BAYESIAN CATEGORIZATION IN TOUCH

EXPLORATION OF A BAYESIAN PROBABILISTIC MODEL FOR CATEGORIZATION IN THE SENSE OF TOUCH

By KYRA ALICE GAUDER, HONS. B.SC.

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

McMaster University
DOCTOR OF PHILOSOPHY (2024)
Hamilton, Ontario (Psychology)

TITLE: Exploration of a Bayesian probabilistic model for categorization in the sense of touch
AUTHOR: Kyra Alice GAUDER , Hons. B.Sc. (McMaster University)
SUPERVISOR: Dr. Daniel GOLDREICH
NUMBER OF PAGES: xiv, 140

Lay Abstract

The process of categorization is an essential part of our daily life as we encounter various things in the world. Here we explore a model that attempts to explain this process. This model is derived using Bayesian inference and was applied to human behavioural data in a categorization task. We found that the model accounted for most of the performance of our participants but consistently outperformed them. We conducted simulations to explore and demonstrate that this difference is primarily due to the presence of sensory noise in participants. Once we accounted for this noise, we found that our model predicted human performance even more accurately. The work in this thesis demonstrates that a Bayesian Categorization Model which accounts for sensory noise is a good fit and predictor for human performance on categorization tasks.

Abstract

Categorization is a complex decision-making process that requires observers to collect information about stimuli using their senses. While research on visual or auditory categorization is extensive, there has been little attention given to tactile categorization. Here we developed a paradigm for studying tactile categorization using 3D-printed objects. Furthermore, we derived a categorization model using Bayesian inference and tested its performance against human participants in our categorization task. This model accurately predicted participant performance in our task but consistently outperformed them, even after extending the learning period for our participants. Through theoretical exploration and simulations, we demonstrated that the presence of sensory measurement noise could account for this performance gap, which we determined was a present factor in participants undergoing our task through a follow-up experiment. Including measurement noise led to a better-fitting model that was able to match the performance of our participants much more closely. Overall, the work in this thesis provides evidence for the efficacy of a tactile categorization experimental paradigm, demonstrates that a Bayesian model is a good fit and predictor for human categorization performance, and underscores the importance of accounting for sensory measurement noise in categorization models.

Acknowledgements

First, I would like to extend a huge thank you to my graduate supervisor Dr. Daniel Goldreich. The mentorship and support I received under his supervision was integral to my learning and growth as a scientist and I would not have made it as far as I did without his guidance. He introduced me to the world of Bayesian inference and his passion inspired me to involve myself academically in pursuit of understanding that in any way I could. Thank you for your guidance over these many years.

Thank you as well to the members of my supervisory committee, Dr. Susan Becker and Dr. Deda Gillespie. Your support was invaluable and it was always reassuring knowing that I had your supervision behind me. I am especially grateful to Deda for joining my committee later in my graduate career.

Thank you to all the people in my family who have been there for me and encouraged me throughout my journey. To my sisters, Teesha and Aneesa Munoz, and my grandparents, for your support and love throughout my life. Thank you to my father, Sean Gauder. I am forever grateful for everything he has done for me in helping me get to this point in life. Without his help and encouragement, I would not have made it through University nor my graduate program. He inspired my initial interest in Psychology and curiosity for understanding the human brain. And thank you to my late grandfather, Sergio Munoz, who I wish I could have been able to share the end of this journey with but know how proud he was of me all the same.

Thank you to everyone in my life who supported my coming out as a transgender woman. Coming out was among the hardest and scariest processes I've ever had to go through. There is always a deep-seated fear that revealing who you are in that way can lead to catastrophic outcomes. I am so grateful for everyone who accepted who I am and that I can submit this thesis as my true self.

Most of all, I would like to thank and dedicate this dissertation to my wife, Maxine Susan Sky Farrell. Skye came into my life when I was struggling the most and provided nothing but love and support to help lift me back up. Thank you for always being there, for your endless kindness and understanding and patience, and for sharing your lives with me. This wouldn't have been possible without you.

Table of Contents

List of Figures

List of Tables

List of Terms, Abbreviations, and Symbols

Term	Description
Observer	General term for any agent that performs a task given available information.
Participant	Human observer who undergoes our experiments.
Simulant	Simulated observer which can be virtually run through a task by responding according to some model as a stand-in for a human observer.
Proximal stimulus	Internal signal from the sensory receptors, such as the visual image on the retina or the sensation along the skin.
Distal stimulus	External object/stimulus which cast the signal onto a sensory system.
Percept	Holistic perception of proximal stimulus, the instantiation of the perceptual process that is held in awareness through attention.
Model	Hypothesized set of computations that may represent the perceptual process which gives rise to an observer's percept and their resulting behaviour. Every Bayesian model consists of an Encoder and Decoder module.

Continued on next page

Table 1 – continued from previous page

Encoder Module	Hypothesized process by which measurement data are generated as vector $\mathbf{f}_m$. The Encoder modules encapsulate the process by which the object is chosen and the process by which the observer obtains a sensory measurement from the object. Also known as the Generative Model or the Forward Model.
Decoder Module	Hypothesized process by which the observer infers the category C given their measured data vector f_m.
Category (C)	Distribution of objects g within the feature distribution space. Different categories are often denoted as either Category A or Category B throughout this work.
Stimulus Object (g)	Physical object from our stimulus set used throughout experiments. Each object g has physical properties or features f associated with it. The vector of all features f associated with object g is $\mathbf{f}_g$.
Feature (f)	Physical property or feature of the object (ie, number of dots or number of sides)
Number of Sides (n)	Physical feature dimension, the number of sides found on our stimulus object. Proportional to the length of each edge.
Dot spacing (d)	Physical feature dimension, the spacing between the raised hemispherical dots found on the surface of our stimulus object. Proportional to the number of dots or the density of dots.
σ	Standard Deviation
μ	Mean

Continued on next page

Table 1 – continued from previous page

Measurement $(f_m; n_m, d_m)$	Internal sensory measurement f_m of the physical property, what is sensed from the feature f. The internal measurement is a random sample from a Gaussian distribution with mean f and standard deviation σ_{fm}. $\mathbf{f}_m$ (bolded **f**) refers to a vector containing both measurements n_m and d_m, which is associated with a particular object $\mathbf{f}_g$.
Perceived measurement $(\mathbf{f}_{mk})$	Hypothesized measurement perceived by an observer's sensory system contaminated by internal sensory noise, where k denotes each measurement bin.
Measurement Noise $(\sigma_{f_m}; \sigma_{n_m}, \sigma_{d_m})$	Variation of measurement f_m. The observer's internal measurement error σ_{fm}, often due to internal sensory neural noise.
Within-Category Feature Distribution $P(f \mid \mu_{Cf}, \sigma_{Cf})$	Statistical distribution of object features within a category.
Within-Category Feature Mean $(\mu_{Cf}; \mu_{Af}, \mu_{Bf})$	Average stimulus object within a particular category.
Within-Category Feature Variability $(\sigma_{Cf}; \sigma_{Af}, \sigma_{Bf})$	Variation of frequency of objects within a given category.
Category Posterior $P(C \mid f_m)$	Observer's belief that, upon sensing an object, it belongs to a particular category.
Category Prior $P(C)$	Prevalence of a category, typically assumed to be 50%.
Category Likelihood $P(f_m \mid C)$	Probability of the vector $\mathbf{f}_m$ within a category distribution.
MAP estimate	*maximum a posteriori* estimate
Exemplar Theory	Classic model for categorization in which each instance of a category is stored and referred back to upon recall.

Continued on next page

Table 1 – continued from previous page

Prototype Theory	Classical model for categorization in which all instances of a category are averaged together into a single example.
2IFC	Two-Interval Forced-Choice; a methodological design where participants are presented with two objects and must make a single choice between them.
SDT	Signal Detection Theory; a general framework for modelling an observer's ability to detect the presence (or absence) of sensory input, especially in the presence of uncertainty or noise.
RMC	Rational Model of Categorization
ZOL	Zero-one loss
left-DLPFC	left-Dorsolateral Prefrontal Cortex
AAC	Anterior Cingulate Cortex
TPC	Temporal-Parietal Cortex
LIP	Lateral Intraparietal Cortex
MT	Middle Temporal Visual Area

Declaration of Academic Achievement

I, Kyra Alice GAUDER, declare that this thesis titled, “Exploration of a Bayesian probabilistic model for categorization in the sense of touch” and the work presented in it are my own. I confirm that:

Chapter 2

I was involved in all aspects of the theoretical derivation of our computational Bayesian Categorization model and the programming related to it. My graduate supervisor Dr. Daniel Goldreich made equally major contributions to this theoretical work and the programming related to it. I was also involved in all aspects of the experimental design established that was used throughout the rest of the thesis includion the creation and 3D printing of the stimuli and programming to run the experiment. My graduate supervisor Dr. Daniel Goldreich as well as Dr. Keon Allen (who was an undergraduate student at the time) provided major contributions towards the experimental design, creation of the stimuli, and programming.

Chapter 3

I was involved in all aspects of the empirical research related to Experiments 1 & 2. This includes the experimental design, programming of experimental software, data collection and analysis. Dr. Daniel Goldreich made major contributions to the experimental design and analysis. Many undergraduate students assisted in the development, data collection and analysis throughout Experiments 1 & 2: Patrick Dans, Salina Mathur, Dezi Ahuja, Behrad Dehnadi, Kritika Soin, Rafi Matin, Shekina Remigio, and Selina Bains.

Chapter 4

I was involved in all aspects of the theoretical work involved in our experiment simulations and analysis. Dr. Daniel Goldreich made major contributions to this theoretical work and the programming related to it. I was also involved in all aspects of the empirical research related to Experiment 3. This includes the experimental design, programming of experimental software, data collection and analysis. Dr. Daniel Goldreich made major contributions to the experimental design and analysis. Two undergraduate students assisted in the development, data collection and analysis of this experiment: Alex Kappen and Grace Arthur.

Chapter 1

General Introduction

1.1 Overview of Categorization

Categorization is a computational problem comprised of two steps: data acquisition and decision-making. Observers acquire information as they move throughout the world and quickly and accurately come to conclusions about the various stimuli they encounter. Often this requires the combination of many features. It is not enough for an observer holding a rock to independently observe that the object is grey, hard, and cool to the touch to determine that it is a rock; any number of other objects may be grey, hard, or a similar temperature. But by combining these features into a holistic percept, and then comparing that percept to previous knowledge of what they know about other rocks—a rock category—an observer can decide whether the object they're holding is, in fact, a rock.

Categories are found throughout daily life. Any object or pattern recognition task could be understood as a categorization problem. Imagine reaching into a bag of the many different polyhedral dice commonly used throughout tabletop role-playing games to pull out a favourite twenty-sided die. The dice might have different numbers of sides and different sets of dice may have different weights and textures. There are many features of data available to categorize and identify the sought-after die. Or perhaps you find yourself eating a bag of nuts with both pecans and walnuts mixed in together and would like to only eat the pecans. Pecans and walnuts can be measured by their size, texture, and knobbliness. While the average pecan is quite different from the average walnut, these features can vary greatly due to natural variations in how these nuts grow. Examining only the size feature to categorize pecans from walnuts may lead to mistaken classifications; while the average pecan is smaller than a walnut, individual pecans may be larger and individual walnuts may be smaller. From identifying bird species by bird calls out of a cacophony of sounds to classifying galaxy shapes from deep space imagery of the night sky, understanding the underlying processes for how categories are formed and how categorizations are made is an important pursuit in understanding information processing.

A simple model for object categorization could be that the brain stores information about every instance of an object it has encountered. A rock category could be a collection of memories of every previous rock an observer has seen. This *Exemplar Theory* for categorization would then require the observer to compare the object they are holding against every instance stored in each known category for similarity; the category with the highest similarity would be the best classification for the object (Nosofsky, 1986). While Exemplar models are mathematically tractable, they pose great difficulties in biological implementation. It would be very costly and unsustainable for the brain to not only store every instance of an object it perceives but then further compare new objects against every stored exemplar to make a decision. The processing required to perform this classification task would grow as more exemplars are added to a category and as more categories are formed.

An earlier seminal model for categorization was *Prototype Theory*, first explored by Posner & Keele (1968). Prototype models propose that categories are abstractions of the objects they represent. Rather than storing all exemplars of a rock, the brain might summarize or extract from the collection of all rock instances a representative of a “prototypical rock”. The classification process would then only require comparing a new object against that rock prototype to determine similarity.

The idea of objects being represented by a more generalized prototype can be traced back many centuries to Plato’s Theory of Forms, making this theory of categorization one of the oldest developed. The Theory of Forms posited that for any conceivable thing you can observe, whether it’s a rock, a chair, or a house, there exists an abstracted perfect Form of that thing. In trying to create these “things” through art or material creation, people are effectively recreating imperfect approximations of their Forms. Conceptually, this bears similarity to *Prototype Theory* in that the brain may have these abstracted representations of categories, and when either trying to identify an object as those categories or create something like those categories, we are referring to that internal representation.

Prototype Theory, however, is not without its faults as well, and it has been contested over the years alongside competing Exemplar theories for categorization. A large point of contention for *Prototype Theory* is that it requires forgetting large swaths of information and detail by averaging everything into a single prototype. Brains are capable of holding specific object information as well as categorical information, which goes against the notion of pure prototypes. The process of summarizing all instances into an abstracted average also poses issues when dealing with a category containing high variability. Rocks do not all look alike and vary wildly in shape, size, sharpness, and colour. If you were to average all these together, you might end up with a smooth, rounded sphere—something that looks very different from the rock exemplars that went into it.

Many models have spawned from the competing Exemplar and Prototype models for categorization, with both types finding success in various areas but neither taking the title as the definitive model for how categorization functions in the brain. Both Exemplar and Prototype models have their strengths and evidence for implementation, but they often oppose each other as incompatible ideas.

Categorization research has a rich history of study from many different perspectives. One paradigm for studying categorization is through what we will refer to as the *Nominal-Ordinal Feature Paradigm*. Under this paradigm, the focus falls on the relationships between specific features making up categories that observers can extract and compare against. Often this is broken down into exploring learned rules that observers may be using to categorize the stimuli presented to them. There are generally two ways that observers may be learning these rules; *Intentional Learning* and *Incidental Learning*. With intentional learning, observers approach the problem of categorization analytically and readily search for single-dimensional rules. Under incidental learning (sometimes referred to as unintentional learning), observers approach categorization non-analytically and use approaches such as *family resemblance* (Brooks, 1978; Kemler Nelson, 1984). The categories themselves are better described through family resemblance approaches, but participants in real-world experimental settings tend not to use these types of rule sets. Instead, participants have been shown to actively search for simple, single-dimensional rules for even complex natural categories, despite this rarely being an optimal strategy (Brooks et al., 2007).

