2) was used to asses explanatory power of a given window size, and the Akaike information criterion (AIC) was used to compare models and their significance. There is no evidence that the human data can be better accounted for by a spatial integration area larger than that implicitly assumed throughout the main analyses in the article.

4.4. Results

Three observers collected 20,000 trials each across two double-pass experiments, with the aim of determining how different types of natural stimulus variability—namely, variation in luminance-contrast patterns and variation in local-depth variation—limit sensory-perceptual performance in a depth-from-disparity discrimination task. Comparing performance between two appropriately designed double-pass experiments enables one to dissect the limits placed on performance by distinct types of stimulus-driven uncertainty versus the limits imposed by noise.

In each of the two double-pass experiments, psychometric data was collected with stimuli sampled from scene locations with two different levels of local-depth variability. There were ten conditions total in each experiment—five fixation disparities (i.e. disparity pedestals) crossed with the two levels of local-depth variability (i.e. disparity-contrast; see Methods). In the first double-pass experiment, all stimuli contained natural luminance-pattern variation and natural local-depth variation. In the second double-pass experiment, "flattened" versions of the stimuli used in the first experiment were used such that local-depth variation was eliminated while leaving luminance patterns essentially unaffected (see Fig. 4.1B).

To obtain the stimuli for the experiments, thousands of stereo-image patches were sampled from a published dataset of stereo-photographs of the natural environment with co-registered laser-based distance measurements at each pixel (Burge et al., 2016). Corresponding points were calculated directly from the range data; a subset of corresponding points is shown in one example stereo-image (Fig. 4.1C). Sampled patches were centered on corresponding points (Fig. 4.1D), with known amounts of fixation disparity at the central pixel, as enforced by a custom stereo-image sampling procedure (A. V. Iyer & Burge, 2018). To quantify local-depth variability (i.e. disparity-contrast), ground-truth disparities were computed at each pixel directly from the distance measurements. The routines upon which the sampling procedures were built were accurate with precision better than ± 5 arcsec (A. V. Iyer & Burge, 2018). Hence, sampling errors are smaller than human stereo-detection thresholds for all but the very most sensitive conditions (Blakemore, 1970; Cormack et al., 1991).

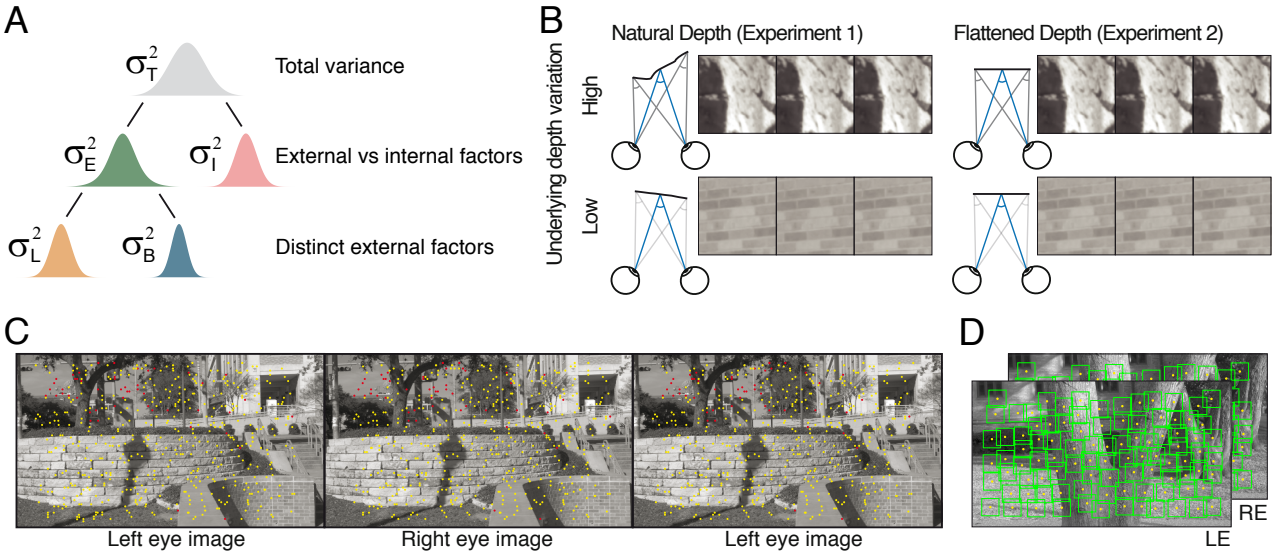

Figure 4.1. Sources of uncertainty in stereo-depth perception, stereo-image database, and experimental stimuli. **A**. The total variance of the decision variable—the variable that signal-detection-theory posits that perceptual decisions are made on the basis of—is contributed to by at least two distinct sources of uncertainty: external (e.g. stimulus-driven) variability σ^2_E and internal noise σ^2_I. The stimulus-driven component can be decomposed into distinct external factors: here, luminance-driven variability σ^2_L and local-depth-driven variability σ^2_B. In natural viewing, luminance-driven variability depends on how luminance-contrast patterns vary across natural stimuli, and depth-driven variability depends on how local-depth structure varies across natural scenes (see *A* and *B*). Traditional psychophysical methods can determine the total variance of the decision variable. Double-pass experiments can partition the total variance into externally- and internally- driven components. The new approach used here can further partition the externally-driven component into distinct external factors. **B**. Two double-pass disparity-discrimination experiments were conducted. Both used images from hundreds of the same natural scene locations. Experiment 1 used stimuli with natural depth profiles (*left*). Local-depth variation, as quantified by disparity-contrast (see Methods Eq. (4.1)), was either high (*top row*) or low (*bottom row*). Experiment 2 used the same stimuli but with flattened versions of the natural depth profiles (*right*). The flattened stimuli (*right*) had the same luminance profiles as the stimuli in Experiment 1, but had no local-depth variation. **C**. Example natural stereo-image from which natural stimuli were sampled for the experiments, with corresponding points overlaid in yellow. Corresponding points were calculated directly from laser-range-based ground-truth distance data at each pixel. Points in one image without a valid corresponding point in the other (e.g. half-occluded scene regions) are colored red. Divergently-fuse the left two images, or cross-fuse the right two images, to see the scene in stereo-3D. **D**. Another example natural stereo-image with patches that were vetted for inclusion in the experimental stimulus set (*boxes*; see Methods). For purposes of visualization, depicted patches are four times wider (4×4°) than those used in the actual experiments (1×1°).

Stimuli were presented using a two-interval, two-alternative forced choice (2AFC) design (Fig. 4.2A). The task was to indicate, with a key-press, which of two natural stereo-image patches, appeared to be farther behind the screen. On each trial, we assume that disparity estimates are obtained for the standard and comparison stimuli: δ_{std} and δ_{cmp}, respectively. Each of these estimates is affected both by properties of the standard and comparison stimuli, and by noise. The decision variable is then obtained by subtracting the standard disparity estimate from the comparison disparity estimates. Distributions of these disparity estimate and decision variable distributions are shown in Figure 4.2B.

4.4.1. Decision-variable correlation

The decision variable underlying performance is given by

$$D = V + W, \tag{4.42}$$

where V is captures the effect of externally-driven, stimulus-based variability and W is a sample of internal noise.

The double-pass experimental design, like a typical (single-pass) experimental design, allows one to estimate the variance of the decision variable. Assuming conditional independence, the total variance of the decision variable is given by

$$\sigma_T^2 = \sigma_E^2 + \sigma_I^2, \tag{4.43}$$

where σ_E^2 is the variance of the externally-driven component and σ_I^2 is the variance of the internally-driven component. The total variance of the decision variable can be computed directly from the discrimination threshold (Fig. 4.2B-C). Specifically, for a certain definition of threshold-level performance (i.e. $d' = 1.0$), which we use here, the total variance of the decision variable simply equals the square of the discrimination threshold (i.e. $\sigma_T^2 = T^2$; see Methods Eq. (4.5)).

The double-pass experimental design, more uniquely, allows one to estimate decision-variable correlation (Fig. 4.2D-F). decision-variable correlation indicates the degree to which the trial-by-trial

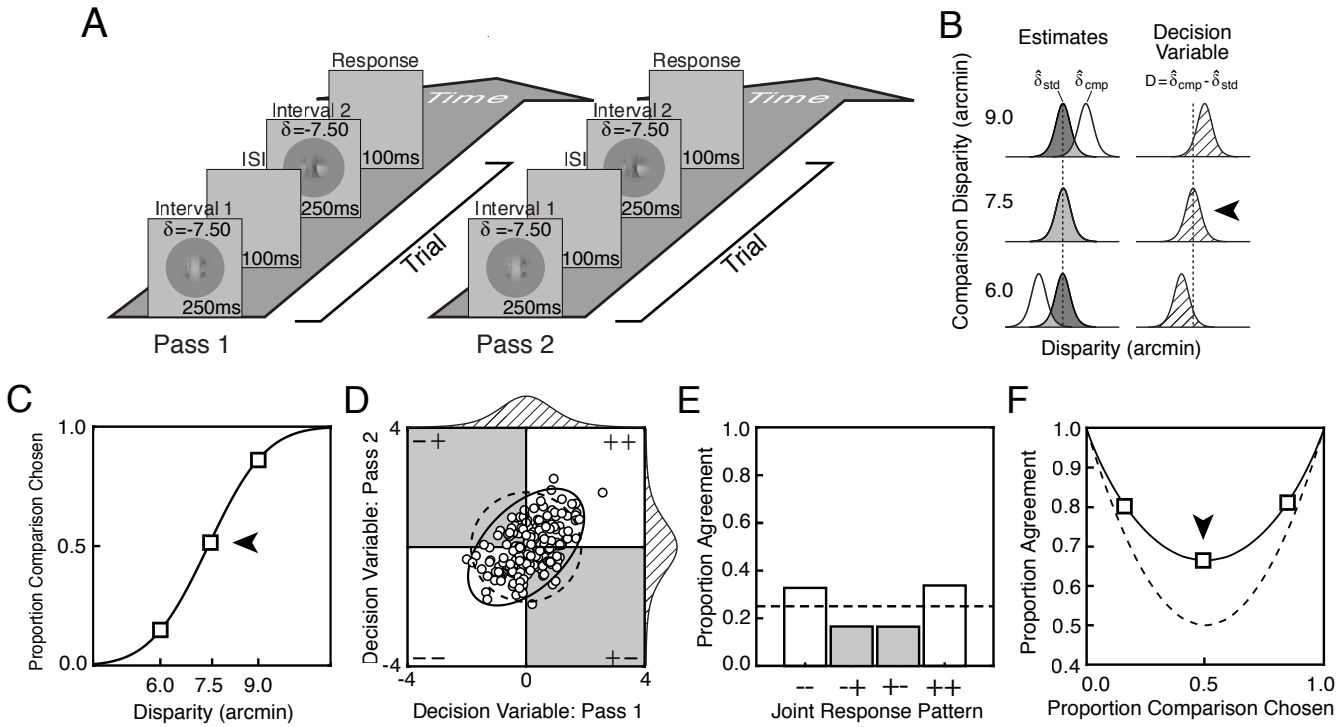

Figure 4.2. Double-pass experimental design. **A**. Each pass of a double-pass experiment is composed of a large number of unique trials, presented one time each. Each trial is composed of a unique pair of natural stimuli. Trials are identical between passes. The task on each trial is to indicate which of two dichoptically presented stimuli appears to be farther behind the screen. **B**. Standard and comparison disparity estimate distributions for each of three comparison disparity levels (*left*), and corresponding decision variable distributions. Each decision variable distribution is obtained simply by subtracting the standard disparity estimate from the comparison disparity estimate on each trial (*right*). **C**. Psychometric data for stereo-depth discrimination with fitted cumulative Gaussian curve, collapsed across both passes of a double-pass experiment. Thresholds or standard deviation of the decision variable are estimated from the variance parameter of the curve. Psychometric data is binary, indicating whether the comparison stimulus was chosen (+) or not (−). Different decision-variable distributions (B) underlie performance at each point on the psychometric function. **D**. Distribution of joint decision variable (*ellipses*) and samples (*dots*) across both passes of a double-pass experiment. Samples in each of the four different quadrants will yield one of four possible joint responses across passes ($--, -+, +-, ++$), two of which indicate agreement ($++$ and $--$). Decision-variable correlations larger than 0.0 evince shared sources of response variability across passes. Dashed ellipse shows joint decision-variable distribution if observer responded completely by chance (correlation of zero). **E**. Histograms show the expected proportion of each of the four joint response types for the joint-decision-variable distribution shown in *B*. **F.** Proportion of between-pass agreement as a function of proportion comparison chosen. Solid line shows best fit to the data. Dashed line shows expected agreement if the observer responded completely by chance (correlation of zero).