Research under the *Nominal-Ordinal Feature Paradigm* typically uses categories of objects or stimuli meant to be perceived holistically. Some for instance use fictional, novel drawings of objects and animals as their stimuli set. These cartoon drawings are often simplified to only contain information about the relevant features separating the categories. An example of such imaginary creatures used in categorization research are the Bleebs and Ramuses found throughout work by Dr. Lee Brooks (Brooks et al., 2007; Hannah & Brooks, 2009). The features defined in these types of stimuli sets are often binary—smooth versus rough, spots versus stripes, or round versus angular.

Alternative to the *Nominal-Ordinal Feature Paradigm* is the *Continuous Feature Paradigm*. Under this paradigm, categorization is treated as a statistical learning problem where the observer measures the statistics of various observable features of the *distal stimulus* and matches it to learned categories. For example, pecans are a category of nut defined by shape, size, and knobbliness. Each of these separated features or cues is integrated to form the percept of the nut, which in succession informs knowledge of the underlying distribution of the category the object originally came from.

Research under the *Continuous Feature Paradigm* typically uses categories of more abstract stimuli with a continuous parameter space. Rather than comparing the binary presence of stripes or spots, categories may be comprised of features like the orientation of bars, size, or the number of dots present (Smith et al., 2012; Ashby & Gott, 1988;

Bankieris et al., 2017). For instance, Smith et al. (2010) explored categorization in both human and monkey observers using circular sine-wave disks which varied by bar width and bar orientation, two continuous parameters that the experimenters could easily manipulate to generate any number and variation of stimuli. The two categories were defined by Gaussian distributions with set means and variances, and stimuli were drawn from these distributions during experiments.

1.2 Perception as an inverse problem

Perception is a general term describing the process by which the brain makes observations about the world through its various sensory systems, resulting in our abilities to see, hear, and feel. A percept, then, is the object that was perceived through these systems—the instantiation of the process that is held in perceptual awareness through attention. The brain is an expert at this process, quickly and efficiently making decisions on what it perceives in its environment and feeding the final, holistic percepts back into the awareness of the observer. However, the percept held in awareness is not a perfect image of the object explored in reality. The sensory system is subject to neural noise, often found to be Poisson-like (Tolhurst et al., 1983; Knill & Pouget, 2004; Beck et al., 2008). Consequently, perception is by nature an inference problem; the task of the brain is to infer the most likely percept from this noisy sensory data (Yuille & Bülthoff, 1996; Feldman, 2012).

We call the *distal stimulus* the object in the external environment, and the *proximal stimulus* the impression of that object on the sensory systems (for instance, light cast onto the retina, or the impression of an object against the skin). The *proximal stimulus* activates sensory receptors and gives rise to a neural signal. It is this neural signal, together with relevant prior knowledge, that the brain acts upon to create an inference which we call the *percept* (Pizlo, 2001; Feldman, 2012). Vision, like all sensory systems, is a system faced with perceptual inference problems, where an observer is tasked with determining the most likely percept from noisy sensory data (Yuille & Bülthoff, 1996).

We can understand the process of perception further by considering Marr's *Levels of Analysis* (Marr, 1982). A perceptual system can be split into three levels: the *computation* of the system, the *algorithm* the system uses to undertake that computation, and the *implementation* of this algorithm (McClamrock, 1991; Marr, 1982). The *computation* refers to the goal or task of the system. In the context of our examination of perceptual processing, this refers to the objective of perceiving the external world and forming accurate percepts of distal stimuli. However, the proximal information available is noisy and the *algorithms* employed by the perceptual system must accommodate for this noise. If the information was not transformed in some way and the perceptual system took the noisy proximal information as is, it would be subject to many mistakes and errors.

Instead, it makes inferences about the distal stimuli based on the proximal stimuli. Computational categorization models often describe various theoretical algorithms that fit into this second level of analysis. The third *implementation* level refers to how these algorithms are physically implemented in the brain, the neural circuitry that carries out these calculations.

The problems faced by perception lend well to be understood as an inverse problem. As Feldman (2012) describes, the problem faced by perception is determining what should be believed about the world from sensory data. The overall goal of the system is to determine the *distal stimulus* presented to the observer, but it does not have direct access to this. Instead, it receives a noisy neural signal from the *proximal stimulus* and must infer the *distal stimulus* that caused it. Bayesian inference provides a natural framework to solve this inverse problem.

1.3 Bayesian inference applied to perception

Bayes' theorem, sometimes called Bayes' rule, was initially laid out by Thomas Bayes in 1763 who noticed the relationship between the conditional probabilities of A and B when $P(A|B)$ could yield a formula for the *inverse* conditional probability $P(B|A)$ (Bayes, 1763; Feldman, 2012):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

This equation allows us to solve the inverse problem faced by perception by calculating the Likelihood of the *proximal stimulus* measured by an observer given a hypothesis for the *distal stimulus*. We can see this is dependent on the Prior probabilities of the hypothesis for the *distal stimulus* and the *proximal stimulus*. Assuming that all the possible hypothetical *distal stimuli* are mutually exclusive, we can calculate $P(B)$ as the sum of the Likelihoods for the *proximal stimulus* and each hypothesis. Replacing A with H_i for the hypothetical *distal stimulus* and B with the *proximal stimulus* data D, we can rewrite Equation 1.1 for a particular hypothesis H_1 as follows:

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{\sum_i P(D|H_i)P(H_i)} \tag{1.2}$$

Through Equation 1.2 an observer can infer the *distal stimulus* with the highest Posterior probability based on their *proximal stimulus* measurements. Perceptual models rooted in Bayesian inference have found a lot of success in explaining many perceptual phenomena and illusions (Ernst & Bülthoff, 2004; Weiss et al., 2002; Knill & Pouget, 2004). For example, Knill (2007) demonstrated that a robust Bayesian model fits human performance in cue-conflict experiments. Cue-conflict was first examined by Rock &

Victor (1964) through a phenomenon coined as *visual capture*. When observers were presented with conflicting visual and haptic information about the size of an object, the perceived size was biased towards the visual information (Rock & Victor, 1964). This phenomenon was later explored through the *ventriloquist illusion* by Alais & Burr (2004) who found that *visual capture* was explained through cue integration models that take into account feature reliability. The visual system often provides more reliable information than other senses and so gets weighted more heavily when combined to form percepts. Degrading visual information (such as by blurring the visual stimuli) allows for the 'capture' of information from other senses. Knill (2007) extended this idea further by applying a Bayesian model that assumed different hypotheses or categories for the nature of the stimuli being perceived by observers, providing strong supporting evidence for a Bayesian model for robust cue integration.

The Bayesian framework can be applied beyond individual percepts to the perception of categories as a whole. The percepts inferred by a Bayesian perceptual system can in turn be sorted into categories. Categorization can be seen as a higher-order perceptual problem. The categories to which a percept may belong provide contextual information which can feed backward to influence perception.

Throughout this thesis, we explore categorization through our own developed model which falls under the *Continuous Feature Paradigm* approach. This model is probabilistic, employing Bayesian inference and treating categorization as a perceptual inference problem driven by feature measurements and the understanding of underlying categorical distributions. We call this model the Bayesian Categorization Model. Furthermore, we focus on stimulus perception and categorization exclusively through an observer's sense of touch.

1.4 Haptic exploration

The sense of touch is essential for our everyday interactions with the world. Through touch we explore our environments, effortlessly and quickly making observations and measurements of the various objects and stimuli we find within them. But the brain doesn't just hold disparate raw measurements of what it observes; decisions are made quickly and effortlessly about the objects we find, combining a multitude of measurements into holistic percepts and categorizing these percepts based on what we know about the world. Often we are unaware of all of the minute calculations and measurements our brains are making—we perceive the categorical judgement without needing to worry about how that judgement was made.

The process of exploring an object to gather information through touch is called *haptic exploration*. The term Haptic refers to the use of the tactile perception system, including cutaneous receptors and proprioceptive feedback from muscles, joints and tendons (Lederman & Klatzky, 1993) A defining characteristic of *haptic exploration* is that it involves active object manipulation. Lederman & Klatzky (1993) observed several different, systematic haptic exploratory procedures people perform that each extract specific information about some aspect of an object. *Lateral Motion*, described as repetitive surface rubbing, extracts information about texture. *Enclosure* and *Contour Following*, on the other hand, extract information about the global shape of an object. These techniques differ with regard to the time they take and the information they provide. *Contour Following* extracts more precise information about shape than *Enclosure* does, but at the cost of time; *Contour Following* can take upwards of 11 seconds to extract all of its information while *Enclosure* only requires 2 seconds, and *Lateral Motion* requires 3.5 seconds. Finally, some techniques are incompatible with each other as they cannot be performed at the same time. *Lateral Motion* cannot effectively be performed alongside *Enclosure*, but can be performed at the same time as *Contour Following*.

Haptic exploration often integrates with other senses, Vision and touch are used side-by-side every day to interact with the world. Klatzky et al. (1993) found that *haptic exploration* of objects is guided by vision; visual preview of an object may provide coarse information that helps direct further exploration by touch. Interestingly, the information provided through touch is not necessarily processed from the same viewpoint as how you see the object (Newell et al., 2001). An observer reaching out and grabbing an upright object will naturally have their fingers fall on its backside, and this is reflected in the viewpoint information. To explore this dynamic, Newell et al. (2001) asked participants to learn objects either visually or haptically and then attempt to recognize the objects through either the congruent or incongruent sensory modality. Participants performed worse when learning haptically and recognizing visually, but this improved significantly when the object was rotated 180 degrees horizontally during recognition. Thus, when they were seeing the backside of the object, they were able to recognize what they previously felt. This suggests the information gathered through *haptic exploration* had the opposite viewpoint from vision.

1.5 Computational categorization models

The Bayesian Categorization Model we developed and explored throughout this thesis is a type of computational model—a proposed abstract algorithm which may represent the process by which a perceptual system computes information. While we believe our approach and analysis to be unique, especially in the context of applying this paradigm to the tactile system, there are many other related computational categorization models and approaches. In this, we are not saying that our Bayesian Categorization Model is itself absolutely unique; it is based on the same principles shared by other many models. Here we will cover a select few computational models which are directly related to our own in some way.

General Recognition Theory

General Recognition Theory suggests observers base their categorical decisions on calculated decision boundaries separating the categories. Ashby & Gott (1988) set up a two-dimensional problem to explore the observer's decision processes in a categorization task. Knowing the parameters of these distributions allows for a decision boundary to be drawn that an observer might use to categorize stimuli from these distributions. They empirically tested whether the features were evaluated independently or integrated, making a more optimal decision boundary. These optimal decision boundaries can be calculated through a few algorithms: a minimum distance classifier, which finds a boundary at the minimum distance between the means of the two distributions and is constrained to be orthogonal to a line passing through both means; a general linear classifier, which is similar to the minimum distance classifier but without constraints on slope or intercept; and an optimal decision rule, which compares the likelihood ratio of the two category distributions against some criterion value. They concluded that the decision bounds of observers were best represented by those calculated through integrating the features and found support for a near-optimal general linear classifier, though needed further work with different category formulations to find evidence for their optimal decision rule model. This paradigm was used later by Smith et al. (2010) when exploring category learning in both humans and macaques to success.

General Recognition Theory is rooted in *Signal Detection Theory* (*SDT*), which is a general framework for modelling an observer's ability to detect the presence or absence of a *distal stimulus*,—especially in the presence of uncertainty or noise. *SDT* is the underlying framework for the field of *psychophysics* and is rooted in understanding perception through the *proximal* measurements made by an observer's sensory systems. By considering the statistics underlying categories as multivariate normal or Gaussian, we can see that General Recognition Theory is an extension of *SDT* applied in multiple dimensions. However, a criticism of this approach is whether it truly represents a two-dimensional problem. The stimuli used by Ashby & Gott (1988) were two lines of varying length, one horizontal and one vertical, connecting perpendicularly at a right angle. The two features defining the categories were the lengths of either line, which were continuous variables. As pointed out by Macmillan & Creelman (2005, Chapter 12), this two-dimensional feature space can be re-contextualized as one-dimensional. A compound decision boundary, referred to as $d`_{compound}$, can be calculated through a simple equation which leverages the Pythagorean Theorem in relating the two feature dimensions:

$$d`_{compound} = \sqrt{d`^2_x + d`^2_y} \tag{1.3}$$

Using Equation 1.3 immediately leaves us with the same decision boundary that Ashby & Gott (1988) calculated through their general linear classifier. Conceptually, we can consider the two features instead through a single feature defined by the difference between the line lengths. This was somewhat addressed through a separate supplemental experiment with a new stimulus of a horizontal line going through a rotated semicircle. In this scenario, one feature was the radius of the semicircle (the length of the horizontal line) and the other feature was the rotation of the semicircle (the angle at the intersection of the semicircle's straight edge and the horizontal line). They argued that radius and rotation were not on the same perceptual scale and thus could not be reduced to a one-dimensional problem. However, there is a way for these stimuli to be represented through a single feature—arc length. The length of the arc formed by the semicircle and horizontal line is directly proportional to both the radius and angle of rotation. Thus, this may yet again be a poor example of a two-dimensional categorization problem.

This raises interesting considerations for categories with multiple feature dimensions. Can any two-dimensional problem be reduced to one dimension by abstracting onto a compound feature space? While the stimuli from Ashby & Gott (1988) can be reduced this way through a mathematical transformation, this is not always so perceptually trivial to do. For instance, Smith et al. (2010) used visual sine-wave disks which varied by bar width and bar orientation. These two features do not have an immediate conceptual one-dimensional analogue. And stimuli that involve measurements from multiple sensory systems, such as visual displays of dots and audio pitches, cannot be easily conceptually reduced in this way. While *SDT* is primarily based on the measurements made by an observer, abstracting from the physical instantiations of these measurements into a generalized parameter space allows for these mathematical transformations outside of conceptual constraints. However, finding approaches that maintain parity with the measurements of the *distal stimuli* is important for ecological validity.

Categorical Sensory Cue Combination

In daily life, humans use information from multiple senses to accurately perceive a scene or explore an object. Each sense provides an informational cue that the brain can use to form an overall percept. These different cues are often not weighted equally (Rock & Victor, 1964). Vision tends to overpower the other senses in localization tasks, leading to a phenomenon called Visual Capture. This phenomenon was explored further by Alais & Burr (2004) through the Ventriloquist Illusion, where they demonstrated that the information provided by vision and audition is weighted by the variance, or reliability, of that information. This reliability is associated with the acuity of the sensory system providing the information, although it can be manipulated by degrading or improving the stimulus (Alais & Burr, 2004; Samad et al., 2015).

Cue combination is a powerful approach to perceptual inference problems where multiple sources of information need to be combined or integrated in some way to form an overall percept. It is a model which again relies on the underlying proximal measurement distributions from a *distal stimulus* input but instead considers how these measurements combine. This model considers the uncertainty or noise of the measurements being made when information is integrated. This is especially important when conflicting information is introduced between measurements. If an observer's sensory system has less certainty about one measurement than the other, it will be down-weighted and the resulting percept will be closer to the other measurement (Alais & Burr, 2004). Cue combination models have been successfully pushed further to predict inferred causal relationships between the cues; events occurring too distant in space should be less likely to be associated together as arising from the same cause, and the Causal Inference Model can match human performance demonstrating this effect (Körding et al., 2007).

Some previous work has explored cue combination integrating tactile cues with those from other sensory modalities. Gepshtein & Banks (2003) demonstrated that observers optimally combined visual and haptic information in size perception. When both the visual system and tactile system provide information about a certain cue, the brain integrates them to form an overall estimate predicted by cue combination formulae. Another studied phenomenon is the *Rubber Hand Illusion*, where the sense of body ownership is manipulated to incorporate an external object (Kalckert & Ehrsson, 2014; Samad et al., 2015). In this phenomenon, an observer's hand is obscured from view and visually replaced with a fake rubber hand. Both the rubber hand and the observer's real hand are stroked with a brush. The visual information of a rubber hand being brushed is integrated with the tactile sensation of the observer's hand being brushed, leading to an illusory perception of body ownership over the rubber hand.

While cue combination tasks are not discussed in terms of categorization, they are nevertheless related. The difference between the sensory cue combination paradigms described above and more generalized categorization tasks is the context of the given task and the objective of the observer. Sensory cue combination generally describes situations where the observer is tasked with making a judgement about a single, typically continuous feature, such as location in space, through multiple senses or measurements. We can consider categorization as an extension of cue combination, introducing an additional level of judgement for the observer to consider in their observations about the world.

Recently, Bankieris et al. (2017) extended the cue combination paradigm into categorization. By defining their categories as multivariate normal distributions and considering the measurement noise of the individual features themselves, they developed a model which accounted for both the measurement noise of the feature measurements and the categories those measurements arose from. They considered the problem as a combination of both the measurement distributions and the category distributions. Effectively, knowledge about the underlying categories should help inform observations of each cue. A decision variable D can be calculated by a weighted average between the proximal signals of the two features which would reflect this combination of distributions.

$$D = w_A S_A + w_B S_B \tag{1.4}$$

$$w_A = \frac{\frac{\Delta\mu_A}{\sigma^2_{A,sense}+\sigma^2_{A,cat}}}{\frac{\Delta\mu_A}{\sigma^2_{A,sense}+\sigma^2_{A,cat}} + \frac{\Delta\mu_B}{\sigma^2_{B,sense}+\sigma^2_{B,cat}}}, w_B = \frac{\frac{\Delta\mu_B}{\sigma^2_{B,sense}+\sigma^2_{B,cat}}}{\frac{\Delta\mu_A}{\sigma^2_{A,sense}+\sigma^2_{A,cat}} + \frac{\Delta\mu_B}{\sigma^2_{B,sense}+\sigma^2_{B,cat}}} \tag{1.5}$$

Bankieris et al. (2017) formed their categories as bivariate Gaussian distributions along two features: a visual array of dots and an auditory pitch. Weber's law—a perceptual rule of thumb that relates the perception of change in a sensory measurement to be inversely proportion to the initial intensity of the measurement—was used to determine the step size within their feature range such that each step would be equally discriminable. The categories were defined by a mean and Standard Deviation (σ) within this two-dimensional feature space. Furthermore, the researchers were able to manipulate the measurement noise of the auditory cue with the addition of pink noise, allowing them to test whether the resulting behaviour of an observer followed their optimal integration which considered both measurement noise and category information in their decision. They found that their model fit human data well, marking one of the first applications of the cue-combination paradigm in a higher-order categorization decision task. Importantly, this is one of the only studies into a categorization model that considers and accounts for the presence of measurement noise, which we will explore in more depth in Chapter 4.

Rational Model of Categorization

A previous Bayesian approach to categorization is the Rational Model of Categorization (RMC) (Anderson, 1991). This model is probabilistic and can be generalized for complex scenarios where the number of categories is unknown. The RMC treats categorization as a parameter estimation problem where an observer attempts to learn the underlying categorical structure of a set of stimuli when they don't know what those categories are nor how many categories there are. If N objects have stimulus features $\mathbf{x}_N$ and category labels $\mathbf{y}_N$, and there are a total of J category labels, then the probability of an object belonging to category j and being labelled $y = j$ given its features $\mathbf{x}$ can be calculated by the following:

$$P(y = j|\mathbf{x}) = \frac{P(y = j)P(\mathbf{x}|y = j)}{\sum_{J_i} P(y = J_i)P(\mathbf{x}|y = J_i)} \tag{1.6}$$

In other words, we can calculate the probability of a new object being correctly labelled j by calculating the prior probability of that being the true label $P(y = j$ and the likelihood that the object would have its observed features given that j is the correct label $P(\mathbf{x}|y = j)$. The prior $P(y = j)$ in this context is heavily dependent on how likely a new object gets grouped into an existing category or not. For this, Anderson (1991) proposed two possibilities which might be considered by the observer: either the object is from an existing category, or it is from a new one. In the first case, the prior $P(y = j)$ can be calculated through the following equation:

$$P(y = j) = \frac{cM_j}{(1 - c) + cN} \tag{1.7}$$

where c is an arbitrarily set coupling probability and M_j is the number of objects currently assigned to category j. This describes a greedy algorithm where existing categories with many objects are more likely to get new objects assigned to them. Intuitively it follows from Equation 1.7 that if an observer has seen most of the objects presented to them assigned to category A and only a few assigned to category B, then their prior probability that the next object would be assigned to category A will be higher than the prior probability that it be assigned to category B.

The second case for the prior $P(y = j)$, where the object may belong to an entirely new category, can be calculated through the following equation:

$$P(0) = \frac{(1 - c)}{(1 - c) + cN} \tag{1.8}$$

With the RMC, an observer can sequentially observe new objects and either categorize them into a pre-existing category or place them in a new category. Interestingly, this Bayesian model is more applicable to the *Nominal-Ordinal Feature Paradigm* as it doesn't involve consideration of the features as continuous distributions. However, it could be further extended to describe the statistical distribution underlying the features making up categories and even involve further parameter estimation of the category distribution means and σ's themselves. However, there is a major limitation of the RMC with the massive computational cost of calculating the entire Posterior distribution.

To get around this problem, Anderson (1991) approximated the Posterior by employing a local *maximum a posteriori* (MAP) algorithm. This way, the algorithm does not need to recalculate the Likelihood of the previous object category labels with every new object. Unfortunately, this makes the model highly biased to the order of object presentations. After several categorizations have been made and the model has more information about what the underlying categories are, it might be reasonable to be able to go back and re-evaluate previous categorizations. Doing so, however, exponentially increases the computation required to calculate the Posterior which quickly becomes unfeasible. This drawback held the RMC back from further work for many years.

More recently, the RMC was revitalized with modern computational techniques, solving the computational cost issues. To accomplish this, Sanborn et al. (2010) recontextualized the RMC as a Dirichlet Process Mixture Model. A Dirichlet distribution is a multivariate form of the Beta distribution. Beta distributions are a convenient function commonly used in Bayesian statistics for calculating broad priors. These distributions strike somewhat resemblance to Gaussian distributions but are unique in that they are constrained to the interval between 0 and 1. This makes them convenient for many Bayesian applications.

Consider a category as a cluster of objects k. $\mathbf{z}_N$ was defined as the set of category cluster assignments, and K was the total number of clusters in $\mathbf{z}_N$. The number of objects assigned to cluster k was defined as M_k. From this, Sanborn et al. (2010) recognized that Equations 1.7 and 1.8 could be redefined by the following equation:

$$P(\mathbf{z}_N) = \frac{(1-c)^K c^{N-K}}{\prod_{i=0}^{N-1}[(1-c)+ci]} \prod_{k=1}^{K}(M_k-1)! \tag{1.9}$$

The researchers observed that Equation 1.9 is equivalent to one derived from Dirichlet processes calculating the probability that a Dirichlet process separates N observations into clusters $\mathbf{z}_N$:

$$P(\mathbf{z}_N) = \frac{\alpha^K}{\prod_{i=0}^{N-1}[\alpha+i]} \prod_{k=1}^{K}(M_k-1)! \tag{1.10}$$

With Equation 1.10 they could apply advanced computational approaches such as Gibb's Sampling or Particle Filtering to calculate a closer approximation for the posterior distribution of the problem, thereby solving the computational cost issues of the RMC while allowing backtracking for re-evaluating previous categorizations.

This Bayesian approach to categorization is compelling because it encapsulates both *Prototype Theory* and *Exemplar Theory* as extreme cases. *Exemplar* models can be described as a case where each object is assigned to its own category, and *Prototype* models can be described as a case where all objects are assigned to the same category (Sanborn et al., 2010). However, the RMC also allows the possibility for in-between cases where some set of objects can be assigned to one category while the others are assigned to another. The value of the RMC and the complexity of the categorization problem it is trying to solve lies in Equations 1.7, 1.8, and 1.10. The RMC is an attempt to model category learning when the number of categories is unknown. This situation is often circumvented by explicitly making the number of categories known to the observer.

For our categorization experiments throughout this thesis, we make clear to observers that there are two categories in our categorization tasks. This allows us to ignore the computational problem of observers learning multiple unknown categories. Instead, we assume throughout that our participants are aware that there are two categories. This constraint simplifies our computational model greatly as we do not need to include the RMC's Prior term in our model.

1.6 Overview of Chapters

Throughout the rest of this thesis, we explore a Bayesian-driven categorization paradigm within the sense of touch, testing the efficacy of a Bayesian observer model against human performance. Chapter 2 describes the main experimental procedure and the derivation of the observer model. Chapter 3 focuses on two foundational experiments testing the model and paradigm in a human participant setting. Chapter 4 explores the role and influence of internal sensory noise in our categorization paradigm, detailing simulations exploring the expected outcomes of the model and an additional experiment exploring the sensory resolution of participants. Finally, Chapter 5 discusses the overall conclusions of our experiments, how our model fits with the greater bodies of categorization research, and the limitations and possible future extensions of our work here.

Chapter 2

Bayesian Categorization Model

2.1 Introduction

The approach to categorization we will explore throughout this thesis is our Bayesian Categorization Model. This is an observer model that takes the information available and uses Bayesian inference to determine the best categorical response based on the posterior probability for each category. We will consider categories as a set of objects that have certain features in common, and an object as a set of features that are combined by an observer into a holistic percept. Observers sense an object's *distal* features using their sensory systems resulting in *proximal* measurements. These measurements are made in some hypothetical parameter space. A rock might have a particular weight measurement, hardness rating, or temperature.

We can consider the variation of these features as part of categorical representations. Variation is found in natural stimuli throughout the world, most often following a Gaussian distribution. For example, while a tree might have an average leaf size, the precise size of every leaf is not identical. If you were to create a histogram of all leaf sizes found on a tree you would find that they follow a distribution. Feature values can then be defined and stored by only two variables: an average value, and the variation about that value.

One complexity with this approach to category representation has to do with an assumption made earlier in the rock example; how does the observer know to compare their object to a rock? There are theoretically infinite numbers of categories someone can form as they go about in real life. It is unrealistic to expect that observers compare the object they are trying to identify with all possible stored category representations. This would be costly both in terms of resources and time. However, observers typically have further information available about an object, including the environment and context in which it was found. If you find yourself in a library, you might expect to encounter desks, tables, bookshelves, books, and computers. If you're on a hiking trail, then you might expect trees, trail markers, birds native to your environment, and rocks. The context in which you expect to find particular objects should provide a good starting point when trying to identify unknown objects. From a Bayesian perspective, all this background information can be considered as the Prior Probability that an object belongs to a particular category.

Treating categorization as a parameter estimation task frames the scenario as an inverse problem: observers make measurements of an object but must then decide what category gave rise to those measurements. Context and background information help constrain the hypothesis space that comparisons are being made from, and the observer can determine the probability that their measurements would occur, given the object belonged to a particular category.

This type of problem lends itself well to a Bayesian approach. Bayes' Theorem is a probabilistic equation that calculates the conditional probability of a hypothesis given prior knowledge and observations made. As we are defining categories with respect to their feature distributions, Bayesian inference provides a probabilistic way to solve the inverse problem we described above.

We propose a Bayesian model capable of resolving the categorization problem we described previously called the Bayesian Categorization Model. We will first discuss the development of a novel stimulus set designed for a tactile categorization experiment paradigm. This will allow for training and testing of categories using the sense of touch in human participants and testing of our Bayesian approach to the problem. We will then derive our Bayesian Categorization Model, the exploration of which will be the focus of the rest of the thesis.

2.2 Physical Stimulus Set for a Categorization Task

To test our Bayesian Categorization Model, we first needed to design a stimulus set of objects and categories. In a typical categorization task, a category is defined by a set of features that an observer must measure and then identify. A common approach is to model these categories as Gaussian distributions: each measurable cue or feature is represented as a single dimension in a multivariate Gaussian distribution. Under this design, a category can be summarized as a set of means and standard deviations for each of these feature dimensions.

To study a 2-Category task within the sense of touch it was necessary to create a novel stimulus set. Exploration within the sense of touch creates a unique methodological challenge that a typical visual or auditory categorization task does not face: since the stimuli need to be physically created and stored, we are limited in the number of stimulus variations we can reasonably present to the participant. Compared to a typical categorization experiment we have relatively few exemplars to train and test participants on. Despite this challenge, we were successful in designing a stimuli set with enough resolution to employ a 2-Category learning experiment.

The novel stimuli created for these experiments were plastic polygonal discs with different numbers of sides and an evenly spaced arrangement of raised hemispherical dots on one of the surfaces, acting as a macro-texture (Figure 2.1). These two features were chosen according to the distinct different extraction methods an observer may use to measure each feature. Following the taxonomy of Lederman & Klatzky (1993), the Number of Sides feature relates most closely to a "shape" feature whose exploration involves *Enclosure* of the hand around the object, and the Dot spacing feature relates most closely to a "texture" feature whose exploration involves *Lateral Motion* of the hand or fingers along the surface of the object. The distinctness of these two features was important to ensure observers would treat them as two independent measurements.