values of the decision variable are correlated across passes. It is given by

$$\rho = \frac{\sigma_E^2}{\sigma_T^2}, \tag{4.44}$$

being equal to the proportion of total variability in the decision variable that is due to factors that are common across repeated presentations of the same trial (e.g. external stimulus variability) versus those that are not (e.g. internal noise). decision-variable correlation is estimated from the repeatability of observer responses across passes (Fig. 4.2D-E; see below). On each trial of each pass, the observer reports either that the comparison stimulus appeared farther away than the standard stimulus (+), or that the comparison stimulus appears closer than the standard stimulus (−). Upon completion of both passes, the observer will have made a particular joint response on each unique trial, out of four possible joint responses (−−,−+, +−, ++). When decision-variable correlation equals zero—as it will be when noise is the only source of variability in the decision variable—response agreement is expected to be at chance levels (see Fig. 4.2D-F, dashed lines). When decision-variable correlation is high—as it will be when external factors (e.g. nuisance stimulus variability) are the dominant source of variance in the decision variable—response agreement is expected to be high.

Decision-variable correlation, like other important quantities in signal detection theory (e.g. d'), must be estimated from a set of binary observer responses (Fig. 4.2D-F). We computed how repeatable observers' responses were (i.e. how often observer responses agreed) across the repeated presentations of the same stimuli in the first and second passes of the double-pass experiment (see Fig. 4.2 and Methods). From the level of response agreement in a given condition, we used maximum likelihood techniques to estimate decision-variable correlation across passes.

Decision-variable correlations reflect the *relative* contributions of each individual source of variability in the decision variable (Eq. (4.42)). A change in decision-variable correlation between conditions could result from an increase in one source of variability, a decrease in the other, or a combination of both. Discrimination thresholds provide an *absolute* measure of the total variance in the decision variable. But they do not indicate the relative contribution of external (e.g. stimulus-driven) and in-

ternal (e.g. noise-driven) sources of variability (Eq. (4.43)). Together, discrimination thresholds and decision-variable correlation can be used to determine the absolute contribution of stimulus-driven and internal-noise-driven sources of variability to the decision variable (see Eqs. (4.10) and (4.18)). From estimates of decision-variable correlation (Eq. (4.44)) and of the total variance of the decision variable (Eq. (4.43)), the variances of the externally- and internally-driven components of the decision variable can be computed (see Methods, and below).

4.4.2. Experiment 1: Natural stimuli with natural depth profiles

Figure 4.3 shows raw data from one individual observer in the first double-pass experiment which used stimuli having natural luminance and natural depth profiles. Psychometric data and function fits showing proportion comparison chosen are presented in Figure 4.3A. The slopes of the psychometric functions decrease systematically both as disparity-contrast increases from low to high (top vs. bottom), and as disparity pedestal increases (psychometric functions, left to right). These patterns show that, as the surfaces to be discriminated become more non-uniform in depth (i.e. have higher disparity-contrast), and as they move farther from the fixated distance, discrimination thresholds increase.

Response agreement data and fits for the same observer are shown in Figure 4.3B. The corresponding estimates of decision-variable correlation in each condition are indicated at the top of each subplot. In all conditions, response agreement is systematically higher than expected under the assumption that decision-variable correlation equals 0.0. Indeed, decision-variable correlation is approximately equal to 0.5, on average across the conditions. Thus, the relative contributions of externally- and internally-driven components to the variance of the decision variable are similar (i.e. $\sigma_E^2 \approx \sigma_I^2$; see Eqs. (4.43) and (4.44)). External and internal sources limit performance near-equally. Further, decision-variable correlation is always higher in the high than in the low disparity-contrast conditions (see the inset values of ρ in each subplot). The increase in decision-variable correlation with the level of disparity-contrast entails that the threshold increases are due to more substantial increases in the variance of the stimulus-driven than of the noise-driven component of the decision variable.

Figure 4.4A shows how stereo-based depth discrimination thresholds change with fixation error (i.e.

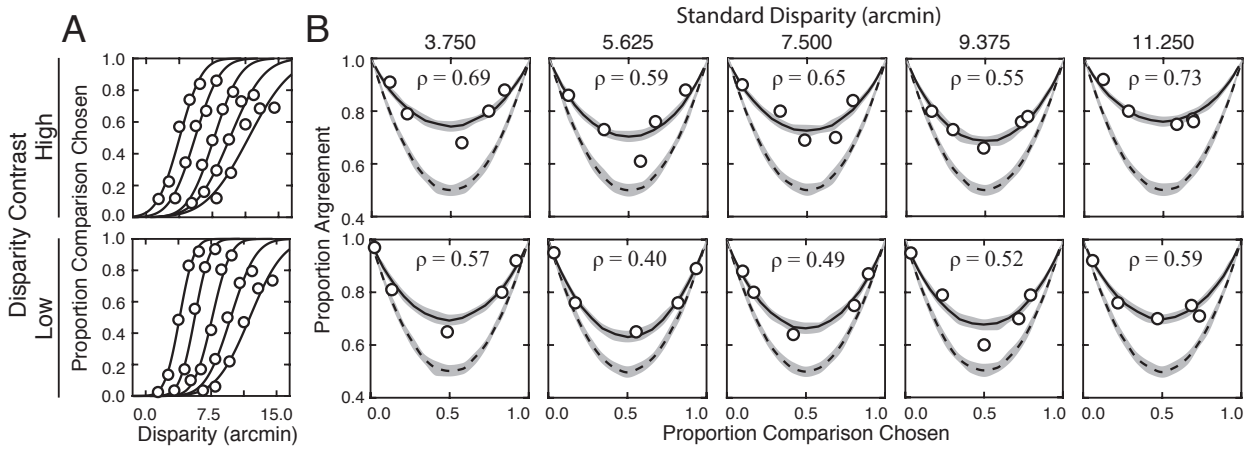

Figure 4.3. Discrimination thresholds, response agreement, and estimates of decision-variable correlation results for one observer. **A**. Response data (*points*) and psychometric curves for each condition. Thresholds increase systematically with disparity pedestal and with disparity-contrast. **B**. Human response agreement and fitted agreement curves for each condition. Thresholds and decision-variable correlation was used to determine relative impact between sources of performance variability.

disparity pedestal) and local-depth variability (i.e. disparity-contrast) for each individual observer, and the observer average. For both disparity-contrast conditions, discrimination thresholds are well-characterized by an exponential function, the signature of which is a straight line on a semi-log plot. This exponential rise in discrimination threshold with pedestal disparity is a classic empirical finding (Badcock & Schor, 1985; Blakemore, 1970; Cormack et al., 1991; McKee et al., 1990; S. B. Stevenson et al., 1992), and is predicted by a normative image-computable model of optimal disparity estimation with natural stereo-images (Burge & Geisler, 2014). The current result provides a psychophysical demonstration that the classic exponential law of human disparity discrimination generalizes to natural stimuli. Because this pattern is robust to the particular stimuli that are used to probe performance, it should be thought of as a feature of how the visual system processes disparity, rather than a consequence of the particular stimuli used to probe performance.

Discrimination thresholds are also higher for stimuli with high disparity-contrast than they are for stimuli with low disparity-contrast. Hence, local-depth variability harms depth discrimination performance. As disparity-contrast increases, thresholds shift vertically in the semi-log plots, such that the two sets of thresholds are parallel to another, indicating that the threshold increases

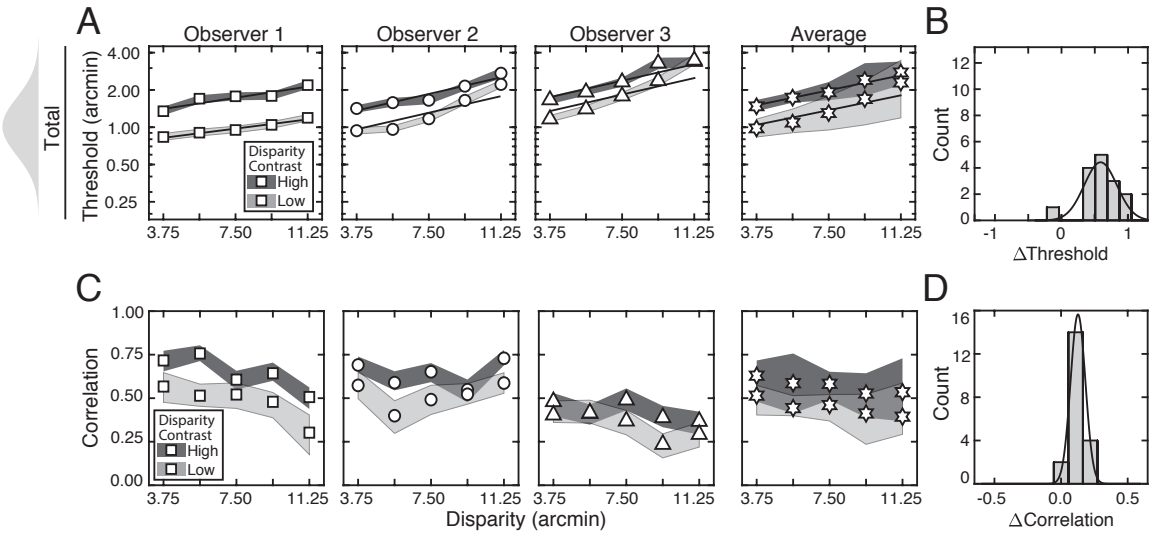

Figure 4.4. Experiment 1 discrimination thresholds and decision-variable correlations. Stimuli in Experiment 1 stimuli had naturally varying local-depth variation. **A**. Discrimination thresholds as a function of disparity pedestals, for different disparity-contrast levels (*shades*), for each observer and the observer average (*columns*). For individual observers, shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped datasets. For the observer average, shaded regions indicate across-observer standard deviations. Lines represent exponential fits to the data in each disparity-contrast condition (see Methods). Discrimination thresholds are equal to the square-root of the total variance of the decision variable (see Eq. (4.5)). **B**. Histogram of threshold differences in the high and low disparity-contrast conditions, collapsed across disparity pedestal and individual observers. Curves indicate best-fit normal distributions to the data. **C**. Estimated decision-variable correlation in the same conditions for each observer and the observer average. **D**. Histogram of differences in decision-variable-correlation differences between the high and low disparity-contrast conditions, collapsed across disparity pedestal and individual observers.

with disparity-contrast are multiplicative. Figure 4.4B visualizes these threshold differences as a histogram, collapsed across all disparity pedestals and observers. Clearly, the histogram of threshold differences is substantially shifted to the right of zero, which confirms that thresholds increase with disparity constraint.

The fact that disparity-contrast degrades discrimination performance should surprise nobody (Banks, 2004; Nienborg et al., 2004; Tyler, 1974; Westheimer, 1979). Increased local-depth variability entails that the left- and right-eye images have more local differences between them. These more pronounced local differences make the stereo-correspondence problem more difficult to solve. The increased difficulty in solving the correspondence problem should, in turn, make stereo-based depth discrimination more difficult. This increase in difficulty is what we observe in our results. However, as we will see, this unsurprising degradation in discrimination performance with disparity-contrast is partly due to a surprising underlying cause (see below).