A set of 25 objects were designed which varied along these two dimensions, allowing for a combination of 5 different Number of Sides and 5 different Dot spacings in 1mm intervals. These objects were modelled in the *openSCAD* program to ensure they were physically identical in all other respects. All of the objects had the same thickness, radius, and roughly the same weight. The objects were created using the *Ultimaker 2 Go* 3D printer in PVC plastic. They were stored within a box aligned in a 5x5 grid (Figure 2.2) for ease of use during the running of an experiment. A list of all 25 objects and their associated features can be found in Table 2.2.

It is worth noting that although here we are treating each of the two features as separate, single dimensions with one type of measurement needed to extract the information, several different measurements can be made for each feature. The Sides feature can be measured in a few different ways; a participant may count the number of sides, measure the length of each side, or measure the vertices of each side. Likewise, the Dots feature can be measured in millimetres of spacing between dots, or the number of dots can be counted, or an observer can feel the frequency of vibrations as they quickly scan their finger across the surface.

We did not instruct participants on the most effective method for extracting each feature measurement and instead encouraged them to acquire this information in whatever way felt natural for them. It is assumed that, given ample time, an observer will be able to make a sufficient measurement for the Sides or Dots features with any of these methods. Thus, for our experiments, these features will be treated as single measurements, although we acknowledge the potentially complex information processing that goes into extracting a single feature with multiple sources of information. Realistically, measuring the two features can be considered feature-combination problems in themselves.

2.3 Category Formation

We defined a category C within our novel stimulus set as a mean Number of Sides μ_{Cn} and Dot spacing (mm) μ_{Cd} with some variance σ_{Cn} and σ_{Cd} along each respective feature dimension. Through this method we established two categories, A and B. Category A had a mean Sides $\mu_{An} = 7$, mean Dot spacing (mm) $\mu_{Ad} = 5mm$, and standard deviations $\sigma_{An} = 1.5$ Sides and $\sigma_{Ad} = 1.5mm$ Dot spacing for each feature. Category B had a mean Sides $\mu_{Bn} = 9$, mean Dot spacing $\mu_{Bd} = 7mm$, and standard deviations $\sigma_{Bn} = 1.5$ Sides and $\sigma_{Bd} = 1.5mm$ Dot spacing for each feature. The two features, Sides and Dot spacing (mm), are conditionally independent given the category. We refer to these distributions as *Within-Category Feature Distributions* or Class-Conditioned Feature distributions. From this point forward, while the Dot spacing feature was set in millimetre spacing between the dots, we will present it as unit-less for easier comparison to the Sides feature.

Figure 2.3 displays 2D intensity plots of the two category distributions described above. You can notice immediately that these categories are perfectly symmetrical. They are represented as 2D Gaussian distributions with equal variance in either feature dimension. However, the means of the two distributions are still close enough that there is heavy overlap between the categories. This means that our two categories are perceptually similar and it is not unlikely that an A object be chosen that looks closer to being a B object. In such a situation, an ideal observer would incorrectly categorize the object as B. This is our expected outcome, and we chose these distribution distances such that an ideal optimal observer would reach a maximum percent correct of 76%. Also to note is that these distributions are truncated due to limitations of physically producing the stimulus set. As we go on to show later, this detail turns out to be mathematically irrelevant to the categorization task.

The category feature distributions (Figure 2.3) can be considered frequency plots for each of the 25 objects within either category. If you were to sample a box filled with only Category A objects, the plot represents the frequency at which each object would be drawn, with the brightest spot being the most common.

We will now derive and describe our Bayesian Categorization Observer Model as it will appear throughout the rest of this work. Although more complex models exist within this framework, such as the Rational Model of Categorization (RMC), we decided to simplify the problem to a special case where the observer knows that there are exactly two categories, thereby foregoing the need for our model to try to determine how many categories there are through the RMC's Prior definition. This was both a convenience for our modelling work as well as a convenience for our experimental design as our aim was to run human participants through this categorization task. It is much simpler for a human participant to learn the true categories of our objects if they know from the start that only two categories exist.

2.4 Bayesian Categorization Model Derivation

Consider two categories, A and B, and a set of feature measurements $\mathbf{f}_m$ (bolded $\mathbf{f}$ referring to a set of feature measurements). In order to determine the probability of category A given $\mathbf{f}_m$, which we will write as $P(A|\mathbf{f}_m)$, Bayes' rule can be applied:

$$P(A|\mathbf{f}_m) = \frac{P(\mathbf{f}_m|A)P(A)}{P(\mathbf{f}_m|A)P(A) + P(\mathbf{f}_m|B)P(B)} \tag{2.1}$$

$P(A)$ and $P(B)$ are the Prior probabilities that the object belongs to either category. $P(\mathbf{f}_m|A)$ and $P(\mathbf{f}_m|B)$ are the Likelihoods of each hypothesis: the probability of observing $\mathbf{f}_m$ given the object belongs to either category.

With our stimulus design, the set of feature measurements $\mathbf{f}_m$ is a combination of two features: n_m Sides and d_m Dot spacing.

$$P(A|n_m, d_m) = \frac{P(n_m, d_m|A)P(A)}{P(n_m, d_m|A)P(A) + P(n_m, d_m|B)P(B)} \tag{2.2}$$

Since features n_m and d_m are set to be conditionally independent, we can consider the probabilities $P(n_m|A)$ and $P(d_m|A)$ separately.

$$P(A|n_m, d_m) = \frac{P(n_m|A)P(d_m|A)P(A)}{P(n_m|A)P(d_m|A)P(A) + P(n_m|B)P(d_m|B)P(B)} \tag{2.3}$$

For the purposes of the experiments in this thesis, we will be considering only two categories with two feature measurements. We will also be experimentally enforcing equal Prior probabilities for the category $P(A) = P(B) = 0.5$ by randomly drawing from the categories with equal probabilities. Thus, Equation 2.1 can be simplified as:

$$P(A|n_m, d_m) = \frac{P(n_m|A)P(d_m|A)}{P(n_m|A)P(d_m|A) + P(n_m|B)P(d_m|B)} \tag{2.4}$$

The category likelihoods $P(n_m|C)$ and $P(d_m|C)$ can be represented as one-dimensional Gaussian probability densities with a mean μ_{Cf} and sigma σ_{Cf} determined by each of the category representations. For instance, Category A was defined to have a mean $\mu_{An} = 7$ and $\sigma_{An} = 1.5$.

We can specify $P(f_m|C)$ using the Gaussian formula:

$$P(f_m|\mu_{Cf}, \sigma_{Cf}) = \frac{1}{\sigma_{Cf}\sqrt{2\pi}} \exp\left(-\frac{(f_m - \mu_{Cf})^2}{2\sigma_{Cf}^2}\right) \tag{2.5}$$

From here we can see that the numeration of Equation 2.4 represents a two-dimensional

Bivariate Gaussian distribution, where $P(n_m|C)$ and $P(d_m|C)$ are independent Gaussian distributions with no covariance ρ between the two features. We can see this more clearly by applying Equation 2.5 to the $P(n_m|C)P(d_m|C)$ within Equation 2.4

$$\begin{aligned} P(n_m|C)P(d_m|C) &= \frac{1}{\sigma_{Cn}\sqrt{2\pi}} \exp\left(-\frac{(n_m-\mu_{Cn})^2}{2\sigma_{Cn}^2}\right)\frac{1}{\sigma_{Cd}\sqrt{2\pi}} \exp\left(-\frac{(d_m-\mu_{Cd})^2}{2\sigma_{Cd}^2}\right) \\ &= \frac{1}{2\pi\sigma_{Cn}\sigma_{Cd}} \exp\left(-(\frac{(n_m-\mu_{Cn})^2}{2\sigma_{Cn}^2} + \frac{(d_m-\mu_{Cd})^2}{2\sigma_{Cd}^2})\right) \end{aligned} \tag{2.6}$$

We can simplify Equation 2.6 further by assuming a common σ_{Cf} for the two features, which follows from our stimulus and category design where the $\sigma_{Cf} = 1.5$ for each feature.

$$P(n_m|C)P(d_m|C) = \frac{1}{2\pi\sigma_{Cf}^2} \exp\left(-(\frac{(n_m-\mu_{Cn})^2 + (d_m-\mu_{Cd})^2}{2\sigma_{Cf}^2})\right) \tag{2.7}$$

Substituting Equation 2.7 into 2.4 and rearranging yields an elegant expression for the posterior probability of Category A which describes the experimental paradigm explored within this thesis, where $\sigma_{Cf} = \sigma_{Cn} = \sigma_{Cd} = 1.5$:

$$P(A|n_m, d_m) = \frac{1}{1 + \exp\left(\frac{(n_m-\mu_{An})^2-(n_m-\mu_{Bn})^2+(d_m-\mu_{Ad})^2-(d_m-\mu_{Bd})^2}{2\sigma_{Cf}^2}\right)} \tag{2.8}$$

The model described in Equation 2.8 applies to the specific categorization task which will be explored further in Chapters 3 and 4. This adheres to specific constraints, such as a common Within-Category Feature Standard Deviation σ_{Cf} between each of the two features for the two categories, and no covariance present between the features. These are simplifications we can enforce in a laboratory setting while running our categorization experiment. In a real-world categorization scenario, there may be more than two categories, more than two features, and it would be unlikely for the σ_{Cf} to be equal.

In a 2-Category/2-Feature task where each feature has a unique σ_f, we can derive a similar equation to 2.8:

$$P(A|n_m, d_m) = \frac{1}{1 + \frac{\sigma_{An}}{\sigma_{Bn}}\frac{\sigma_{Ad}}{\sigma_{Bd}} \exp\left((\frac{(n_m-\mu_{An})^2}{2\sigma_{An}^2} + \frac{(d_m-\mu_{Ad})^2}{2\sigma_{Ad}^2}) - (\frac{(n_m-\mu_{Bn})^2}{2\sigma_{Bn}^2} + \frac{(d_m-\mu_{Bd})^2}{2\sigma_{Bd}^2})\right)} \tag{2.9}$$

2.5 Generalized Bayesian Categorization Models

In a multi-Category/2-Feature task, we can see that Equation 2.4 would expand to be a comparison between many Bivariate Gaussian Distributions. Suppose a hypothetical third *Category H* was introduced into the experimental paradigm. Assuming equal priors as before, by following the same derivation as shown above, we would end with the following equation for the posterior probability of Category A:

$$\begin{gathered}
P(A|n_m, d_m) = \frac{1}{1 + \frac{P(n_m,d_m|B)}{P(n_m,d_m|A)} + \frac{P(n_m,d_m|H)}{P(n_m,d_m|A)}} \\
\frac{P(n_m, d_m|B)}{P(n_m, d_m|A)} = \frac{\sigma_{An}}{\sigma_{Bn}}\frac{\sigma_{Ad}}{\sigma_{Bd}} \exp\left((\frac{(n_m - \mu_{An})^2}{2\sigma^2_{An}} + \frac{(d_m - \mu_{Ad})^2}{2\sigma^2_{Ad}}) - (\frac{(n_m - \mu_{Bn})^2}{2\sigma^2_{Bn}} + \frac{(d_m - \mu_{Bd})^2}{2\sigma^2_{Bd}}) \right) \\
\frac{P(n_m, d_m|H)}{P(n_m, d_m|A)} = \frac{\sigma_{An}}{\sigma_{Hn}}\frac{\sigma_{Ad}}{\sigma_{Hd}} \exp\left((\frac{(n_m - \mu_{An})^2}{2\sigma^2_{An}} + \frac{(d_m - \mu_{Ad})^2}{2\sigma^2_{Ad}}) - (\frac{(n_m - \mu_{Hn})^2}{2\sigma^2_{Hn}} + \frac{(d_m - \mu_{Hd})^2}{2\sigma^2_{Hd}}) \right)
\end{gathered} \tag{2.10}$$

We can see that as the number of category comparisons increases, the number of terms in the denominator increases in a regular, predictable fashion. Generalizing this for x number of categories would yield the following equation, where C_i are the categories in the comparison:

$$P(C_1|n_m, d_m) = \frac{1}{1 + \sum_{i=2}^{x} \frac{\sigma_{C_1 n}}{\sigma_{C_i n}}\frac{\sigma_{C_1 d}}{\sigma_{C_i d}} \exp\left((\frac{(n_m - \mu_{C_1 n})^2}{2\sigma^2_{C_1 n}} + \frac{(d_m - \mu_{C_1 d})^2}{2\sigma^2_{C_1 d}}) - (\frac{(n_m - \mu_{C_i n})^2}{2\sigma^2_{C_i n}} + \frac{(d_m - \mu_{C_i d})^2}{2\sigma^2_{C_i d}}) \right)} \tag{2.11}$$

In a 2-Category/multi-Feature task, we can see a similar regularity would arise where we are multiplying many independent Gaussian distributions together. Following a similar approach as above, we would arrive at the following equation, where y is the number of features:

$$P(A|f_{m_j}) = \frac{1}{1 + (\prod_{j=1}^{y} \frac{\sigma_{Af}}{\sigma_{Bf}}) \exp\left(\sum_{j=1}^{y} \frac{(f_{m_j} - \mu_{Af_j})^2}{2\sigma^2_{Af_j}} - \sum_{j=1}^{y} \frac{(f_{m_j} - \mu_{Bf_j})^2}{2\sigma^2_{Bf_J}} \right)} \tag{2.12}$$

Expanding to the multi-Category/multi-Feature task scenario we can arrive at a generalized formula to handle any number of categories x and features y, assuming equal priors amongst the categories and no covariance between the features:

$$P(C_1|f_{m_j}) = \frac{1}{1 + \sum_{i=2}^{x} (\prod_{j=1}^{y} \frac{\sigma_{C_1 f_j}}{\sigma_{C_i f_j}}) \exp\left(\sum_{j=1}^{y} \frac{(f_{m_j} - \mu_{C_1 f_j})^2}{2\sigma^2_{C_1 f_j}} - \sum_{j=1}^{y} \frac{(f_{m_j} - \mu_{C_i f_j})^2}{2\sigma^2_{C_i f_j}} \right)} \tag{2.13}$$

Generalized Bayesian Categorization Model for Unequal Priors

Up to this point, we have been assuming the simplified categorization scenario where the priors between the categories are equal. This is a fair assumption within the context of our experiments as we artificially set these priors to be equal. However, a true generalized Bayesian categorization formula would need to account for any set of priors. This would become especially important in future possible studies where the prior between the Categories may be changed or manipulated, or measured from the participant as a free parameter.

We will start with Equation 2.3 and follow a similar derivation as before, applying Equation 2.6 and leaving $P(A)$ and $P(B)$ as variables.

$$P(A|n_m, d_m) = \frac{1}{1 + \frac{P(B)}{P(A)} \frac{\sigma_{An}}{\sigma_{Bn}} \frac{\sigma_{Ad}}{\sigma_{Bd}} \exp\left((\frac{(n_m - \mu_{Bn})^2}{2\sigma^2_{Bn}} + \frac{(d_m - \mu_{Bd})^2}{2\sigma^2_{Bd}}) - (\frac{(n_m - \mu_{An})^2}{2\sigma^2_{An}} + \frac{(d_m - \mu_{Ad})^2}{2\sigma^2_{Ad}}) \right)} \tag{2.14}$$

From Equation 2.14 we can see that including the category priors in our calculations leaves us with the prior ratio $\frac{P(B)}{P(A)}$. This equation will reduce to Equation 2.8 under the conditions that there are two categories, two features within each category, the σ_f feature standard deviation is equal in each category for each feature, and the priors are equal. We can apply the derivation in Equation 2.14 to Equation 2.13 to come to a generalized formula for x categories and y features including the priors.

$$P(C_1|\mathbf{f}_m) = \frac{1}{1 + \sum_{i=2}^{x} \frac{P(C_i)}{P(C_1)} (\prod_{j=1}^{y} \frac{\sigma_{C_1 f_j}}{\sigma_{C_i f_j}}) \exp\left(\sum_{j=1}^{y} \frac{(f_{m_j} - \mu_{C_1 f_j})^2}{2\sigma^2_{C_1 f_j}} - \sum_{j=1}^{y} \frac{(f_{m_j} - \mu_{C_i f_j})^2}{2\sigma^2_{C_i f_j}} \right)} \tag{2.15}$$

Truncated Gaussian in a Categorization Task

It might be noticed that our usage of the standard Gaussian formula for $P(\mathbf{f}|C)$ in Equation 2.5 is technically incorrect in the context of our experiment as the stimulus set is not a continuous function. With an upper and lower bound truncating the distribution, the Truncated Gaussian formula would be more proper to use in our calculations:

$$\psi(f_m; \mu, \sigma, a, b) = \frac{P(f_m|\mu, \sigma)}{F(b; \mu, \sigma) - F(a; \mu, \sigma)} \tag{2.16}$$

where $P(f_m|\mu, \sigma)$ is the Gaussian formula as shown in Equation 2.5 and $F(x; \mu, \sigma)$ is its cumulative distribution function, where x is either the upper or lower bound. This equation applies within the range $(a \leq f_m \leq b)$, and returns 0 anywhere outside of this range. $F(x; \mu, \sigma)$ can be expanded as the following:

$$F(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt \tag{2.17}$$

An important detail about the Truncated Gaussian which becomes apparent upon inspection is that the denominator for 2.16, which is a difference of two cumulative distribution functions set at either end of the distribution, is a constant term not dependent on the feature measurement f_m. As long as the categories being compared share the same feature distribution spaces, this common denominator will be divided out of the equation and not impact further derivations. Therefore, while it is technically improper to use the standard Gaussian equation in our derivations, the result will be the same regardless and so it is safe to use the simpler form.

Posterior Probability Matrix

Using Equation 2.8 we can calculate the posterior probability that any of the 25 objects, which we will also refer to as $\mathbf{f}_g$ (where g denotes one of the 25 objects), belongs to either category. $\sigma_{Cf} = \sigma_{Cn} = \sigma_{Cd}$ was set to 1.5 as this was the true *Within-Category Feature Variability* of our two categories in either feature dimension. We will do this in terms of the posterior probability for Category A, $P(A|n, d)$, as the posterior probabilities for Category B will simply be $P(B|n, d) = 1 - P(A|n, d)$.

Table 2.1 and Figure 2.4 display the posterior probabilities that each of the 25 objects $\mathbf{f}_g$ belong to Category A. As either the Sides or Dots features increase, the probability that the object belongs to Category A decreases. This is to be expected, as Category B was defined as having more Sides and larger Dot spacing than Category A.

The diagonal line of 50% probabilities along the posterior distribution (Figure 2.4) is where the combinations of the features are equidistant to the means of the categories. These objects are equally likely to appear in either the A or B categories and should be the most confusing for observers trying to categorize them. We would expect observers to guess purely by chance for these objects which category they belong to, as they have no additional information informing them which category is more likely.

The change in posterior probability from one end of the stimulus space to the other is not linear, but sigmoidal. This can be seen by examining Equation 2.8 and noticing that it follows the common logistic function:

$$\frac{1}{1 + Ze^{-x}}$$

An important takeaway from this visualization is that the most probable A object is not the most common representative of Category A. This becomes especially apparent comparing Figure 2.4 to the category feature distributions in Figure 2.3, where the most common A object was where $\mu_{An} = 7$ and $\mu_{Ad} = 5$. This object is in fact *not* the most probable A object in our categorization task. The object with the highest posterior probability $P(A|n, d)$ is the corner object where $\mu_{An} = 6$ and $\mu_{Ad} = 4$.

2.6 Noisy vs Noiseless Bayesian Categorization Model

The model we have derived here forms the basis of our analysis throughout Experiments 1 and 2 (described in detail in Chapter 3). However, the model is missing a crucial component for it to be a rigorous observer model. The basic version of the model in Equation 2.8 only considers *Within-Category Feature Variability*, which is learned throughout the experiment and shown as the variances in Figure 2.3. We refer to this model as a *noiseless* Bayesian Categorization Model. To extend this to be a more complete model we must consider a secondary source of variability: the *Internal Measurement Noise* of the observer.

An observer does not have direct access to the *distal stimulus* they are trying to observe. Instead, this distal stimulus casts projections onto the observer's nervous system from which a proximal measurement is made. We refer to this as the Encoding process; the nervous system encodes the distal stimulus into internal measurements. However, the nervous system is not a perfect measurement system, and this proximal measurement is subject to internal sensory *Measurement Noise*. This *Measurement Noise* is introduced at every step of the measurement process, from the dynamics of sensory receptors in the fingertips up to the somatosensory cortical regions responsible for processing. This leads to uncertainty in the measurements being made.

A rigorous computational treatment of this sensory noise might involve modelling every stage of the observer's nervous system from initial skin deformation and activation of the cutaneous receptors (the rapid-adapting afferents RA1, the slowly-adapting type1 and type 2 afferents SA1 and SA2, and Pacinian Corpuscles PC), to the somatosensory cortex and beyond. Sensory noise from these afferents, and the sensory noise at further stages of this process, could be modelled as a Poisson-like spike count noise (Sripati, 2006; Tolhurst et al., 1983). However, we can simplify this modelling by instead considering the overall result of this noise. We can take advantage of the Central Limit Theorem which states that, given enough random samples from a population, the distribution of the sample means will approximate a normal distribution. Therefore, this *Measurement Noise* uncertainty can be modelled as probabilistic Gaussian measurements.

The inclusion of *Measurement Noise* in our model gives us what we will refer to as the *noisy* Bayesian Categorization Model. This model will include two parts: the Encoder model and the Decoder model.

2.6.1 Encoder model

The Encoder model is also sometimes called a *generative model*. This is the model by which the nervous system produced the proximal measurement f_m of the distal feature f, complete with both *Within-Category Feature Variability* and *Measurement Noise*. In our categorization task, this encapsulates both the process by which the stimulus object is chosen from either category and the process by which the observer obtains a sensory measurement from the object. From this sensory measurement, the observer's task is to infer which category the object was drawn from. This inference is made through the Decoder model. However, the observer is not necessarily aware of the Encoder model used to generate the measurements, and the Decoder model may vary.

$$P(f_m|A) = \sum_i P(f_m|f, A)P(f|A) = \sum_i P(f_m|f)P(f|A) \tag{2.18}$$

$P(f_m|f)$ refers to the *Measurement Noise* present throughout the nervous system, and does not depend on the category of the object. It is modelled here as a Gaussian distribution with the mean set at f and σ_{fm} representing the internal *Measurement Noise*.

$$P(f_m|f, \sigma_{fm}) = \frac{1}{2\pi\sigma_{fm}^2} \exp\left(-\frac{(fm - f)^2}{2\sigma_{fm}^2}\right) \tag{2.19}$$

With Equation 2.19, repeated exposure to the same object with feature value f would result in slightly different measurements f_m, with the spread of this noisy measurement depending on σ_{fm}.

A true ideal observer would be aware of the entire Encoder model and have information about both the external *Within-Category Feature Variability* and internal *Measurement Noise* of their own system. This may not be the case and the observer may only be partially aware of the Encoder model.

2.6.2 Decoder models

As Bayes' Rule shows (Equation 2.1), in order to determine the posterior probability of $P(A|\mathbf{f})$ the observer must first calculate the likelihoods $P(\mathbf{f}|A)$ and $P(\mathbf{f}|B)$. The possible categories provide context for what feature measurements the observer can expect, so getting different measurements assuming that they come from a particular category will yield different likelihood probabilities. We will now discuss three possible Decoder models describing how an observer may interpret their sensory measurement of a feature given the likelihood that the stimulus belonged to a particular category $P(f_m|C)$. The purpose of these sub-models is to disentangle the impact of the observer's knowledge about the *Within-Category Feature Variability* and the *Measurement Noise* on $P(f_m|C)$.

Decoder model 1

Decoder model 1 (labelled as $D1$) represents an observer which is only aware of the *Within-Category Feature Variability* and is ignorant of their internal sensory *Measurement Noise*. Upon being presented with a stimulus object, the observer measures the features **f** considering the likelihood that they would be drawn from either category's feature distribution. In other words, this observer assumes (incorrectly) that its measurement equals the true feature value. This observer assumes (incorrectly) that the trial-by-trial variation in its measurements, given a particular category, is produced entirely by the *Within-Category Feature Variability*.

$$P(f_m|C) = P(f|C) = \frac{1}{\sigma_{Cf}\sqrt{2\pi}} \exp\left(-\frac{(f-\mu_f)^2}{2\sigma_{Cf}^2}\right) \tag{2.20}$$

Decoder model 2

Decoder model 2 (labelled as $D2$) represents the counter-example to Decoder model 1; an observer which is only aware of their internal sensory *Measurement Noise* and is ignorant of the *Within-Category Feature Variability*. In effect, an observer using this decoder model would believe there are only two objects with no σ_{fC} present—one object represents each Category, $C = \mu_{fC}$. The variation that the observer measures trial-by-trial is assumed to be produced entirely by this *Measurement Noise*.

$$P(f_m|C) = P(f_m|\mu_{Cf}) = \frac{1}{\sigma_{fm}\sqrt{2\pi}} \exp\left(-\frac{(f_m-\mu_{Cf})^2}{2\sigma_{fm}^2}\right) \tag{2.21}$$

Decoder model 3

Decoder model 3 (labelled as $D3$) represents the ideal observer which is aware of both the *Within-Category Feature Variability* and their internal sensory *Measurement Noise*. This model is optimal as it matches the Encoder model (Equation 2.18).

$$P(f_m|C) = \sum_i P(f_m|f_i)P(f_i|C) \tag{2.22}$$

We can expand this equation further by substituting the Gaussian distributions for $P(f_m|f_i)$ and $P(f_i|C)$.

$$P(f_m|C) = \sum_i \frac{1}{\sigma_{f_m}\sqrt{2\pi}} \exp\left(-\frac{(f_m-f_i)^2}{2\sigma_{f_m}^2}\right) \frac{1}{\sigma_{Cf_i}\sqrt{2\pi}} \exp\left(-\frac{(f_i-\mu_{Cf_i})^2}{2\sigma_{Cf_i}^2}\right) \tag{2.23}$$

One unresolved question about the implementation of Equation 2.23 is how to treat the measurement feature $\mathbf{f}$ distributions. The stimulus set used to train participants on the categories contained 25 objects, with 5 levels per feature. Thus there are only 5 possible feature measurements that the observer might see. But is the observer aware that the feature $\mathbf{f}$ distributions are not continuous?

For the analyses present in this paper, we will assume that the observer is aware of the discrete distributions for the features. When calculating the *Within-Category Feature Variability* $P(f_i|C)$, the observer will consider only the 5 levels of each feature f_i. When calculating the internal sensory *Measurement Noise* $P(f_m|f_i)$, the observer will consider a continuous distribution for the measurement of the feature f_m, but marginalize across the 5 levels of each feature f_i.

Alternative Decoder models

The three Decoder models $D1$, $D2$, and $D3$ defined here mark three possibilities for how an observer may process the measurement $\mathbf{f}_m$ which they would decide on. D1 and D2 have only a partial understanding of the Encoder model and are consequently sub-optimal, whereas $D3$ correctly understands the full Encoder model and is consequently an optimal observer. However, this is by no means an exhaustive list of possible models. There are potentially infinite different sub-optimal Decoder models that we could generate that could theoretically be used by an observer to perform our categorization task. For instance, a *guessing* model Dr where an observer responds to every trial by simply guessing with equal probability between the two categories is a valid potential Decoder model. We could also define Decoder models which ignore one of the two features entirely (Ds and Dd). While we can include select alternative Decoder models in our analyses, we will not realistically be able to exhaustively compare every Decoder model possible.

2.6.3 Analytical Solution

The Encoder and Decoder models described above will be implemented directly through simulations in Chapter 4. It is also possible to derive an analytical solution for our Decoder model. This approach loses some validity for the context of our Experiments as it requires us to assume our *Within-Category Feature Distributions* are infinitely continuous, whereas our Categorization paradigm has a very limited feature range. Nevertheless, an analytical solution to this problem can prove insightful in understanding the dynamics of the Decoder models.