Decision-variable correlations in each condition for each observer, and for the observer average are shown in Figure 4.4C. In each and every condition, decision-variable correlation is higher in the high disparity-contrast condition than in the low disparity-contrast condition (Fig. 4.4D). This consistent pattern of results indicates that as disparity-contrast increases and the task becomes harder, there is an increase in the proportional impact of external, stimulus-driven components on the decision variable—that is, observer responses become more repeatable, not less.

The externally- and internally-driven contributions to threshold were computed from the estimates of decision-variable correlation and the total variance of the decision variables (i.e. discrimination-thresholds squared (see Eqs. (4.18) and (4.19)), and are shown in Figure 4.5. As with the discrimination thresholds (see Fig. 4.4A)—which reflect the total variance of the decision variable—these individual components also tend to increase exponentially with disparity pedestal (i.e. linearly on semi-log axes; see Fig. 4.5A). However, disparity-contrast impacts these two components differently. The variance of the external component scales with disparity-contrast (Fig. 4.5A top row), and substantially so, whereas the variance of the internally-driven component changes more modestly (Fig. 4.5A bottom row). Thus, the increase in discrimination thresholds with disparity-contrast

can be attributed primarily to increases in the variance of the externally-driven (i.e. stimulus-driven) component of the decision variable. The histograms in Figure 4.5B emphasize this point. They show histograms of the difference in variance between the high and low disparity-contrast conditions in each component, across all observers and disparity pedestals. Clearly, the effect of disparity-contrast on the externally-driven component is more pronounced than the effect on the internally-driven component.

As noted, the fact that discrimination thresholds increase with local-depth variability is to be expected (Banks, 2004). What is unexpected is that a substantial portion of the threshold increases are attributable to factors that make responses more repeatable on successive presentations of the same stimulus. The implication is that, in natural scenes, local-depth variability does not simply make disparity-based depth discrimination noisier, as might be expected if local-depth variability simply made the binocular matching process more unreliable. Rather, the results suggest that local-depth variability biases the observer, stimulus-by-stimulus, to perceive more or less depth in a manner that is repeatable across repeated stimulus presentations. The results therefore imply that, at least in principle, observer errors on each individual stimulus should be predictable. Developing image-computable models that enable stimulus-by-stimulus prediction of depth estimation performance in depth-varying natural scenes is an interesting direction for future work (Burge, 2020).

One potential source of observer repeatable error was that observers were not making disparity estimates based on the very most central pixels of each stimulus. Instead, observers could have been averaging disparities within a window of spatial integration. We investigated this possibility using logistic regression (see Methods), by asking whether disparities averaged within spatial integration windows of fixed size, across a range of sizes, could better account for the observer responses than the disparities associated with the central pixel of each patch. We found that all window sizes accounted for the data equally well. Changing the size of the spatial integration window produced no improved ability to account for explainable variance (R^2). And the Akaike information criterion (AIC) indicated that none of tested spatial integration window sizes produced a significantly better account of the data than the smallest window size that was implicitly assumed throughout the rest

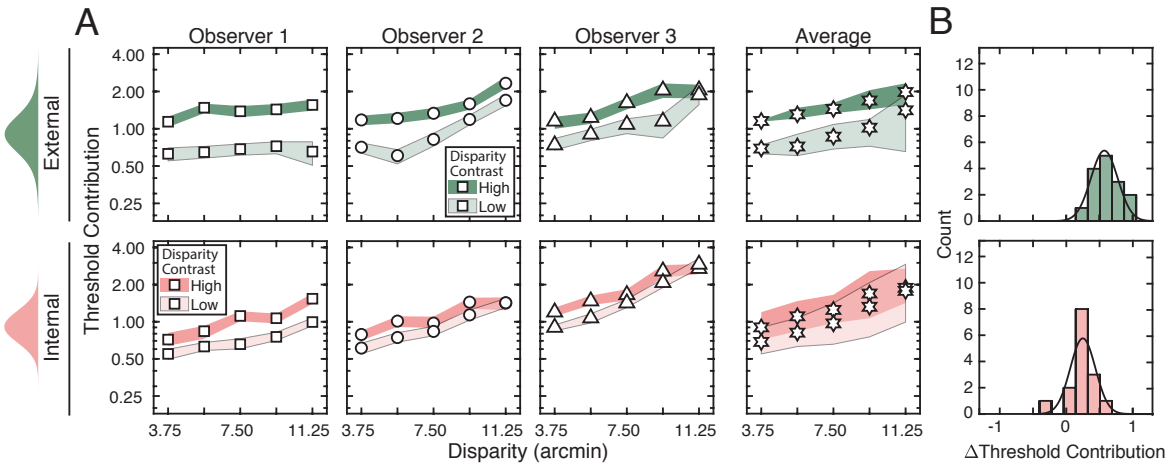

Figure 4.5. External stimulus-driven and internal noise-driven contributions to thresholds in Experiment 1. **A**. Estimated external stimulus-driven (*top row*) and internal stimulus-driven (*bottom row*) contributions to threshold, at all disparity and disparity-contrast conditions, for each observer and the observer average. For observers, bounds of shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped samples. For the observer average, bounds indicate standard deviations. Threshold contribution reflects the variances σ_E^2 and σ_I^2 of the stimulus-driven and internal noise-driven components of the decision variable, respectively (see Methods). **B**. Histograms of differences between high and low disparity-contrast conditions for both externally- and internally-driven components (*top row* and *bottom row* respectively).

of the paper.

Another way to investigate the degree to which stimulus-based variability is predictable is to examine between-observer performance similarities. We assessed whether between-observer-threshold variability is more attributable to differences in the effect of external factors (e.g. stimulus-based variability) or internal factors (e.g. noise) across observers. Figure 4.6 shows how the external, stimulus-based and internal, noise-based contributions to threshold vary across observers relative to the between-observer mean. Between-observer variation in the externally driven-component of the decision variable is substantially smaller than in the internally-driven component (Fig. 4.6). The stimulus-driven component of the decision variable is very similar across human observers, and does not contribute substantially to between-observer differences in discrimination threshold. Because the external drive to the decision variable is consistent across observers, it implies that the stimulus-specific computations performed by the human visual system are stable across observers (also see below). Hence, between-observer variability is primarily due to differences in internal noise.

4.4.3. Experiment 2: Natural stimuli with flattened depth profiles

The second double-pass experiment made use of natural stimuli having "flattened" depth profiles (see Fig. 4.1B). The luminance profiles of these stimuli are essentially unchanged from those in the first experiment, because they were derived from the exact same scene locations, but the disparity-contrasts of all stimuli were set equal to zero. Thus, in Experiment 2, the nominal "high disparity-contrast" and "low disparity-contrast" stimuli had zero disparity-contrast, even though the luminance profiles were drawn from scene regions originally associated with high and low levels of local-depth variability.

The primary aim of the second double-pass experiment is to make it possible to partition the effects of variation in natural luminance contrast patterns and local-depth variation in limiting stereo-depth discrimination. Doing so requires analyzing the data from both experiments simultaneously. Before turning to this joint analysis of the psychophysical data from both double-pass experiments, we first present the results of the second experiment on their own.

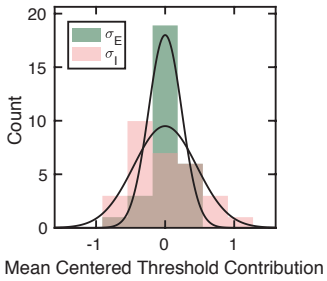

Figure 4.6. Between-observer variability is primarily attributable to differences in internal noise. Observer-mean subtracted estimates of externally-driven σ_E^2 (green) and internally-driven σ_I^2 (pink) components of the decision variable, histogrammed across conditions. Black lines represent best-fit normal distributions. Across the high and low disparity-contrast conditions, the fraction of between-observer variance explained by the internally-driven component for Experiment 1 was 0.81 ($p = 2.0 \times 10^{-4}$, $F = 0.23$ where F is the test statistic of a two-sample F-test).

Figure 4.7 shows discrimination thresholds (i.e. the square-root of the total variance of the decision variable), and decision-variable correlations across all conditions in Experiment 2, for each individual observer and the observer average. There is one marked change in the patterns in the data as compared to the first experiment. Discrimination thresholds (Fig. 4.7A-B) and decision-variable correlations (Fig. 4.7C-D) are now largely unaffected by nominal disparity-contrast. There are also consistent decreases in thresholds and decision-variable correlations, as compared to Experiment 1 (see Fig. 4.4). These results imply that a source of stimulus-driven variance in the decision variable that increases response agreements across repeated stimulus presentations, has been removed from the stimuli.

Analysis of the external (stimulus-driven) and internal (noise-driven) contributions to threshold lead one to the same conclusion: flattening the stimuli removes a stimulus-driven source of variance in the decision variable that is due to local-depth variability (Fig. 4.8). Neither the external drive to threshold (Fig. 4.8, top row), nor the internal drive to threshold (Fig. 4.8, bottom row), is affected by nominal disparity-contrast.

Of course, this change in the pattern of results makes sense. The "high disparity-contrast" and "low disparity-contrast" stimuli in Experiment 2 had been associated with depth varying regions of natural scenes in Experiment 1, but they were flattened for the current experiment. So the result

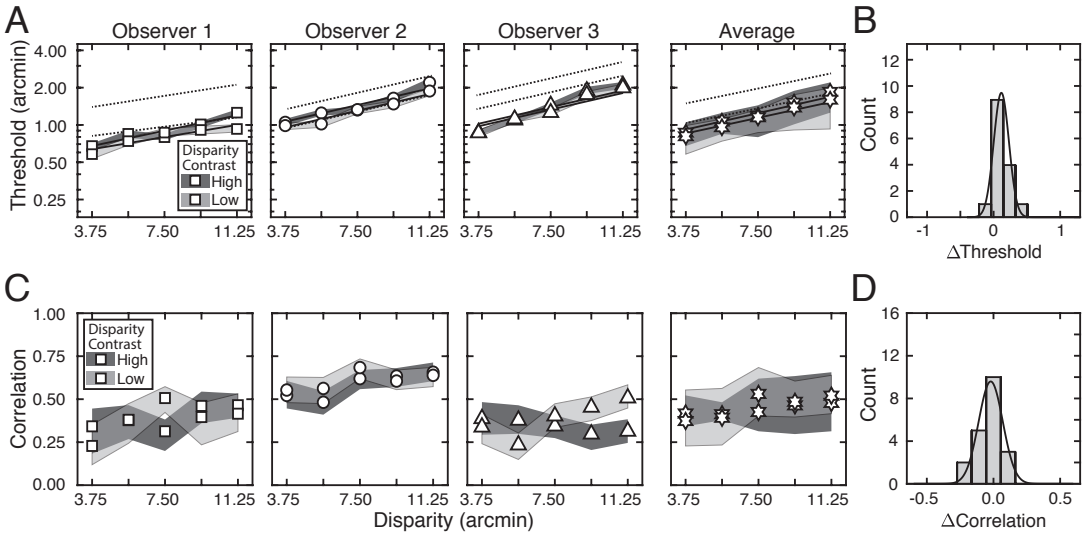

Figure 4.7. Experiment 2 disparity discrimination thresholds and decision-variable correlation. Experiment 2 stimuli were flattened (i.e. had zero local-depth variability), but otherwise had the same luminance contrast patterns as those in Experiment 1. **A**. Discrimination thresholds as a function of disparity pedestal, for different disparity-contrast levels (*shades*), for each observer and the observer average (*columns*). Unlike in Experiment 1, there is little to no effect of nominal disparity-contrast on threshold. For individual observers, shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped datasets. For the observer average, shaded regions indicate standard deviations. Solid lines represent exponential fits to the data. Dotted lines represent the exponential fits to the threshold data from Experiment 1 (see Fig. 4.5A). **B**. Histogram of threshold differences in the high and low disparity-contrast conditions, collapsed across disparity pedestal and individual observers. Curves indicate best-fit normal distributions to the data. **C**. Estimated decision-variable correlation in the same conditions for each observer and the observer average. Decision-variable correlations are systematically lower than those in Experiment 1 (see Fig. 4.4). **D**. Histogram of decision-variable-correlation differences in the high and low disparity-contrast conditions, collapsed across disparity pedestal and individual observers.

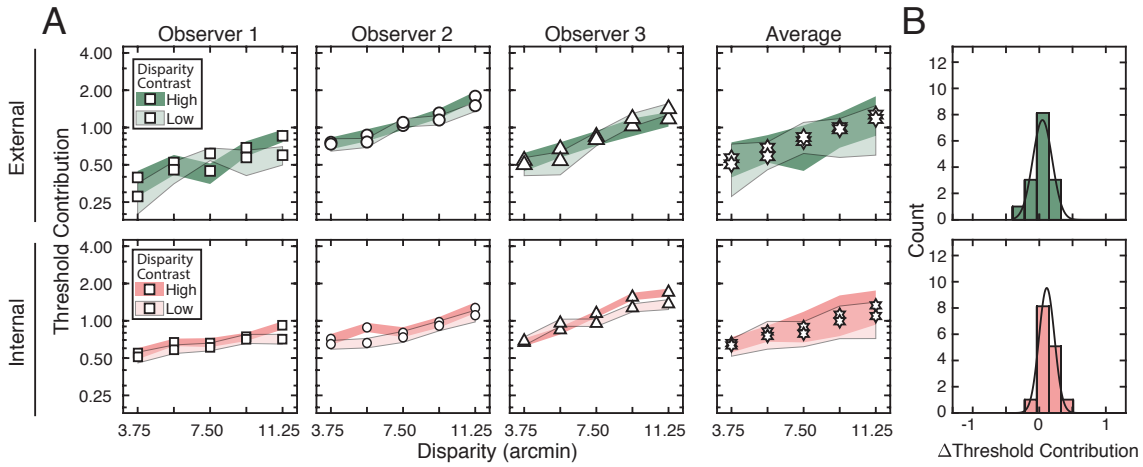

Figure 4.8. External stimulus-driven and internal noise-driven contributions to thresholds in Experiment 2. **A**. Estimated external stimulus-driven (*top row*) and internal stimulus-driven (*bottom row*) contributions to threshold, at all disparity and disparity-contrast conditions, for each observer and the observer average. For observers, bounds of shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped samples. For the observer average, bounds inidcate across-observer standard-deviations. Threshold contribution reflects the variances σ_E^2 and σ_I^2 of the stimulus-driven and internal noise-driven components of the decision variable, respectively (see Methods). Note that, in comparison to the results of Experiment 1, there is hardly any effect of disparity-contrast on the stimulus-driven contributions to threshold. **B**. Histograms of differences between high and low disparity-contrast conditions for both externally- and internally-driven components (*top row* and *bottom row* respectively).

is not unexpected. But it is also not guaranteed. The effect of natural depth variability in bumpier (higher disparity-contrast) scene regions on the decision variable could have been correlated with the effect of natural luminance contrast patterns such that, even with flattened stimuli, the luminance profiles associated with the high disparity-contrast regions of the scene would have generated higher discrimination thresholds. That is, luminance profiles associated with scene locations having greater local-depth variability could themselves have been more difficult to discriminate, even after stimulus-flattening. The current results suggest that this is not the case.

Because of the fact that, in the first double-pass experiment, high disparity-contrast stimuli yielded high levels of externally-driven variance in the decision variable and low disparity-contrast stimuli yielded lower levels of externally-driven variance (see Fig. 4.5A), the current results strongly imply that a stimulus-driven, and repeatable, source of variability has been removed from the decision variable. The flattened stimuli of the second double-pass experiment also yield the lowest levels of externally-driven variability in the decision variable. Together, these results imply that stimulus flattening removes a distinct source of variability to the decision variable. This idea is tested more rigorously below.

4.4.4. Partitioning sources of variability in natural stimuli

Here, we show that stimulus-driven variability in the decision variable can be partitioned into separate factors that depend on natural luminance and natural depth structure. These sources of variability—natural luminance structure and natural depth structure—have distinct and largely separable effects on human performance. To determine the importance of these two factors, and to test whether these factors interact, we compared human performance across the four passes of the two double-pass experiments with flattened and natural stimuli. We refer to this comparative analysis as a quasi-quadruple-pass analysis (see Methods). (Note that typical quadruple-pass experiments—to the extent that quadruple-pass experiments are ever typical—present *exactly* the same stimuli across all four passes. Our experiments presented similar, but not identical, stimuli across the four passes of the two double-pass experiments, hence the "quasi-quadruple-pass" moniker.)

Luminance-contrast pattern variability was essentially the same in both double-pass experiments, and was thus the same across all four passes. However, because the second double-pass experiment used flattened stimuli—which prevents local-depth variability from directly influencing the variance of the decision variable—natural luminance variation is the only remaining stimulus factor that can contribute to the decision variable because natural depth variability has been eliminated. The quasi-quadruple-pass analysis allows one to determine how these two factors combine and/or interact to limit performance.

To understand the reasoning behind the quasi-quadruple-pass analysis, it is useful to write out expanded expressions for the decision variable (also see Eq. (4.42) above). The expanded expression for the decision variable is given by

$$D = \overbrace{(L + B)}^{V} + W, \tag{4.45}$$

where L and B are luminance profile driven and local-depth-variability driven contributions to the decision variable (which sum to the total stimulus-driven contribution V), and W is a sample of internal noise.

In the double-pass experiment with natural luminance and depth profiles (Exp. 1), the expressions for the total variance of the decision variable and for decision-variable correlation across passes, in terms of the variance of these newly articulated components (i.e. L and B in Eq. (4.45)), are given by

$$\sigma^2_{T_*} = \overbrace{(\underbrace{\sigma^2_L + \sigma^2_B + 2\text{cov}[L, B]}_{\text{unknowns}})}^{\sigma^2_{E_*}} + \sigma^2_{I_*}, \tag{4.46}$$

$$\rho_{**} = \frac{\sigma^2_{E_*}}{\sigma^2_{T_*}} = \frac{\sigma^2_{E_*}}{\sigma^2_{E_*} + \sigma^2_{I_*}}, \tag{4.47}$$

where σ^2_L and σ^2_B are the variances of the components driven by luminance profile and local-depth

variability, the interaction term cov$[L, B]$ is the covariance between them (if it exists), $\sigma^2_{E_*}$ is the external (stimulus-driven) variance, and $\sigma^2_{I_*}$ is the variance of internal noise. The external stimulus-driven- and internal noise-driven variances can be solved from the equations for total variance and decision-variable correlation (Eqs. (4.46) and (4.47)). But there are not enough equations to separately determine the values of the three unknown factors: the variance σ^2_L of component driven by luminance-pattern variability, the variance σ^2_B of component driven by local-depth variability, and the covariance cov$[L, B]$ between the luminance and depth driven components. Fortunately, the second double-pass experiment allows one of these unknown factors—the variance of the luminance-driven component of the decision variable—to be determined.

In the second double-pass experiment with natural luminance profiles and flattened depth profiles (Exp. 2), the expanded expression for the decision variable is given by

$$D = \overbrace{(L \qquad)}^{V} + W. \tag{4.48}$$

Note that the disparity-contrast driven component B that is present in the first experiment does not appear in Eq. (4.48), because disparity-contrasts were set equal to zero when the stimuli were flattened. The corresponding expressions for the variance of the decision variable, and decision-variable correlation, are given simply by

$$\sigma^2_{T\dagger} = \overbrace{(\sigma^2_L \qquad\qquad)}^{\sigma^2_{E\dagger}} + \sigma^2_{I\dagger}, \tag{4.49}$$

$$\rho_{\dagger\dagger} = \frac{\sigma^2_{E\dagger}}{\sigma^2_{T\dagger}} = \frac{\sigma^2_{E\dagger}}{\sigma^2_{E\dagger} + \sigma^2_{I\dagger}}, \tag{4.50}$$

where, again, σ^2_L is the luminance profile driven variance, $\sigma^2_{E\dagger}$ is the external stimulus-driven variance, and $\sigma^2_{I\dagger}$ is the internal-noise-driven variance associated with the flattened stimuli. Just as before, the external and internal variances can be estimated from Equations 4.49 and 4.50. Now, the variance of the luminance-pattern-driven component σ^2_L is easily obtained because it exactly

equals the variance of the externally-driven component. Also note that in this experiment, because local-depth variability is absent, the variance of the disparity-contrast-driven component is zero. But there are still two remaining unknowns.

Here is where the quasi-quadruple-pass analysis proves useful. By computing decision-variable correlation across passes of the two different double-pass experiments, an additional equation is obtained. Decision-variable correlation between passes across experiments is given by

$$\rho_{\dagger *} = \frac{\sigma_L^2 + \mathrm{cov}[L, B]}{\sigma_{T\dagger}\ \sigma_{T*}}. \tag{4.51}$$

With this expression, we now have the number of equations necessary to determine the unknowns. Using maximum likelihood techniques, we fit all three decision-variable correlations ($\hat{\rho}_{\dagger\dagger}$, $\hat{\rho}_{**}$, and $\hat{\rho}_{\dagger *}$) simultaneously from the data in both experiments with the quasi-quadruple-pass analysis (see Eq. (4.28)), and then solved algebraically the system of equations specified by Eqs. (4.46), (4.47) and (4.49) to (4.51) for the unknown parameters. This approach guarantees that shared factors between equations are consistent with one another.

Before proceeding to the main results, we briefly note that we have already estimated decision-variable correlation across passes in the first experiment and in the second experiment—$\hat{\rho}_{**}$ and $\hat{\rho}_{\dagger\dagger}$, respectively—, in each case using data only from the respective experiment in isolation. When carrying out the quasi-quadruple-pass analysis, the estimates of the within-experiment decision-variable correlations (i.e., $\hat{\rho}_{**}$ and $\hat{\rho}_{\dagger\dagger}$) and the variances of the externally- and internally-driven components (i.e., σ_E^2 and σ_I^2) are not guaranteed to be the same as when they are estimated with the data from only one isolated experiment (see Fig. 4.5 and 4.8. Reassuringly, however, the estimates from the quasi-quadruple-pass analysis are very similar to those previously estimated. This consistency supports the claim that factors assumed to be common to both experiments are in fact common to both experiments (see Fig. 4.9). The consistency which these parameters vary across experiments and observers, suggests that each component of the decision variable is indeed driven by the natural-image property—or a tight co-variate of the property—that is said to drive it.

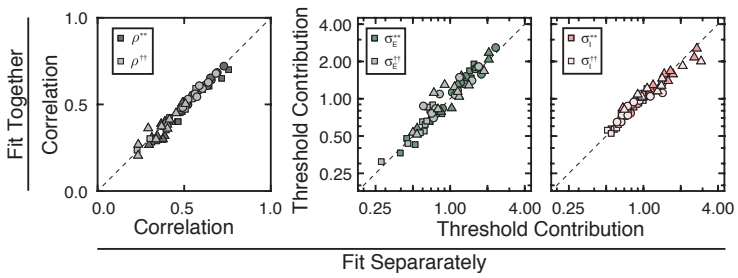

Figure 4.9. Robustness of fitting methods. Comparison of values obtained from fitting data from Experiments 1 and 2 separately (see Figures 4.5 and 4.8) versus together with a quasi-quadruple pass analysis. For decision-variable correlation (*left*), and threshold contributions by stimulus-driven factors (*middle*) and internally-driven factors (*right*), results are consistent regardless of the analytical approach. The consistency of the results indicates the validity and robustness of the quasi-quadruple pass analysis.

Figure 4.10 shows the recovered values of the luminance- and depth-driven components of the decision variable—σ_L^2 and σ_B^2, respectively—, and their interaction term $\mathrm{cov}[L, B]$, that were obtained from the quasi-quadruple-pass analysis (see above; also see Methods). The variances of both the luminance-driven and local-depth-driven components clearly increase with disparity pedestal for all conditions and observers. This pattern is similar to the patterns in all previous plots. More interestingly, whereas the luminance-driven component is very nearly unaffected by the level of disparity-contrast (Fig. 4.10A-B top row), the local-depth-driven component has substantially higher variance with high than for with low disparity-contrast stimuli (Fig. 4.10A-B bottom row).