First, we can consider the Encoder model as an integral across all possible values of feature f. Each f gives rise to possible measurements f_m. By integrating the probability of each f_m given f and the probability of each f given category C over all possible f, we have a similar equation for the Likelihood $P(f_m|f)$.

$$P(f_m|C) = \int P(f_m|f)P(f|C)df \tag{2.24}$$

$$P(f_m|C) = \int \frac{1}{\sigma_{f_m}\sqrt{2\pi}} \exp\left(-\frac{(f_m - f_i)^2}{2\sigma^2_{f_m}}\right) \frac{1}{\sigma_{Cf}\sqrt{2\pi}} \exp\left(-\frac{(f - \mu_{Cf})^2}{2\sigma^2_{Cf}}\right) ds \tag{2.25}$$

After expanding these terms and rearranging, and resolving the integral as the Gaussian Distribution formula always integrates to 1, we come to the following simplification:

$$P(f_m|C) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{1}{\sigma^2_{Cf} + \sigma^2_{fm}}} \exp\left(-\frac{(\mu_f - f_m)^2}{2(\sigma^2_{Cf} + \sigma^2_{fm})}\right) \tag{2.26}$$

Here we can define a new σ^*_f, which represents the combination of the *Within-Category Feature Variability* σ_{Cf} and the *Measurement Noise* σ_{fm}. This turns out to be the sum of the two variances, which follows intuitively from the relationship between the *Within-Category Feature Distribution* and *Measurement Noise*. First, feature f is drawn from the *Within-Category Feature Distribution*. Second, a measurement of this feature f_m is drawn from a *Measurement Noise* distribution centred on feature f. As such, the final measurement is a sum of two Gaussian variables and thus forms s Gaussian distribution as well whose variance is the sum of the two individual variances.

$$\sigma^*_{Cf} = \sqrt{\sigma^2_{Cf} + \sigma^2_{fm}} \tag{2.27}$$

Substituting Equation 2.27 into Equation 2.26 gives us the following new equation which illustrates that the two-step process results in a new Gaussian distribution:

$$P(f_m|C) = \frac{1}{\sigma^*_{Cf}\sqrt{2\pi}} \exp\left(-\frac{(f_m - \mu_{Cf})^2}{2{\sigma^*_{Cf}}^2}\right) \tag{2.28}$$

Now we can expand Equation 2.28 into two dimensions to accommodate the features of our category design, Sides n and Dots d. Following a similar approach for a 2-Category/2-Feature task as in Equations 2.6 through 2.8, we can derive a compact formula for the posterior probability $P(A|n_m, d_m)$:

$$P(A|n_m, d_m) = \frac{1}{1 + \exp\left(\Delta_n \frac{M_n - n_m}{\sigma^{*2}_n} + \Delta_d \frac{M_d - d_m}{\sigma^{*2}_d}\right)} \tag{2.29}$$

where $M_f = \frac{\mu_{Af} + \mu_{Bf}}{2}$, and $\Delta_f = \mu_{Af} - \mu_{Bf}$. From here, we can see an important dynamic of our Encoder and Decoder models; when the term in the exponential equals 0, $P(A|n_m, d_m) = 0.5$. When this term is > 0, Equation 2.29 will favour *Noek*, and then this term is < 0, Equation 2.29 will favour *Elyk*. Thus, we can consider this term as a Decision variable whose solution tells us immediately how an observer using our Decoder model should respond in a categorization task.

$$D = \Delta_n \frac{M_n - n_m}{\sigma_n^{*2}} + \Delta_d \frac{M_d - d_m}{\sigma_d^{*2}} \tag{2.30}$$

With Equation 2.29 we can now very quickly determine how increasing the *Measurement Noise* might change how an observer categorizes different objects in our 2-Category/2-Feature task.

2.7 Conclusion

Here we have described in detail the Bayesian Categorization Model which we will focus on exploring through the rest of this thesis. Chapter 3 will explore the efficacy of the stimulus set developed for this categorization task in a human experimentation setting and evaluate a *noiseless* Bayesian Categorization Model against human performance. Chapter 4 will explore the introduction of *Measurement Noise* through both theoretical simulations and a sensory measurement Two-Interval Forced-Choice (2IFC) experiment, finally re-evaluating the observational data from Chapter 3 with a more complete observer model.

2.8 Figures and Captions

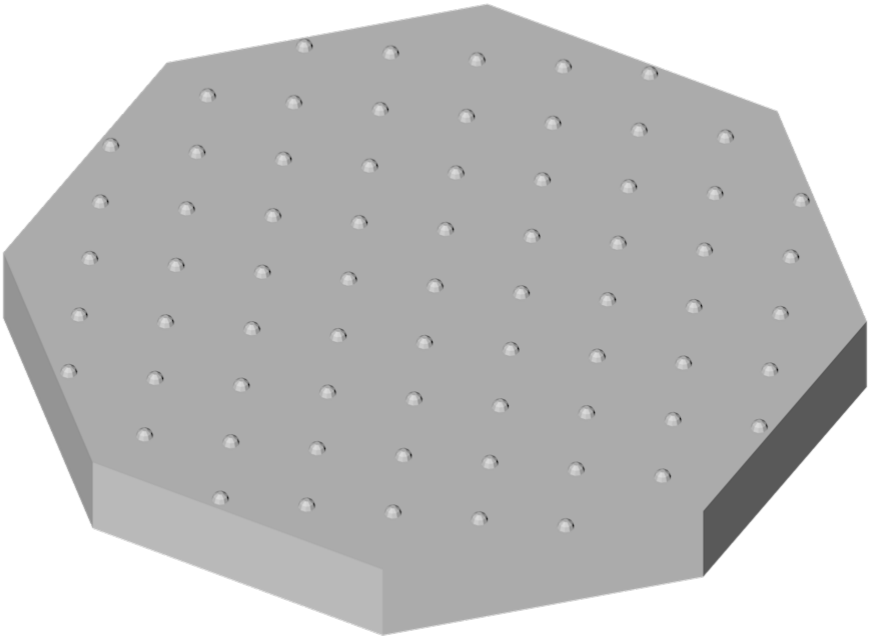


FIGURE 2.1: 3D render of the stimuli created and used for our haptic categorization task. Objects could have different numbers of sides and spacing of raised hemispherical dots on their top surface. All of the objects were generated using the program *openSCAD* and physically created using the *Ultimaker 2 Go* 3D printer in PVC plastic.

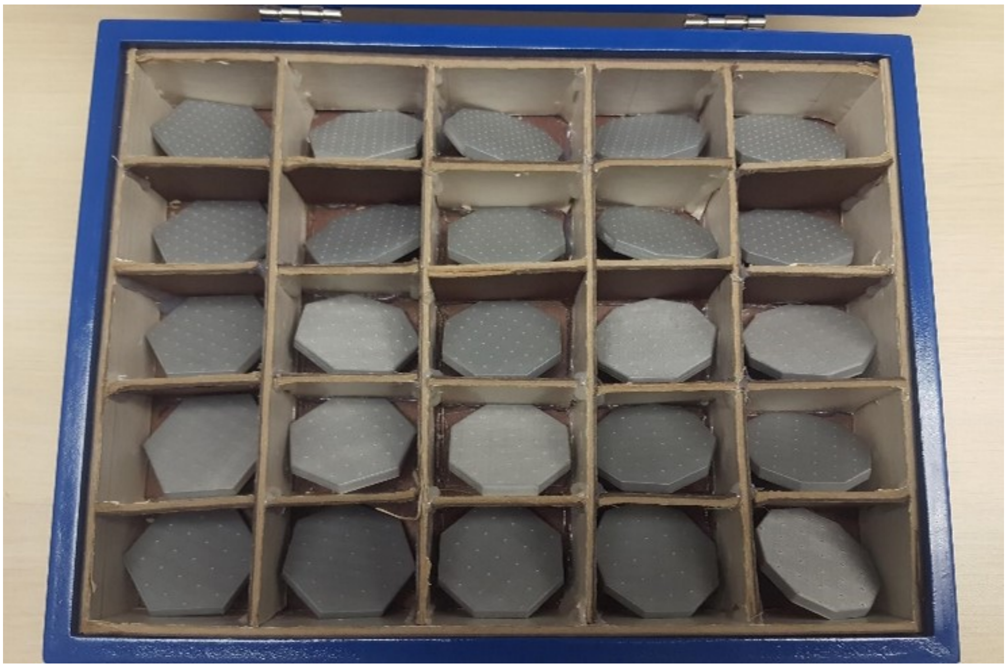


FIGURE 2.2: Box of stimuli used throughout participant experiments described in this paper. The box contained 25 plastic objects each in a smaller drawer. The objects were organized in a 5x5 grid. Objects were arranged according to their sides feature along the columns and their dots feature along the rows.

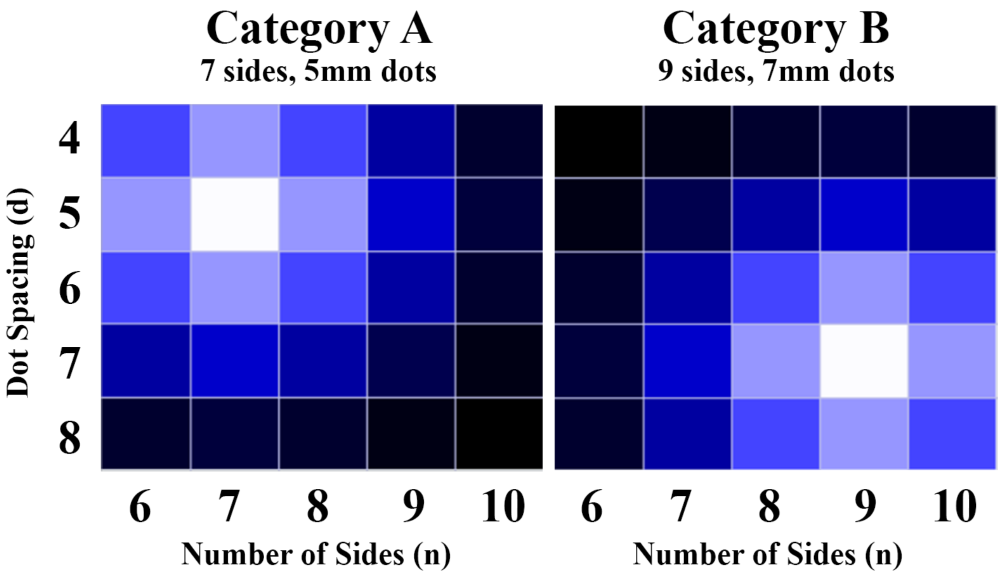


FIGURE 2.3: Two-dimensional intensity plot displays of the two category feature distributions. This can be calculated using Equation 2.7. Category A was defined as having a mean of 7 Sides and 5mm Dot Spacing with a 1.5 standard deviation in either feature dimension. Category B was defined as having a mean of 9 sides and 7mm Dot Spacing with a 1.5 standard deviation in either feature dimension. Note that the Dots feature was generated according to the space between each dot measured in millimetres and is coded accordingly, however, this is also directly proportional to the Number of Dots or Dot Density present. These are organized to match the arrangement of our stimulus object set seen in Figure 2.2.

TABLE 2.1: Table of Posterior Probability for Category A $P(A|n,d)$ for each of the 25 objects, calculated using Equation 2.8. This is organized to match the arrangement of our stimulus object set seen in Figure 2.2.

		Number of Sides				
		6	7	8	9	10
Dot Spacing	4	0.972	0.935	0.855	0.709	0.500
	5	0.935	0.855	0.709	0.500	0.291
	6	0.855	0.709	0.500	0.291	0.145
	7	0.709	0.500	0.291	0.145	0.065
	8	0.500	0.291	0.145	0.065	0.029

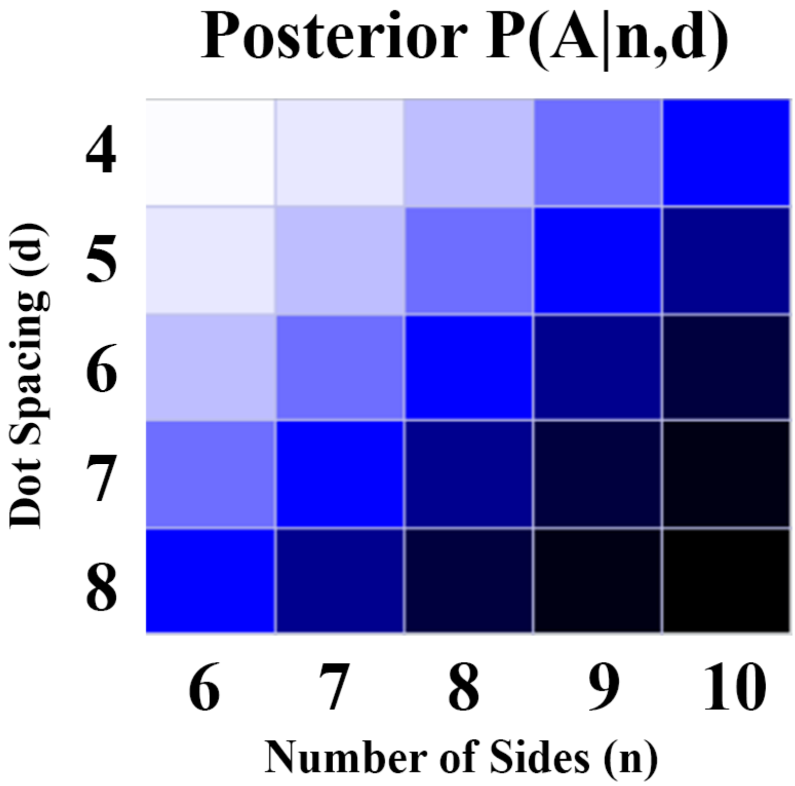


FIGURE 2.4: Intensity plot of the Posterior Probability$P(A|n,d)$ for *Category A* calculated using Equation 2.8 arranged in a 5x5 grid. The brightness intensity of each square is directly proportional with the probability, with the highest probability being 0.972 (located on the upper-leftmost square). This is organized to match the arrangement of our stimulus object set seen in Figure 2.2.

Table 2.2: Table of the 25 object stimuli used throughout our experiments. Listed is each object's label $\mathbf{f}_g$ as well as their associated Sides and Dots features. Also listed are the likelihoods for $P(n, d|A)$ and $P(n, d|B)$ calculated using Equation 2.7, where $\sigma_{Cf} = 1.5$ for each feature, $\mu_{An} = 7$ and $\mu_{Ad} = 5$ for the Category A *Within-Category Feature Means*, and $\mu_{Bn} = 9$ and $\mu_{Bd} = 7$ for the Category B *Within-Category Feature Means*. The final column lists the posterior $P(A|n, d)$ calculated using Equation 2.8 for each object.