These points are emphasized by histograms of the differences in the values of these components in the low and high disparity-contrast conditions. Although the luminance-pattern-driven component is essentially invariant to it (Fig. 4.10B), the local-depth-driven component changes substantially with disparity-contrast (Fig. 4.10D). From these results we conclude that the variance of luminance-driven component of the decision variable is a function of pedestal disparity but not disparity-contrast $\sigma_L^2(\delta_{std})$, and that the local-depth-driven component is a function of both factors $\sigma_B^2(\delta_{std}, C_\delta)$, a finding that strongly suggests that the components are not substantively affected by a potential common cause (e.g. local-depth variability). Overall, these results support the conclusion that natural luminance-pattern variability and natural local-depth variability in real-world scenes have separable effects on stereo-based depth discrimination performance.

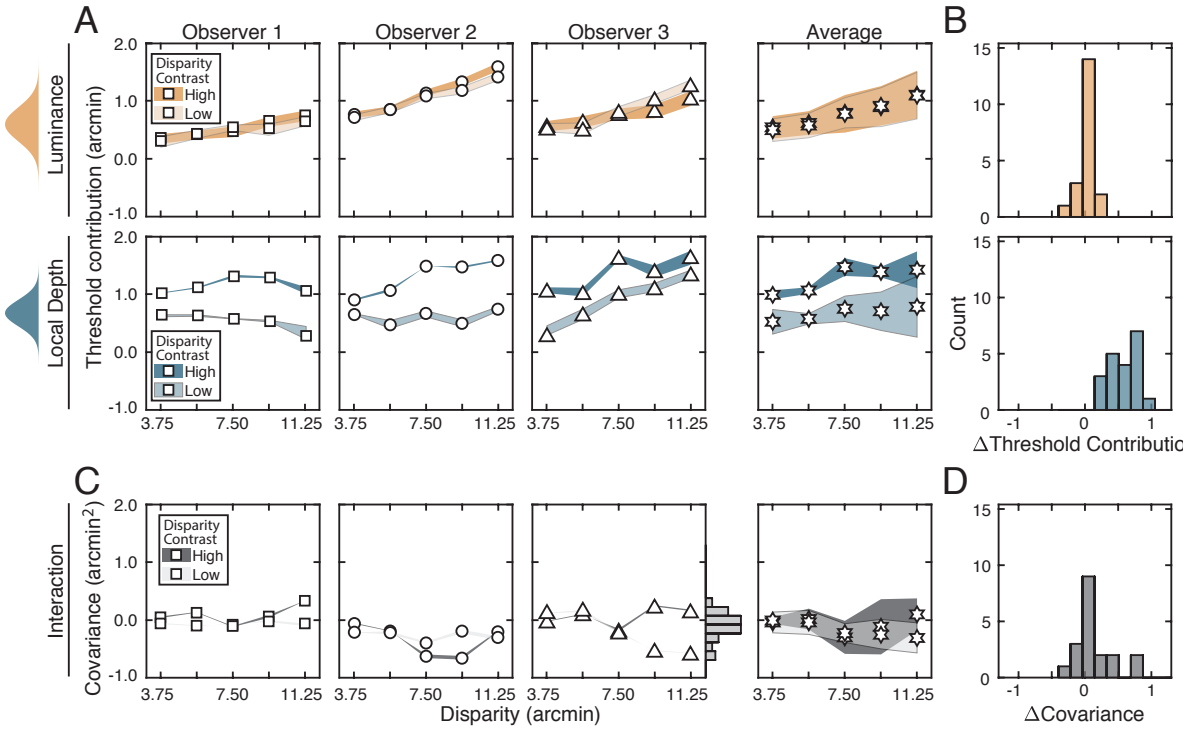

Figure 4.10. Contributions of distinct stimulus-specific factors to thresholds, as revealed by the quasi-quadruple-pass analysis. **A**. Contribution of luminance-contrast pattern variability (*top*) and variability in local-depth structure (*bottom*) to threshold as a function of disparity pedestal at different disparity-contrast levels (*shades*), for each observer and the observer average. For individual observers, bounds of shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped samples. For the observer average, bounds indicate across-observer standard-deviations. **B**. Histogram of differences in luminance-pattern-driven and local-depth-driven threshold contributions across high and low disparity-contrast conditions, collapsed across disparity pedestals and individual observers. **C**. Same as *A*, but for the interaction term (i.e. $\text{cov}[L, B]$). Histogram of the interaction terms collapsed across all disparity pedestals, disparity-contrasts, and individual observers is shown on the rightmost y-axis of the third column (mean=-0.11, sd=0.23). **D**. Histogram of differences in the interaction term (i.e. $\text{cov}[L, B]$) across high and low disparity-contrast conditions, collapsed across disparity pedestals and individual observers. Data in *C-D* indicate that the interaction term is near-zero in all conditions.

Note that the value of the interaction term is near-zero for all conditions (Fig. 4.10C-D). Refitting the data with a model that fixes the interaction term to zero yields estimates of σ_L^2, σ_B^2, $\sigma_{I_\dagger}^2$, and $\sigma_{I_*}^2$ that are robust to whether the constraint on the interaction term is imposed during fitting; any qualitative description that applies to one set of fitted results applies to the other. Fits with and without the constraint also yield near-identical log-likelihoods. Just as the fitted results are robust to whether data from the two double-pass experiments are fit together (with the quasi-quadruple-pass analysis) or separately (see Fig. 4.9)), the estimates of whether luminance-pattern- and local-depth-driven sources of variance do not covary with one another.

One might have expected different results. A given scene location gives rise to the luminance patterns in the left- and right-eye images, and to the pattern of binocular disparities between them. So, one might predict that the effect of a given luminance contrast pattern (i.e. photographic content) would be tightly correlated with the effect of the corresponding local-depth variation on disparity discrimination performance, for the simple reason that they have a potential common cause: both are largely determined by the same location in the scene. The current results show that this is not the case. Rather, the results strongly suggest that each of these natural-stimulus-based sources of variability in the decision variable are near-independent of one another.

4.4.5. Shared stimulus drive between observers

Earlier, we presented data showing that between-observer variability (i.e. threshold differences) was driven more by observer-specific differences in internal noise than by observer-specific differences in stimulus-driven variability (see Fig. 4.6). We speculated that this result was due to a high degree of similarity between the computations that different humans use to extract useful information from each stimulus for the task. Here, we present data from between-observers decision-variable correlations that bolster the case.

Between-observers decision-variable correlation quantifies the similarity of the decision variable in two different observers across repeated presentations of the same stimuli. If different human observers are using the same computations to estimate and discriminate stereo-defined depth from natural stimuli, stimulus-by-stimulus disparity estimates from one human should be correlated with

those from a second—that is, between-observers decision-variable correlation will be substantially larger than zero (assuming internal noise is not too large). On the other hand, if subjects are using quite different computations to process stimuli, stimulus-by-stimulus estimates or trial-by-trial responses from one observer will provide no information about estimates or responses from another, and between-subjects decision-variable correlation should equal zero.

We computed between-observers decision-variable correlation from response agreement data by straightforward adaptation of the quasi-quadruple-pass analysis (see Methods). However, because between-observers correlation is impacted by internal noise, its value does not transparently reflect the level of shared stimulus drive. The partial correlation does. Partial correlation is given by

$$\rho_{12 \cdot W} = \frac{\rho_{12}}{\sqrt{\rho_{11}\rho_{22}}} = \frac{\mathrm{cov}[S_1, S_2]}{\sigma_{E1}\sigma_{E2}}, \tag{4.52}$$

where ρ_{12} is between-observers decision-variable correlation, ρ_{11} and ρ_{22} are the within-observer decision-variable correlations, S_1 and S_2 are the stimulus-driven components of the decision variable that are shared between the two observers, and σ_{E1} and σ_{E2} are the standard-deviations of the stimulus-driven components of the decision variable in the two observers. This partial correlation provides more unvarnished information about what we are most interested in, because it is unaffected by internal noise. It quantifies the level of correlation in the stimulus-driven component of the decision variable between observers (see Methods).

Between-observers partial correlations are shown in Figure 4.11. Across all conditions and observer pairs, between-observers partial correlations are substantially above zero. In the high disparity-contrast conditions of Experiment 1, which are the conditions in which local-depth variability has its largest effects, between-observers partial correlations are 0.79 on average, with some values approaching the maximum possible value (i.e. 1.0). In the low disparity-contrast conditions of Experiment 1, the average value is 0.59. In Experiment 2, the average partial correlations for the high and low disparity-contrast conditions are 0.56 and 0.53, respectively (Fig. 4.11 bottom row). Histograms of the differences between the high- and low-disparity-contrast conditions are shown in Figure 4.11B. And histograms of the raw values are shown in Figure 4.11C.

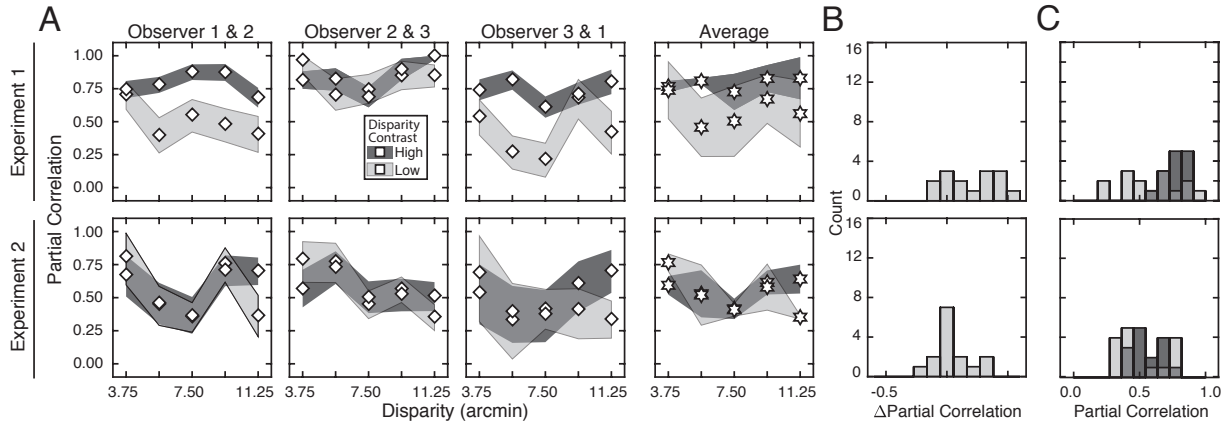

Figure 4.11. Between-observer correlation in the stimulus-driven component of the decision variable, as revealed by the quasi-quadruple-pass analysis. **A**. Estimated partial correlation values, controlling for (i.e. removing) the affect of internal noise, between all observer pairs, for each experiment (*rows*), at all disparity and disparity-contrast levels. Averages across observer-pairs are shown in *column 4*. With the effect of internal noise removed, only the stimulus-driven component of the decision variable drives between-observer correlation. For observer pairs, bounds of shaded regions indicate 95% confidence intervals for each condition from 1,000 bootstrapped datasets. For the across-observer-pair average, bounds of shaded region indicates across-pair standard deviations. **B**. Histogram of differences in partial correlation across high- and low-disparity-contrast conditions shown in *A*, collapsed across disparity pedestals and observer pairs. **C**. Histograms of the raw partial correlations for each observer pair in *A*. In Experiment 1, the mean partial correlations are 0.79 and 0.59 in the high- and low-disparity-contrast conditions, respectively. In Experiment 2, the values are 0.56 and 0.53. The majority of the stimulus-driven variance is shared between observers.

These results indicate that the majority—and, in one case, the strong majority—of the stimulus-driven component of the decision variable is shared between observers. That is, natural stimulus variability associated with different stimuli having same the latent variable (i.e. disparity) causes similar stimulus-by-stimulus over- and under-estimations of disparity-defined-depth in different humans. We conclude that the deterministic computations that the human visual system performs on individual stimuli are largely consistent across observers.