Object g	$\mathbf{f}_g$	Sides n	Dots d	$P(n, d\|A)$	$P(n, d\|B)$	$P(A\|n, d)$
1	$\mathbf{f}_1$	6	4	0.045	0.001	0.972
2	$\mathbf{f}_2$	6	5	0.057	0.004	0.935
3	$\mathbf{f}_3$	6	6	0.045	0.008	0.855
4	$\mathbf{f}_4$	6	7	0.023	0.01	0.709
5	$\mathbf{f}_5$	6	8	0.008	0.008	0.5
6	$\mathbf{f}_6$	7	4	0.057	0.004	0.935
7	$\mathbf{f}_7$	7	5	0.071	0.012	0.855
8	$\mathbf{f}_8$	7	6	0.057	0.023	0.709
9	$\mathbf{f}_9$	7	7	0.029	0.029	0.5
10	$\mathbf{f}_{10}$	7	8	0.01	0.023	0.291
11	$\mathbf{f}_{11}$	8	4	0.045	0.008	0.855
12	$\mathbf{f}_{12}$	8	5	0.057	0.023	0.709
13	$\mathbf{f}_{13}$	8	6	0.045	0.045	0.5
14	$\mathbf{f}_{14}$	8	7	0.023	0.057	0.291
15	$\mathbf{f}_{15}$	8	8	0.008	0.045	0.145
16	$\mathbf{f}_{16}$	9	4	0.023	0.01	0.709
17	$\mathbf{f}_{17}$	9	5	0.029	0.029	0.5
18	$\mathbf{f}_{18}$	9	6	0.023	0.057	0.291
19	$\mathbf{f}_{19}$	9	7	0.012	0.071	0.145
20	$\mathbf{f}_{20}$	9	8	0.004	0.057	0.065
21	$\mathbf{f}_{21}$	10	4	0.008	0.008	0.5
22	$\mathbf{f}_{22}$	10	5	0.01	0.023	0.291
23	$\mathbf{f}_{23}$	10	6	0.008	0.045	0.145
24	$\mathbf{f}_{24}$	10	7	0.004	0.057	0.065
25	$\mathbf{f}_{25}$	10	8	0.001	0.045	0.028

Chapter 3

Human Performance In A Two-Category Learning Paradigm

3.1 Introduction

Haptic exploration refers to the use of the tactile system to extract sensory information about an object and its features. Through physical manipulation and exploration, an observer can learn the underlying statistics of categories, allowing an explored object to be categorized (Klatzky et al., 1993). This process often involves the full suite of sensory systems including vision and audition to quickly and accurately distinguish different objects. But touch can work alone to extract enough information to make accurate categorizations, both in sighted and blind individuals (Klatzky et al., 1985; Norman et al., 2011; Norman et al., 2015). For example, reaching into a mixed bag of various-sided polyhedral dice would reveal a set of similar-feeling objects, yet an observer is capable of distinguishing an 8-sided die from a 20-sided die.

Prior research has explored the ability of the tactile system to extract feature information from objects both alone and in cooperation with the other senses. Klatzky et al. (1993) observed several stereotyped procedures that human observers tend to use when exploring objects, with each procedure optimized to extract specific information. For instance, *Lateral Motion*, described as repetitive surface rubbing, extracts information about texture, while *Enclosure* or *Contour Following* extracts information about the global shape of an object. They also found that haptic exploration is naturally guided by vision when visual information is present; a visual preview of an object provided coarse information that helped direct further exploration using touch. However, the information extracted from the senses does not need to overlap and can even be processed from different frames of reference (Newell et al., 2001). When an observer grasps an upright object, touch naturally provides information about the back side while vision provides information about the front side.

Bayesian Categorization with Two Known Categories

In a typical two-categorization task, the objective of the observer is to determine the posterior probability of the category of an explored object given its feature measurements $\mathbf{f}_m$ (bolded $\mathbf{f}$ referring to a set of features). The observer is assumed to know there are two categories and the features to measure. We can write this posterior probability as $P(A|\mathbf{f}_m)$ (written in terms of category A, with the posterior for category B being $1 - P(A|\mathbf{f}_m)$). The observer gathers information about the stimulus through sensory inputs. If the observer is assumed to know the two categories ahead of time, they are able to calculate the likelihoods (i.e., the probabilities) of measuring the features $\mathbf{f}_m$ given either category, $P(\mathbf{f}_m|A)$ and $P(\mathbf{f}_m|B)$. Refer to Chapter 2 for a full derivation of our Bayesian Categorization Model.

The underlying categorical structure is assumed to be understood by the observer as Gaussian probability distributions. We designed stimuli to have only two distinguishing features (n and d) and to be identical in every other respect. We also set the *Within-Category Feature Variability* of each dimension and each category to be equal and independent.

We assume that the observers apply an equal prior between categories A and B, $P(A) = P(B) = 0.5$, according to Laplace's *Principle of insufficient reason*. This allows us to use Equation 2.8 as our *noiseless* Bayesian Categorization Model observer. In the present study's experimental design, this assumption was reinforced by randomly drawing objects from either category with equal probabilities.

In the present study, we investigated human category learning of objects using only the sense of touch and compared results to our Bayesian model. The stimuli were a set of novel 3D-printed polygonal objects with pseudo-continuous features. Categories were artificially defined out of this set of objects as 2D Gaussian distributions. With this study, we aimed to investigate three questions: 1) Is human haptic categorization Bayes-optimal? 2) Which training method is most effective for category learning? and 3) Does categorization improve over longer-term training?

3.2 Experiment 1

The purpose of Experiment 1 was to explore human performance in a two-category learning paradigm using tactile exploration of physical objects. Two categories were defined and objects were artificially drawn from these categories using computer software. Participants were physically handed objects corresponding to the ones drawn by the software one at a time to freely explore. The goal was to train participants on the two underlying categories to the point of reaching their highest possible accuracy in a categorization task. A set of novel objects was created for the purposes of this experiment.

In order to test different methods for training participants on the categories, three separate training regimens were established: block training with feedback, block training without feedback, and continuous testing with feedback. The goal was to determine the effectiveness of free exploratory learning (where observers had a chance to explore many objects from a category where they knew which category the objects belonged to) against corrective feedback learning (where observers learned while trying to categorize unknown objects).

3.2.1 Methods

Participants

Forty-five participants (18-26 yr old, median age 19 yr) were recruited from the McMaster University community. Exclusion criteria included any of the following conditions known to adversely affect tactile acuity or performance on psychophysical tasks: diabetes, nervous system disorders, cognitive impairment, learning disabilities, dyslexia, attention deficit disorder, carpal tunnel syndrome, and arthritis of the hands. All participants were naïve to the experiment and the unique stimuli used. Participants were compensated with either course credit or $20. Participants completed a short demographic and handedness questionnaire prior to participation, and an exit interview after the experiment concluded.

Materials and Stimuli

The stimulus objects were created using an Ultimaker 2 Go 3D printer. The objects were designed using the open-source Computer-Aided Design (CAD) modeller OpenSCAD. They were designed to be similar polygonal prisms with only two differentiating features—the number of sides and the spacing of a dot-matrix texture that covered one of the faces. All the objects were otherwise identical with a radius of $25mm$ and a thickness of $5mm$.

The dot-matrix texture consisted of evenly spaced $0.5mm$ radius hemispheres protruding from the surface. We established 5 different levels for dot spacing and 5 levels for sides, creating 25 unique objects. The center-to-center spacing between neighbouring dots on an object ranged from $4mm$ to $8mm$ in $1mm$ increments. The number of sides of each object ranged from 6 to 10 sides in increments of 1 side (see Figures 2.1 and 2.2).

Object Categories

Two categories were defined using the twenty-five objects detailed in Chapter 2. These categories were defined by symmetric, 2D Gaussian distributions with a set mean number of Sides and Dot Spacing, and a Standard Deviation (σ) of 1.5. The first category was set at a mean of 7 Sides and $5mm$ Dot spacing. The second category was set at a mean of 9 Sides and $7mm$ Dot Spacing. Participants were given different names to distinguish these categories. For half of the participants, the categories were named *Elyk* and *Noek* respectively, and for the other half, these names were reversed. These pseudo-words were chosen to be phonetically similar while avoiding any possible confounding effects of previous biases to real words or effects of ordering (for instance, naming the categories 'A' and 'B' might imply a sense order as A comes before B). The Gaussian distribution parameters were chosen in such a way that the *noiseless* Bayesian Categorization Model observer would be correct on average 76% of the time.

Throughout this chapter, we will refer to the first object category with means of 7 Sides and $5mm$ Dot spacing simply as Category A, and the second object category with means of 9 Sides and $7mm$ Dot Spacing as Category B.

As discussed in Chapter 2, these categories are not truly continuous Gaussian distributions. This is a physical limitation; we could not practically produce a continuous range of stimuli across Dot spacing, and the number of Sides of a polygon is constrained to be integers. Furthermore, the distributions are clipped at the ends and consequently are more similar to truncated Gaussian distributions. Regardless, we believe the range of stimuli presented to participants is sufficient to convey the underlying ideal distributions.

We wanted to simulate the effect of physically drawing random objects from a box with an equal number of Category A and B objects mixed in together without having to 3D print multiple copies of each object. In reality, and unknown to the participants, we had only a single 3D print of each of the 25 objects. During the experiment, the computer randomly (with 50% probability) chose to present either a Category A or B object, and, having made that choice, randomly sampled an object from the corresponding category distribution.

Software

The experiment ran on a 2006 iMac using a custom program written in LabVIEW 2017 (National Instruments). While remaining hidden from the participant, this software managed the progress of the experiment and data collection. A human experimenter was required to present and retrieve stimuli and input participant responses, but the computer handled sorting participants pseudo-randomly into different training regimen groups, selecting an object every trial, and providing corrective feedback to the participant.

A GoPro Hero 4 camera was used throughout the experiment to record participants' hands. This camera was mounted on an articulating arm and pointed down for a top-down view. Video capture was handled automatically by the experiment software. One out of every 5 trials was recorded and saved as a separate .mp4 video file.

Experiment instructions were pre-recorded on audio files and played for the participant before beginning the experiment.

Procedure

At the start of the experiment, participants were played a pre-recorded script. All participants were told that their task was "to try to learn two different categories of objects using only touch". They were told the names of the two categories, *Elyk* and *Noek*, and that they "differ by both shape and the density of dots on top". They were explicitly told to try to pay attention to both the Shape and Dots features. They were also told that "half the objects are *Elyk* and half are *Noek*". The script varied slightly depending on their assigned training group as the instructions included a description of any training blocks and corrective feedback.

Unknown to the participants, they were pseudo-randomly assigned to one of three training regimens. Participants assigned to the training regimen Group 1 completed a set of two training blocks and a testing block with corrective feedback. Participants assigned to the training regimen Group 2 completed a set of two training blocks and a testing block without corrective feedback. Participants assigned to the training regimen Group 3 completed a set of three testing blocks with corrective feedback. As participants in Group 3 were given no prior training or exposure to the stimuli and categories, it was expected that they would be guessing the first few trials. This was conveyed to them during the pre-recorded script to avoid confusion.

Each block comprised 40 trials, and the set of three blocks was repeated three times for a total of 9 blocks (360 trials total). In each trial, the participant was given an object to explore for five seconds. The experiment took a maximum of 2.5 hours to complete.

Participants were counterbalanced for the categories naming convention; for half the participants, *Elyk* meant an average of 7 sides and 5mm spacing while *Noek* meant an average of 9 sides and 7mm spacing, and for the other half, *Elyk* and *Noek* meant the opposite. Additionally, we also counterbalanced the order of training blocks (only applicable for participants in Groups 1 and 2). This led to four possible combinations of naming conventions and training block order. These counterbalances controlled against possible effects of learning order or the object category names.

During a training block, the participant was told what category they were learning from for the entire block. The computer randomly drew with replacement from the indicated category and instructed the experimenter to present that object.

During the testing block, participants were given an object and told it was randomly pulled from either category. The computer first randomly chose between Category A and B, and then randomly drew an object from the chosen category. It then instructed the experimenter to present that object. After five seconds of exploration, the participant reported which category they thought the object belonged to (using the naming convention of *Elyk* or *Noek*) and their confidence in their response from 50%-100%.

In the testing block with corrective feedback, the participants received auditory feedback and so learned on every trial whether their answer was correct or incorrect. In the testing block without corrective feedback, participants did not receive this feedback and did not know whether their response was correct or incorrect.

In an exit interview at the end of the experiment, the experimenter asked the participants five questions: Did they think they knew how *Noek* and *Elyk* differed (and if so, how), did they pay attention equally to the two features or more to one than the other, did the categories become clearer as they progressed, did they have any strategy or rule, and how confident were they that they understood how *Noek* and *Elyk* differed.

3.2.2 Results

The primary aim of this experiment was to explore human categorization capabilities for objects learned through the tactile modality. A secondary aim was to investigate the efficiency of various training methods. To this end, three different training regimens were compared.

In Experiment 1, forty-five participants were split across three training Groups. Across the Groups were three testing blocks where performance could be compared; Blocks 3, 6, and 9. Participant responses on these three testing blocks were calculated against that of a *noiseless* Bayesian observer (one which does not consider *Measurement Noise*, using Equation 2.8) whose responses matched the decision made by the *maximum a posteriori* (MAP) estimate of the model. Overall, participants in Group 3, who were given repeated testing blocks with corrective feedback), outperformed the participants in Groups 1 and 2.

For all analyses, unless otherwise stated, we excluded trials where one of five object stimuli were presented to the participant: $\mathbf{f}_5, \mathbf{f}_9, \mathbf{f}_{13}, \mathbf{f}_{17}$, and $\mathbf{f}_{21}$ (refer to Table 2.2). These objects were all those in which our *noiseless* Bayesian Categorization Model calculated a posterior probability $P(A|n_m, d_m) = 0.5$. The Bayesian Categorization Model cannot reasonably categorize these objects and would have to simulate "guessing" by choosing randomly. This prevents meaningful trial-by-trial comparison between humans and our model on those trials.

By Block 9 of the experiment, Group 1 reached an average percent correct of 65.6% (SD=13.5), Group 2 reached an average of 64.5% (SD=14.7), and Group 3 reached an average of 73.2% (SD=7.9) (Figure 3.1). Looking at the average responses between Category A and B (where Category A was coded as 1 and Category B was coded as 2), Group 1's average response was 1.47, Group 2's average response was 1.52, and Group 3's average response was 1.52, indicating no significant bias between the categories. Performing a repeated-measures ANOVA between the three Groups and across Blocks 3, 6, and 9 revealed a significant effect of Group, $F(2, 42) = 3.634$, $p = .035$. Mauchly's Test of Sphericity indicates the sphericity assumption had not been violated, $X^2 = 0.871$, $p = .059$.

The average confidence reported by participants in Groups 1, 2, and 3 were 68.1% (SD=12.6), 72.9% (SD=10), and 66.7% (SD=13.5) respectively. A Repeated Measures ANOVA revealed no significant effect of confidence ratings between Groups, $F(2, 42) = 1.531$, $p = .228$. Mauchly's Test of Sphericity indicated that the sphericity assumption had been violated, $X^2 = 14.842$, $p < .001$, and so we corrected for sphericity using the Greenhouse-Geisser correction.