Chin and Burge (2020), in the the domain of speed discrimination, came to a similar conclusion using a related approach. By comparing human performance to that of an image-computable ideal observer, they found that differing levels of human inefficiency are near-exclusively attributable to different levels of internal noise. Like the current findings, this finding entailed that the variance of the stimulus-driven component of the decision variable is quite similar across different human observers, and is consistent with the visual systems of different human observers performing the same deterministic computations on the stimuli. The dovetailing evidence from the current study of disparity-based depth discrimination and the previous study of speed discrimination suggest that natural stimulus variability (natural variation in luminance pattern and/or depth-structure) has consistent effects on the visual systems of different human observers. These results suggest that evolution has tightly honed the details of how visual systems compute so that they extract the most useful task-relevant information from natural stimuli.

4.5. Discussion

In this article, using a natural-stimulus dataset, two double-pass experiments, and a series of analyses, we investigated human stereo-depth discrimination in natural scenes, with specific emphasis on how natural-stimulus variability limits performance. We sourced stimuli from a natural stereo-image database with a constrained sampling procedure, and computed ground-truth disparities directly from laser-range data at each pixel. Fixation (or pedestal) disparity, and local-depth variability—as quantified by disparity-contrast—were tightly controlled. Luminance-contrast patterns and local-depth structures were allowed to vary naturally across the hundreds of unique stimuli that were sampled for each condition.

We find that the exponential law of disparity discrimination holds for human vision in natural scenes. We find that stimulus-driven variability and noise-driven variability have near-equal roles in setting these thresholds, and that the stimulus-based sources of variability make responses more repeatable (and thus potentially more predictable) across repeated stimulus-presentations. We find that one of two underlying causes of the stimulus-driven variability is attributable to local-depth variation, multiplicatively increases discrimination thresholds, and is largely separable from luminance-contrast-pattern variation. And we find that different subjects make correlated stimulus-by-stimulus over- and under-estimations of disparity, suggesting that the different human visual systems process individual natural stimuli with computations that are largely the same.

The approach developed here extends the rigor and interpretability that has been integral to progress in more traditional psychophysics and neuroscience experiments to more natural-stimulus sets Chin and Burge, 2020; Sebastian et al., 2017; Ziemba et al., 2016. In the real world, perceptual, and behavioral variability is driven by both external and internal factors. A comprehensive account of perceptual and behavioral variability, and the neural activity underlying it, must identify and describe the impact of all significant sources of performance-limiting variability. Encouragingly, the current results raise the prospect that an appropriate image-computable model may, in principle, be able to predict a substantial proportion of stimulus-by-stimulus variation across natural images.

4.5.1. Progress and limitations

Progress in science is often incremental. Many times, it occurs by way of relaxing one experimental design element, while holding others fixed. We and others have investigated perceptual performance with stimuli sampled from natural scenes—which are atypical of laboratory experiments—, while using conventional, tightly controlled, laboratory tasks (Burge & Geisler, 2015; Chin & Burge, 2020; Sebastian et al., 2017; Ziemba et al., 2016). Others have investigated performance with atypical tasks (e.g. free viewing and unconstrained eye-movements), while using conventional (e.g. Gabor) stimuli (Yates et al., 2023). Both approaches have increased the ecological validity of the experimental conditions, and have provided new insights into the properties of neural computations underlying sensory-perceptual performance. But there are always limitations.

The stimuli used in the current experiments were foveally presented and subtended only 1° of visual angle, the approximate size of foveal receptive fields in early visual cortex. Foveal presentation of spatially-limited stimuli is common in psychophysical experiments, but doing so prevents the assessment of peripheral visual processing or how performance is affected by the dynamic interplay between eye, head, and body movements occurring in natural viewing. Limiting stimulus size to one degree also limits the degree to which contextual effects can affect performance. In the context of this task, however, there was no evidence that the visual system was spatially integrating over areas any larger than the very most central pixels of each stimulus (see Resutls). Experiments—possibly with larger stimuli, that are specifically designed to examine contextual effects could be an interesting topic for future work.

Related issues concern the two-alternative forced choice (2AFC) procedure used in the current experiments. Although commonly employed, the rigid trial structure imposed by such designs is not well-aligned with how perceptual estimates, perception-driven decisions, and perception-guided action are inter-related in natural viewing. Alternative methods, such as continuous psychophysics, that more closely reflect the continuous interplay of perception and action in natural viewing, could complement the current findings (Bonnen et al., 2015; Burge & Cormack, 2020; Chin & Burge, 2022).

Despite these limitations, the current experiments showed that the natural variation of luminance-contrast patterns and local-depth structures have large, distinct, and identifiable effects on performance. Developing methods that guide the judicious choice of stimulus sets and tasks that strike an appropriate balance between fully natural and tightly constrained, that are well-suited to available analytical methods, and are well-matched to the specific research question under study, will be increasingly important as the science becomes more focused on understanding how neurons respond and how perception works in the natural environment.

4.5.2. Performance variation and prediction

An ultimate goal of perception science is to be able to predict, from an individual stimulus, the neural activity and subsequent perceptual estimate, whether it will be accurate or inaccurate, and

whether it will be reliable or unreliable. The degree to which this goal is achievable hinges on the degree to which the stimulus-by-stimulus estimates are controlled by the properties of the stimulus, as opposed to noise. If the strong majority of performance variation is noise-driven, such efforts will be futile. So, before undertaking to develop and test models that make stimulus-by-stimulus predictions, it is prudent to demonstrate that a substantial proportion of performance variation is stimulus driven. In the current stereo-depth discrimination experiments, natural-stimulus-based sources of response variability account for approximately half of all performance-limiting variability (see Fig. 4.4), a substantial proportion of which was shared across observers (see Fig. 4.11).

However, while the stimuli—stereo-photographs of natural scenes—were allowed to vary naturally in many respects, the mean luminance was fixed to a comfortable photopic level, and luminance-contrast was set to the median contrast in natural scenes (see Methods) (Frazor & Geisler, 2006; A. Iyer & Burge, 2019). Both properties are known to impact stereo-depth discrimination performance (Cormack et al., 1991), and stimulus detection performance in general (Burgess et al., 1981; Legge & Foley, 1980; Mueller, 1951; Nachmias & Sansbury, 1974; Sebastian et al., 2017). Indeed, as mean-luminance and luminance-contrast increase, neurons respond more vigorously, signal-to-noise ratios increase, and performance becomes more reliable (Frazor & Geisler, 2006; Mante et al., 2005). Hence, if luminance and contrast had been allowed to vary more naturally, the proportional contribution of stimulus-based factors to performance-limiting variability is likely to increase. The current estimates of stimulus-based contributions to the decision variable may therefore be underestimates of the total impact that stimulus-based factors would have in less tightly controlled circumstances. This speculation is supported by the fact that between-observers partial correlations are near the maximum possible values in the conditions in which natural stimulus variability was highest (see Fig. 4.11).

The power of empirical datasets to help develop, constrain, and evaluate models can be improved by presenting unique stimuli on each trial. Many models can yield similar predictions of performance if only summary statistics (e.g. bias and precision) are used to evaluate the models' successes and failures. Image-computable models that predict decision-variable correlation and stimulus-by-

stimulus estimates (or discriminations), in addition to bias and/or precision, can provide increased power for evaluating hypotheses about the neural activity and sensory-perceptual computations underlying performance (Burge, 2020; Chin & Burge, 2020; W. S. Geisler, 2018).

4.5.3. Noise and its impact on performance

In this article, we sought to partition the influence on performance of stimulus-driven from noise-driven variability, and to further partition the effects of two distinct types of natural-stimulus variability: luminance-pattern and local-depth variability. We made no attempts to determine different potential sources of noise (i.e. stimulus-independent sources variability), or to partition the influence of each on performance. As a consequence, any source of variance that led to less repeatable responses in the current experiments contributed to the estimate of noise variance. We conceptualized the noise as occurring at the level of the decision variable. But there are multiple stages in the chain of events preceding perceptual estimatation, both external and internal to the organism, where such variability could have originated and that would be consistent with the results.

Variation due to noise could have occurred during the initial encoding of the retinal image, in early visual cortex, at the decision stage (e.g. in the placement of the criterion), or a combination of these possibilities. Potential sources of such variation include the noisy nature of light itself (Hecht, 1942), random fixational errors (Ukwade et al., 2003), neural noise (Tolhurst et al., 1983; Tomko & Crapper, 1974), and trial-sequential dependencies (Laming, 1979). Higher-level factors could also manifest as noise, including stimulus-independent fluctuations in alertness, attention, or motivation (J. I. Gold & Ding, 2013; Lu & Dosher, 1998; Mitchell et al., 2009; Zhang et al., 2018).

Experimental and computational methods that can determine the contribution of different types of stimulus-independent sources of variation are of interest to systems neuroscience (Chin & Burge, 2020; Goris et al., 2014). There are clear steps that could be taken to identify and account for some of these potential sources of noise. Psychophysical methods have the potential to distinguish some of them. High-resolution eye-tracking would allow one to condition performance on the fixational state of the eyes (Bowers et al., 2019; Rucci & Poletti, 2015; S. Stevenson et al., 2016). Parametrically varying performance-contingent reward can systematically alter motivational state (Zhang et al.,

2018). But neurophysiological methods would be required to identify and partition sources of noise internal to the nervous system that may arise at various stages of the visual processing and perceptual decision making pipeline. Paradigms that blend the advantages of the current approach for partitioning stimulus-based variation with neurophysiological and computational methods for partitioning noise would be a useful way forward (Charles & Pillow, 2018; Goris et al., 2014; Ziemba et al., 2016).

4.5.4. External limits to human performance

Broadly construed, the current work continues in the tradition of the classic 1942 study of Hecht, Shlaer, and Pirenne. Its two most widely appreciated results are that, when fully dark adapted, i) the absorption of a single photon reliably elicits a response from a rod photoreceptor and ii) the absorption of five to seven photons in a short period of time reliably causes a reportable sensation of the light. Slightly less widely appreciated is the finding that the limits of the human ability to detect light (i.e. light detectability thresholds) are attributable to the stochastic nature of light itself, a performance-limiting factor that is external to the organism. On a given trial at a given stimulus intensity, whether or not subjects reported that they had seen the stimulus depended near-exclusively on whether or not the requisite number of photons had been absorbed. That is, if the numbers of photons in proximal stimulus was identical, humans would respond identically. Performance was thus very tightly yoked to the variability of the external stimulus.

The results the current study suggest that, just as rod photoreceptors support performance in a very similar manner across different human observers, the computational mechanisms supporting the estimation and discrimination of depth in natural scenes are very similar across observers. In the current study, we showed that stimulus-based limits to performance become increasingly important as stimuli become ever more natural. If this pattern holds, it may that stimulus-based limits to performance are by far the dominant factor as organisms engage with the natural environment. If true, image-computable models will have the potential to achieve remarkable predictive power from analysis of the stimulus alone. Such models, in which the underlying computations are made explicit, would have tremendous practical applications and deepen our understanding of how vision

works in the real world.

CHAPTER 5

SUMMARY OF CONTRIBUTIONS

5.1. Contribution of Chapter 1

The primary goals of vision science include both a detailed understanding of *what* is perceived, but also *how* this is accomplished. Theories proposed by Marr and Poggio popularized a philosophy and psychophysical approach directed towards uncovering the algorithmic structure of the visual system. Many aspects of their theory have received criticism based upon empirical findings, resulting in important changes in how algorithm-oriented psychophysical investigations are approached. There is one remaining aspect of these approaches that requires further scrutiny though: the view that visual cues are able to be distilled into fundamental, independent units of computation. Contemporary neuroscience continues to reveal an increasingly complex process of visual computation, where holism and contextualization are paramount, suggesting the idealized unitary view of visual processing seems increasingly less likely. This suggests that an alternative approach to algorithm-oriented psychophysics be taken into consideration, one which is indifferent to idealized notions of representation and views visual processing in the context of its complex and holistic nature.

5.2. Contribution of Chapter 2

Approaches that seek to uncover algorithms of perception, face the challenge of investigating highly complex systems. Model fitting used by traditional methods provides accurate descriptions of the data they generate but are insufficient to describe underlying algorithms; many algorithms are able to describe the same performance data, even when performance is measured in response to complex images. Alternatively, algorithmic theory provides a better route for the investigation of algorithms, where algorithms are themselves assessed, rather than statistics regarding their performance. Here, data generated by the model and the observer are used as incomplete descriptions of their respective systems, which fully characterize the algorithms of their system, as their descriptions approach infinity. Therefore, in the noiseless case each individual data is used as a model testing criteria rather than expected values across large groupings of data. For noisy systems such as human vision, model

assessment is based upon criteria that assesses similarity between observer and model responses while accounting for noise. Unique and informative criteria—i.e. diverse stimuli and a sufficiently complex set of stimuli—can help identify models closer to the true underlying algorithm by ruling out classes of algorithms. Repeated criteria testing can help reduce uncertainty of an individual test due to noise. Importantly, this approach relies on balancing algorithm space exploration and hedging against certainty by finding an appropriate medium between stimulus diversity and similarity. Natural images and multiple pass procedures provide means of balancing this tradeoff. The diversity of natural images, in both content and criteria, are an excellent substrate for algorithm testing criteria. Multiple pass procedures coupled with correlation methods are a viable way of simultaneously inferring noise impact and taking full advantage of testing criteria.

5.3. Contribution of Chapter 3

Natural images are highly relevant to understanding how vision works in the real world—they are the images that the visual system evolved to process (Felsen & Dan, 2005). Current definitions of natural images are either vague or unuseful in the context of human vision. A more flexible and relevant definition is graded on how likely they are to be generated from the real world, and are robust to cases where the degree of naturalness of an image is not absolute. In context of psychophysical investigation, these definitions allow the contextualization of various types of stimuli—artificial, naturalistic, etc.—and the investigations in which they were used in terms of their relevancy.

A key feature of natural images is their variability, which is derived from the variability of scene content. The limitations of what can be represented in a natural image, coupled with the high degree of variability, make inferences about the scene particularly difficult. The ability of the visual system to deal with natural image variability and the complexity it imposes should be considered its primary feature.

A difficulty with using natural images in psychophysical tasks is presenting them in a controlled manner without the loss of natural content and context. Fully naturalistic approaches are extremely burdensome and produce data whose variability is hard to account for, whereas traditional psychophysics approaches are highly unnatural, but produce highly interpretable data. Hybrid,

naturalistic approaches offer a fair medium.

Previous investigations have used naturalistic images in psychophysics experiments under algorithm-oriented approaches. Some of these investigations have used a highly diverse set of naturalistic stimuli. One such study also incorporated multiple-pass procedures that meet the experimental recommendations proposed in Chapter 2 (Burge & Geisler, 2014). While their analysis provided stronger support for their model than what is typically done, their methods of analysis were limited, thus unable to fully utilize the data available for more robust model testing criteria.

5.4. Contribution of Chapter 4

Chapter 2 recommends an approach for algorithm-oriented psychophysical investigation of the visual system, which includes the use of naturalistic images and their diversity and multiple-pass procedures. Chapter 4 consists of a psychophysics-based investigation, conducted by the author. This study acts as a proof of concept for the the approach recommended in Chapter 2, but also provides a rich picture of stimulus factors that contribute to human perceptual performance in natural scenes.

5.4.1. Contribution to behavioral visual science

This study investigated stereo-depth discrimination performance to natural image patches, and is the first study to use stereo-based photography with naturally derived depth structure to probe stereo-depth perception. This study involved two human stereo-depth discrimination experiments with a comprehensive dataset of 20,000 trials for each of the three observers. The methodological rigor and the large volume of data collected ensure that the findings are robust and reliable.

Vision science has a great understanding of how simple stimuli are perceived, but a limited understanding of natural and naturalistic stimuli. This study replicates a performance pattern from simple-stimulus based studies found in the classic literature: discrimination thresholds increase exponentially as targets move farther in depth from fixation.

In the real world, perceptual and behavioral variability is driven by both external and internal factors. Internal-noise, while interesting in its own right, confounds the investigation of stimulus-

driven effects—the component attributable to deterministic aspects of visual computation. Previous work has partitioned performance into internal-noise-driven and stimulus-driven components (Chin, n.d.). Here, we do the same but in context of stereo-depth discrimination performance. Our findings indicate that these natural-stimulus-based sources of variability are significant determinants of perceptual performance limits in natural environments. We show that performance limits are increasingly attributable to stimulus variability (rather than internal noise) as the stimuli used to probe performance become more natural.

A comprehensive account of perceptual and behavioral variability, must identify and describe the impact of all stimulus-driven sources of performance-limiting variability. How do these sources contribute? We develop methods that allow the further partitioning of the stimulus-driven component into its constituent components. Through this method, we show that two distinct types of natural-stimulus variability—luminance-pattern variation and local-depth variation—have distinct and largely separable effects on human performance.

Internal noise not only confounds analysis of stimulus-driven effects, but it also prevents investigation into how performance between observers differs. This work provides methods that allow similarity in observer response to be estimated while accounting for internal noise. Analysis generated through application of this method revealed observer responses to be highly consistent. These results suggest that human perceptual systems use similar computational mechanisms when processing natural stimuli, highlighting a commonality in perceptual processing. Here, we additionally find that as stimulus variation becomes more severe, the absolute impact of that stimulus-by-stimulus variation on performance becomes more severe and also becomes more uniform across human observers. Encouragingly, these results raise the prospect that an appropriate image-computable model may in principle be able to predict a substantial proportion of stimulus-by-stimulus variation across natural images and observers.

5.4.2. Contribution to algorithmic perceptual science

Previous work has lacked the methods required to make strong model assessment of visual models in terms of its algorithmic structure. Two factors make this particularly difficult: internal noise

and between-observer differences. Model-observer comparisons need to control for noise but also contextualize between-observer differences. This study provides means of making algorithm-based assessments by taking these into consideration.

While this study develops and provides means for model-comparison, it is not made explicitly clear how this is accomplished. This is described in short here. In order to make a model comparison with human data, observer-differences are treated as a random variable. A model that captures the algorithmic structure of the human based algorithm will have its model-human partial-correlation values statistically indistinguishable from human-human partial-correlations. Thus an appropriate model comparison will test this hypothesis.

The methods developed in this work also provide means for more coarsely grained analyses based in terms of performance. In particular three levels of analyses can be made, each progressively more informative than the previous: general performance, externally-driven based performance, and specific externally-driven based impact. Model comparisons in terms of these analyses are likely to provide a much needed benchmark when models do not meet the higher algorithmic-based testing criteria.

BIBLIOGRAPHY

Anderson, J. R. (2013, January 11). *The Adaptive Character of Thought* (1st ed.). Psychology Press. https://doi.org/10.4324/9780203771730

Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., & Bressloff, P. C. (2017). Circuits and Mechanisms for Surround Modulation in Visual Cortex. *Annual Review of Neuroscience*, *40*(1), 425–451. https://doi.org/10.1146/annurev-neuro-072116-031418

Antonopoulos, C., Fokas, A., & Bountis, T. (2016). Dynamical complexity in the C.elegans neural network. *The European Physical Journal Special Topics*, *225*(6-7), 1255–1269. https://doi.org/10.1140/epjst/e2016-02670-3

Badcock, D. R., & Schor, C. M. (1985). Depth-increment detection function for individual spatial channels. *Journal of the Optical Society of America A*, *2*(7), 1211. https://doi.org/10.1364/JOSAA.2.001211

Baddeley, R. (1997). The Correlational Structure of Natural Images and the Calibration of Spatial Representations. *Cognitive Science*, *21*(3), 351–372. https://doi.org/10.1207/s15516709cog2103_4

Banks, M. S. (2004). Why Is Spatial Stereoresolution So Low? *Journal of Neuroscience*, *24*(9), 2077–2089. https://doi.org/10.1523/JNEUROSCI.3852-02.2004

Blakemore, C. (1970). The range and scope of binocular depth discrimination in man. *The Journal of Physiology*, *211*(3), 599–622. https://doi.org/10.1113/jphysiol.1970.sp009296

Bonnen, K., Burge, J., Yates, J., Pillow, J., & Cormack, L. K. (2015). Continuous psychophysics: Target-tracking to measure visual sensitivity. *Journal of Vision*, *15*(3), 14. https://doi.org/10.1167/15.3.14

Bowers, N. R., Gibaldi, A., Alexander, E., Banks, M. S., & Roorda, A. (2019). High-resolution eye tracking using scanning laser ophthalmoscopy. *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 1–3. https://doi.org/10.1145/3314111.3322877

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Brainard, D. H., Cottaris, N. P., & Radonjić, A. (2018). The perception of colour and material in naturalistic tasks. *Interface Focus*, *8*(4), 20180012. https://doi.org/10.1098/rsfs.2018.0012

Burge, J., & Geisler, W. S. (2011). Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences*, *108*(40), 16849–16854. https://doi.org/10.1073/pnas.1108491108

Burge, J., & Geisler, W. S. (2014). Optimal disparity estimation in natural stereo images. *Journal of Vision*, *14*(2), 1–1. https://doi.org/10.1167/14.2.1

Burge, J. (2020). Image-Computable Ideal Observers for Tasks with Natural Stimuli. *Annual Review of Vision Science*, *6*(1), 491–517. https://doi.org/10.1146/annurev-vision-030320-041134

Burge, J., & Cormack, L. K. (2020, August 6). *Target tracking reveals the time course of visual processing with millisecond-scale precision* (preprint). Neuroscience. https://doi.org/10.1101/2020.08.05.238642

Burge, J., & Geisler, W. S. (2012, January 22). Optimal defocus estimates from individual images for autofocusing a digital camera. In S. Battiato, B. G. Rodricks, N. Sampat, F. H. Imai, & F. Xiao (Eds.). https://doi.org/10.1117/12.912066

Burge, J., & Geisler, W. S. (2015). Optimal speed estimation in natural image movies predicts human performance. *Nature Communications*, *6*(1), 7900. https://doi.org/10.1038/ncomms8900

Burge, J., McCann, B. C., & Geisler, W. S. (2016). Estimating 3D tilt from local image cues in natural scenes. *Journal of Vision*, *16*(13), 2. https://doi.org/10.1167/16.13.2

Burgess, A. E., & Colborne, B. (1988). Visual signal detection IV Observer inconsistency. *Journal of the Optical Society of America A*, *5*(4), 617. https://doi.org/10.1364/JOSAA.5.000617

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of Human Visual Signal Discrimination. *Science*, *214*(4516), 93–94. https://doi.org/10.1126/science.7280685

Carlsson, G., Ishkhanov, T., De Silva, V., & Zomorodian, A. (2008). On the Local Behavior of Spaces of Natural Images. *International Journal of Computer Vision*, *76*(1), 1–12. https://doi.org/10.1007/s11263-007-0056-x

Carrasco, M. (2018). How visual spatial attention alters perception. *Cognitive Processing*, *19*(S1), 77–88. https://doi.org/10.1007/s10339-018-0883-4

Charles, A., & Pillow, J. (2018). Additive Continuous-time Joint Partitioning of Neural Variability. *2018 Conference on Cognitive Computational Neuroscience*. https://doi.org/10.32470/CCN.2018.1087-0

Chin, B. M., & Burge, J. (2020). Predicting the Partition of Behavioral Variability in Speed Perception with Naturalistic Stimuli. *The Journal of Neuroscience*, *40*(4), 864–879. https://doi.org/10.1523/JNEUROSCI.1904-19.2019

Chin, B. M., & Burge, J. (2022). Perceptual consequences of interocular differences in the duration of temporal integration. *Journal of Vision*, *22*(12), 12. https://doi.org/10.1167/jov.22.12.12

Chin, B. M. (n.d.). Computational mechanisms underlying perception of visual motion.

Cormack, L. K., Stevenson, S. B., & Schor, C. M. (1991). Interocular correlation, luminance contrast and cyclopean processing. *Vision Research*, *31*(12), 2195–2207. https://doi.org/10.1016/0042-6989(91)90172-2

Deco, G., & Rolls, E. T. (2005). Attention, short-term memory, and action selection: A unifying theory. *Progress in Neurobiology*, *76*(4), 236–256. https://doi.org/10.1016/j.pneurobio.2005.08.004

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience*, *8*(12), 1643–1646. https://doi.org/10.1038/nn1608

Frazor, R. A., & Geisler, W. S. (2006). Local luminance and contrast in natural images. *Vision Research*, *46*(10), 1585–1598. https://doi.org/10.1016/j.visres.2005.06.038

Geirhos, R., Meding, K., & Wichmann, F. A. (n.d.). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency.

Geisler, W. S., & Perry, J. S. (2011). Statistics for optimal point prediction in natural images. *Journal of Vision*, *11*(12), 14–14. https://doi.org/10.1167/11.12.14

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, *96*(2), 267–314. https://doi.org/10.1037/0033-295X.96.2.267

Geisler, W. S. (2003, November 21). Ideal Observer Analysis. In L. M. Chalupa & J. S. Werner (Eds.), *The Visual Neurosciences, 2-vol. set* (pp. 825–837). The MIT Press. https://doi.org/10.7551/mitpress/7131.003.0061

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*(7), 771–781. https://doi.org/10.1016/j.visres.2010.09.027

Geisler, W. S. (2018). Psychometric functions of uncertain template matching observers. *Journal of Vision*, *18*(2), 1. https://doi.org/10.1167/18.2.1

Geisler, W., Perry, J., Super, B., & Gallogly, D. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*(6), 711–724. https://doi.org/10.1016/S0042-6989(00)00277-7

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, *38–38*(1), 173–198. https://doi.org/10.1007/BF01700692

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature*, *402*(6758), 176–178. https://doi.org/10.1038/46027

Gold, J. I., & Ding, L. (2013). How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology*, *103*, 98–114. https://doi.org/10.1016/j.pneurobio.2012.05.008

Goris, R. L. T., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, *17*(6), 858–865. https://doi.org/10.1038/nn.3711

Hecht, S. (1942). Energy, quanta, and vision. *The Journal of General Physiology*, *25*(6), 819–840. https://doi.org/10.1085/jgp.25.6.819

Hegde, J. (2008). Time course of visual perception: Coarse-to-fine processing and beyond. *Progress in Neurobiology*, *84*(4), 405–439. https://doi.org/10.1016/j.pneurobio.2007.09.001

Held, R. T., & Banks, M. S. (n.d.). Misperceptions in Stereoscopic Displays: A Vision Science Perspective, 10.

Iyer, A., & Burge, J. (2019). The statistics of how natural images drive the responses of neurons. *Journal of Vision*, *19*(13), 4. https://doi.org/10.1167/19.13.4

Iyer, A. V., & Burge, J. (2018). Depth variation and stereo processing tasks in natural scenes. *Journal of Vision*, *18*(6), 4–4. https://doi.org/10.1167/18.6.4

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision*, *3*(7), 4. https://doi.org/10.1167/3.7.4

Kant, I. (2009). *The critique of pure reason* (15th printing). Cambridge University Press.

Kim, S., & Burge, J. (2018). The lawful imprecision of human surface tilt estimation in natural scenes (J. L. Gallant, Ed.). *eLife*, *7*, e31448. https://doi.org/10.7554/eLife.31448

Kreiman, G., & Serre, T. (2020). Beyond the feedforward sweep: Feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, *1464*(1), 222–241. https://doi.org/10.1111/nyas.14320

Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica*, *43*(3), 199–224. https://doi.org/10.1016/0001-6918(79)90026-X

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America*, *70*(12), 1458. https://doi.org/10.1364/JOSA.70.001458

Lu, Z.-L., & Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research*, *38*(9), 1183–1198. https://doi.org/10.1016/S0042-6989(97)00273-3

MacLean, K. A., Aichele, S. R., Bridwell, D. A., Mangun, G. R., Wojciulik, E., & Saron, C. D. (2009). Interactions between endogenous and exogenous attention during vigilance. *Attention, Perception, & Psychophysics*, *71*(5), 1042–1058. https://doi.org/10.3758/APP.71.5.1042

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*(1), 147–155. https://doi.org/10.1017/S0952523808080905

Mante, V., Frazor, R. A., Bonin, V., Geisler, W. S., & Carandini, M. (2005). Independence of luminance and contrast in natural scenes and in the early visual system. *Nature Neuroscience*, *8*(12), 1690–1697. https://doi.org/10.1038/nn1556

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press. https://doi.org/10.7551/mitpress/9780262514620.001.0001

Marr, D., & Poggio, T. (1976). From understanding computation to understanding neural circuitry.

McKee, S. P., Levi, D. M., & Bowne, S. F. (1990). The imprecision of stereopsis. *Vision Research*, *30*(11), 1763–1779. https://doi.org/10.1016/0042-6989(90)90158-H

Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2009). Spatial Attention Decorrelates Intrinsic Activity Fluctuations in Macaque Area V4. *Neuron*, *63*(6), 879–888. https://doi.org/10.1016/j.neuron.2009.09.013

Mueller, C. G. (1951). Frequency of seeing functions for intensity discrimination at various levels of adapting intensity. *Journal of General Physiology*, *34*(4), 463–474. https://doi.org/10.1085/jgp.34.4.463

Nachmias, J., & Sansbury, R. V. (1974). Grating contrast: Discrimination may be better than detection. *Vision Research*, *14*(10), 1039–1042. https://doi.org/10.1016/0042-6989(74)90175-8

Nakayama, K., Moher, J., & Song, J.-H. (2022). Rethinking Vision and Action.

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research*, *46*(16), 2465–2474. https://doi.org/10.1016/j.visres.2006.02.002

Nienborg, H., Bridge, H., Parker, A. J., & Cumming, B. G. (2004). Receptive Field Size in V1 Neurons Limits Acuity for Perceiving Disparity Modulation. *The Journal of Neuroscience*, *24*(9), 2065–2076. https://doi.org/10.1523/JNEUROSCI.3887-03.2004

Pashler, H., Johnston, J. C., & Ruthruff, E. (2000). ATTENTION AND PERFORMANCE.

Peebles, D., & Cooper, R. P. (2015). Thirty Years After Marr's *Vision* : Levels of Analysis in Cognitive Science. *Topics in Cognitive Science*, *7*(2), 187–190. https://doi.org/10.1111/tops.12137

Pillow, J. W. (2024). Cross Talk opposing view: Marr's three levels of analysis are not useful as a framework for neuroscience. *The Journal of Physiology*, *602*(9), 1915–1917. https://doi.org/10.1113/JP279550

Radonjic, A., Cottaris, N. P., & Brainard, D. H. (2015). Color constancy in a naturalistic, goal-directed task. *Journal of Vision*, *15*(13), 3. https://doi.org/10.1167/15.13.3

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471. https://doi.org/10.1016/0005-1098(78)90005-5

Rissanen, J. (2007). *Information and complexity in statistical modeling*. Springer.

Rucci, M., & Poletti, M. (2015). Control and Functions of Fixational Eye Movements. *Annual Review of Vision Science*, *1*(1), 499–518. https://doi.org/10.1146/annurev-vision-082114-035742

Sebastian, S., Abrams, J., & Geisler, W. S. (2017). Constrained sampling experiments reveal principles of detection in natural scenes. *Proceedings of the National Academy of Sciences*, *114*(28), E5731–E5740. https://doi.org/10.1073/pnas.1619487114

Sebastian, S., Burge, J., & Geisler, W. S. (2015). Defocus blur discrimination in natural images with natural optics. *Journal of Vision*, *15*(5). https://doi.org/10.1167/15.5.16

Seilheimer, R. L., Rosenberg, A., & Angelaki, D. E. (2014). Models and processes of multisensory cue combination. *Current Opinion in Neurobiology*, *25*, 38–46. https://doi.org/10.1016/j.conb.2013.11.008

Siu, C., Balsor, J., Merlin, S., Federer, F., & Angelucci, A. (2021). A direct interareal feedback-to-feedforward circuit in primate visual cortex. *Nature Communications*, *12*(1), 4911. https://doi.org/10.1038/s41467-021-24928-6

Soto, D., Greene, C. M., Kiyonaga, A., Rosenthal, C. R., & Egner, T. (2012). A Parieto-Medial Temporal Pathway for the Strategic Control over Working Memory Biases in Human Visual Attention. *The Journal of Neuroscience*, *32*(49), 17563–17571. https://doi.org/10.1523/JNEUROSCI.2647-12.2012

Stevenson, S., Sheehy, C., & Roorda, A. (2016). Binocular eye tracking with the Tracking Scanning Laser Ophthalmoscope. *Vision Research*, *118*, 98–104. https://doi.org/10.1016/j.visres.2015.01.019

Stevenson, S. B., Cormack, L. K., Schor, C. M., & Tyler, C. W. (1992). Disparity tuning in mechanisms of human stereopsis. *Vision Research*, *32*(9), 1685–1694. https://doi.org/10.1016/0042-6989(92)90161-B

Tkačik, G., Garrigan, P., Ratliff, C., Milčinski, G., Klein, J. M., Seyfarth, L. H., Sterling, P., Brainard, D. H., & Balasubramanian, V. (2011). Natural Images from the Birthplace of the Human Eye (D. C. Burr, Ed.). *PLoS ONE*, *6*(6), e20409. https://doi.org/10.1371/journal.pone.0020409

Tolhurst, D., Movshon, J., & Dean, A. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, *23*(8), 775–785. https://doi.org/10.1016/0042-6989(83)90200-6

Tomko, G. J., & Crapper, D. R. (1974). Neuronal variability: Non-stationary responses to identical visual stimuli. *Brain Research*, *79*(3), 405–418. https://doi.org/10.1016/0006-8993(74)90438-7

Turner, M. H., Sanchez Giraldo, L. G., Schwartz, O., & Rieke, F. (2019). Stimulus- and goal-oriented frameworks for understanding natural vision. *Nature Neuroscience*, *22*(1), 15–24. https://doi.org/10.1038/s41593-018-0284-0

Tyler, C. W. (1974). Depth perception in disparity gratings. *Nature*, *251*(5471), 140–142. https://doi.org/10.1038/251140a0

Ukwade, M. T., Bedell, H. E., & Harwerth, R. S. (2003). Stereothresholds with simulated vergence variability and constant error. *Vision Research*, *43*(2), 195–204. https://doi.org/10.1016/S0042-6989(02)00409-1

Warren, W. H. (2012). Does This Computational Theory Solve the Right Problem? Marr, Gibson, and the Goal of Vision. *Perception*, *41*(9), 1053–1060. https://doi.org/10.1068/p7327

Westheimer, G. (1979). Cooperative neural processes involved in stereoscopic acuity. *Experimental Brain Research*, *36*(3). https://doi.org/10.1007/BF00238525

Yates, J. L., Coop, S. H., Sarch, G. H., Wu, R.-J., Butts, D. A., Rucci, M., & Mitchell, J. F. (2023). Detailed characterization of neural selectivity in free viewing primates. *Nature Communications*, *14*(1), 3656. https://doi.org/10.1038/s41467-023-38564-9

Zhang, P., Hou, F., Yan, F.-F., Xi, J., Lin, B.-R., Zhao, J., Yang, J., Chen, G., Zhang, M.-Y., He, Q., Dosher, B. A., Lu, Z.-L., & Huang, C.-B. (2018). High reward enhances perceptual learning. *Journal of Vision*, *18*(8), 11. https://doi.org/10.1167/18.8.11

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences*, *113*(22), E3140–E3149. https://doi.org/10.1073/pnas.1510847113