Recalculating the data in terms of the percent of trials matched to the observer model and performing a Repeated Measures ANOVA revealed a significant effect of Group, $F(2, 42) = 7.187$, $p = .002$, as well as a significant effect of Block, $F(2, 84) = 4.961$, $p = .009$. By Block 9, Group 1 reached a percent match of 74.2% (SD=14.9), Group 2 a percent match of 71.9% (SD=14.2), and Group 3 a percent match of 83.1% (SD=7.6).

The average percent correct of the five excluded object stimuli ($\mathbf{f}_5$, $\mathbf{f}_9$, $\mathbf{f}_{13}$, $\mathbf{f}_{17}$, and $\mathbf{f}_{21}$) was between 49% and 50% across the three groups by Block 9, indicating that participants were guessing when presented with these objects as predicted by *noiseless* Bayesian Categorization Model. The average confidence for these five objects by Block 9 was 67.1% for Group 1, 69.3% for Group 2, and 64.4% for Group 3, potentially indicating a bias in self-reported confidence.

To examine whether participants exhibited a feature bias between Sides and Dots, we can examine the average response of objects $\mathbf{f}_9$ and $\mathbf{f}_{17}$ to see if they are skewed towards either category. Human observers may not always respond optimally and there could be many potential sources for this sub-optimal behaviour. A feature bias would be equivalent to the observer incorrectly assuming the *Within-Category Feature Variability* was greater for one feature than the other, and this would have the effect of down-weighting that particular feature in the inference (as shown in Equation 2.9). In terms of objects $\mathbf{f}_9$ and $\mathbf{f}_{17}$, a Dots feature bias would result in $\mathbf{f}_9$ being more often categorized as Category A and $\mathbf{f}_{17}$ being more often categorized as Category B, while a Sides feature bias would result in the reverse. If a "Category A" response is coded as 1, and a "Category B" response is coded as 2, we can take the average of all responses as a coarse indicator for response bias, where an average response of 1.5 would indicate no bias presence. The average responses for object $\mathbf{f}_9$ for Groups 1, 2, and 3 were 1.4, 1.45, and 1.29, while the average responses for object $\mathbf{f}_{17}$ were 1.5, 1.54, and 1.62. This could indicate a bias for the Dots feature, as $\mathbf{f}_9$ was more often categorized as A and $\mathbf{f}_{17}$ more often categorized as B. This bias can also be visually appreciated in Figure 3.5.

The data was further analyzed to compare the frequency of category responses per each of the 25 objects against the posterior probability of each object belonging to Category A (Figure 3.2). This was done for three different sub-models of our Bayesian Categorization Model: one that only considers the Sides feature, one that only considers the Dots feature and one that considers both features. Linear regressions were calculated for each of these models per Group. In Groups 1 and 2, the Dots-only observer model had a slope closest to 1. In Group 3, the expert model had the slope closest to 1.

This object-wise analysis was also calculated for the participants' reported confidence that the object belonged to Category A. Since participants were only asked to report their confidence in their response for each trial, the participant's confidence was in respect of Category A in trials where they responded "A", and in respect of Category B in trials where they responded "B". In order to put everything in terms of Category A, the confidence in trials where the participant responded "B" was recalculated as $1 - confidence$. Performing a linear regression against the posterior probability of each object belonging to Category A reveals a slope of approximately 0.33 for all three Groups (Figure 3.3).

3.3 Experiment 2

The purpose of Experiment 2 was to investigate the effect of a longer training period. A key assumption in analyzing Experiment 1 was that participants had fully learned the categories by the end of the experiment and were performing at their highest possible performance by the last testing block. By tracking performance across five consecutive testing days, we aimed to determine if this was a valid assumption. In Experiment 2, only the third training regimen was used as it was determined through Experiment 1 to lead to the best categorization performance. The daily testing procedure was the same as that used in Experiment 1, except that a slight modification was made on days 2 through 5 with the inclusion of two additional objects.

3.3.1 Methods

Participants

Ten participants were recruited from McMaster University (18-23 yr old, median age 20.5 yr). Participants were screened in the same way as participants in Experiment 1 and were naïve to the experiment and the unique stimuli used throughout. Participants were compensated $20 per day. Participants completed a short demographic and handedness questionnaire prior to the start of the experiment, and an exit interview after the experiment concluded.

Materials and Stimuli

The same twenty-five stimuli from Experiment 1 were used in Experiment 2 with the same category distributions. Additionally, two new objects were introduced from the second day onward with parameters outside of the range previously seen. One object had 11 sides and $7mm$ dot spacing and the other had 9 sides and $9mm$ dot spacing. The objects were chosen to have the same theoretical probability of belonging to Category B as object $\mathbf{f}_{25}$ (10 sides and $8mm$ dot spacing). The experiment was run on the same custom program as Experiment 1.

Procedure

All participants were trained under the training regimen of Group 3 which consisted of testing blocks with corrective feedback. The first day of the experiment was identical to the procedure in Experiment 1, consisting of 9 blocks of 40 trials taking a maximum of 2.5 hours to complete. Participants were played the same pre-recorded script used for Group 3 participants in Experiment 1 at the start of each day. After the first day, two modifications were made: First, the two new objects were discretely included in the experiment using pseudo-randomization that ensured that each was presented once within the block. Trials with these objects provided no corrective feedback to the participant. In this way, participants would not be able to learn from these objects and their understanding of the underlying categories should not have changed.

Second, each trial was given a 10% chance to not give the corrective feedback to the participant, regardless if the object was from the original set or not. This was done to blind the participant as to which trials included these two new objects. Participants were informed that some trials from the second day onward would not give corrective feedback. They were not told that new objects were being added to the experiment.

The experiment was run over five consecutive days, 2.5 hours each day, for a maximum of 12.5 hours. The same exit interview from Experiment 1 was held at the end of the last day of the experiment.

3.3.2 Results

The aim of Experiment 2 was to determine the maximum performance attainable by participants in our categorization task. Ten participants were tested over five consecutive days (two hours each day, 45 blocks in total). One participant's data was removed from the analysis as their performance fell significantly below all other participants with a total average of 44.9% across the five days of the experiment. Trials where the participants were expected to guess randomly (where $P(A|n_m, d_m) = 0.5$) were also removed from this analysis.

A table of the results calculated on the last block of each day can be found in Table 3.1. A Repeated Measures ANOVA revealed no significant effect of experiment day on average percent correct, $F(4, 28) = 0.396$, $p = .810$. The sphericity assumption was tested using Mauchly's Test which was not significant, $X^2 = 11.036$, $p = .295$. Figure 3.7 shows the average percent match to the model across the five-day experiment, with 81.7% (SD=7.8) on Day 1, 85.3% (SD=7.5) on Day 2, 79.4% (SD=11.2) on Day 3, 81.2% (SD=11.3) on Day 4, and 77% (SD=9.4) on Day 5.

A Repeated Measures ANOVA over the average confidence ratings across the five did not reveal a significant effect of experiment day. Mauchly's test indicated that the assumption of sphericity had been violated, $X^2 = 22.520$, $p = .010$, and so we corrected with the Greenhouse-Geisser correction, $F(1.664, 11.647) = 0.443$, $p = .618$.

Performance was calculated as the percentage of trials in which the participant responded similarly to our *noiseless* Bayesian Categorization Model (Equation 2.8). A Repeated Measures ANOVA found no significant effect of experiment day, $F(4, 28) = 1.347$, $p = .277$. The sphericity assumption tested using Mauchly's Test was not violated, $X^2 = 9.222$, $p = .439$. In general, performance reached a similar level to that of participants in Experiment 1 after the first day of the experiment and did not improve with subsequent days of training.

We again reanalyzed the data in terms of observer response behaviour for each of the 25 objects. The frequency of each object being categorized as A was compared against the posterior probability of that object belonging to the A category as calculated by Equation 2.8. Figure 3.8 shows the result of this analysis with a calculated slope close to 1, indicating a strong match between response behaviour and the output of the observer model.

The total average percent correct of the five excluded object stimuli ($\mathbf{f}_5$, $\mathbf{f}_9$, $\mathbf{f}_{13}$, $\mathbf{f}_{17}$, and $\mathbf{f}_{21}$) across all trials was 57%, indicating that participants were likely guessing when presented with these objects as predicted by *noiseless* Bayesian Categorization Model.

Two additional objects were introduced randomly throughout days 2 through 5 in order to test how participants would categorize new objects drawn from the categories. This tested the possibility that participants were not learning the overall categories and were instead learning the categories of each object individually. If this were the case, then participants would be expected to guess with a 50% probability when presented with the new objects. This was not observed in our results; participants had no problem categorizing the novel objects and categorized them as we would expect according to an understanding of the underlying statistics of the categories. The average percent correct for both objects was 87.4%. These two objects were comparable to $\mathbf{f}_{25}$ according to our *noiseless* Bayesian Categorization Model, which had a total average percent correct of 92.6%.

Experiment 2 demonstrated that participants adequately learned the categories after a single day of the experiment. This offers validity to Experiment 1 and the assumption that participants had been trained sufficiently enough on the categories to compare their response behaviour to our model.

3.4 Advanced Analysis Approaches

The analysis shown in Figures 3.2 and 3.8 are compelling as they allow for a direct comparison between the participants' response behaviour and the expected output of our Bayesian Categorization Model. However, this does not quite follow what we would expect from our *noiseless* Bayesian Categorization Model. Our observer uses a MAP estimate to make categorical decisions, leading to a deterministic response behaviour where its response would be the same for any given object. This would result in a step function when plotting $P(''A''|\mathbf{f}_g)$ against $P(A|\mathbf{f}_g)$ where:

$$P(''A''|\mathbf{f}_g) = \begin{cases} 1 & P(A|\mathbf{f}_g) > 0.5 \\ 0 & P(A|\mathbf{f}_g) < 0.5 \end{cases} \tag{3.1}$$

This is not what we observed in our human participants. Instead, we found variability in response behaviour for how each object was categorized. There are a few factors which could contribute to the stochasticity of human performance in our experiment, one of which being the presence of *Measurement Noise* which we will explore in detail in Chapter 4. An alternative possibility we can immediately apply to this analysis is that of *Bayesian Sampling*. Under this method, an observer responds in such a way that is proportional to the underlying posterior probability; if a particular object had a 75% posterior probability for Category A, an observer using *Bayesian Sampling* would respond "A" 75% of the time. Applying this to our analysis, as the posterior probability for an object belonging to Category A $P(A|\mathbf{f}_g)$ increases we'd expect a higher probability of the observer responding "A", $P(''A''|\mathbf{f}_g)$. This was seen in both Experiments 1 and 2 with linear regressions revealing slopes close to 1.

There is a major flaw, however, with this analysis. To achieve this object-wise analysis, the response frequency of Category A needed to be calculated for each of the 25 objects. Each object was only shown a handful of times throughout the experiment for a single participant, and some particular objects may have only been shown very few times or not at all. Figure 3.4 shows a histogram of the frequency each object was drawn, where we can see this discrepancy in stimulus presentation across our participant pool. As such, there was not enough data within a single participant to calculate the frequency of Category A responses per object. To calculate this, all participant responses had to be aggregated together. While still informative, this limitation in the analysis makes it difficult to draw conclusions.

To address this, we attempted a more advanced analysis procedure taking advantage of the probabilistic nature of our model. Consider the posterior probability space $P(A|\mathbf{f}_g)$. Each object has a posterior probability as calculated by our Bayesian Categorization Model in Figure 2.9. Furthermore, we will use a *Bayesian Sampling* method where $P(A|\mathbf{f}_g) = P(''A''|\mathbf{f}_g)$ for our *noiseless* Bayesian Categorization Model. While we cannot directly measure the posterior $P(A|\mathbf{f}_g)$ in our human participants, we can measure $P(''A''|\mathbf{f}_g)$.

The main question in our analysis is whether or not the probability space for the 25 objects described by our model matches the posterior probability $P(A|\mathbf{f}_g)$. We can consider the simple linear regression model $y = mx$ as a mapping function, relating the posterior $P(A|\mathbf{f}_g)$ in the x-axis to the probability response $P(''A''|\mathbf{f}_g)$ in the y-axis. In the case that these two probability spaces map onto each other perfectly, we expect a slope along the identity line $x = 1$. Thus our linear regression model will be the following:

$$P(''A''|\mathbf{f}_g) = xP(A|\mathbf{f}_g) \tag{3.2}$$

Consider that participant responses are binary, either *Elyk* or *Noek*. As such, we can treat each trial as if it is a coin flip, where the probability of "landing heads" is given by $P(''A''|\mathbf{f}_g)$. We can expand this for all trials in a single run through the experiment by employing the probability "and" rule to find the likelihood of all of our observed data given a particular slope x:

$$P(D|x) = \prod_t^n P(''A''|\mathbf{f}_g)_t = \prod_t^n p_t^{r_t}(1-p_t)^{1-r_t} \tag{3.3}$$

where D is all of the participant's responses in the experiment, n is the total number of trials in the experiment, t is a particular trial, and r_t is the response of the participant on that trial. On trials where the participant responds "A", $r_t = 1$, and on trials where the participant responds "B", $r_t = 0$. p_t is the critical modification for our analysis; this is the probability of the participant responding "A" as given by the model in Equation 2.9 for each particular trial. As such, it will depend on the feature measurements of the given object handed to the participant on that trial.

$$p_t = xP(A|n_m, d_m)_t \tag{3.4}$$

Finally, we can apply Bayes' formula with a uniform Prior to calculate an estimate for a particular slope x_i given our observed data D for each participant. Using this logic we can sidestep the issue of data resolution we were facing previously and recalculate our object-wise analysis for each participant. Objects that are presented fewer times will be naturally down-weighted in their influence over the estimate for the slope.

$$P(x_i|D) = \frac{P(D|x_i)}{\sum_k P(D|x_k)} \tag{3.5}$$

Using Equation 3.5 we calculated the posterior probability density function for m across a range of slopes x from -1 to 1 for each participant across each block of Experiment 1. Figure 3.6 shows the most probable slope, corresponding with the mode of the posterior distribution, across the three Groups and Blocks of Experiment 1. Participants in Group 3 reached an average slope of 1.02 by the end of the experiment, while participants in Groups 1 and 2 only reached slopes of 0.67 and 0.71 respectively. This confirms what we had found with our previous flawed analysis more rigorously: participants in Group 3 matched closest to the Bayesian observer model with a slope approaching 1 by the end of the experiment.

One issue with this analysis is our reliance on the use of *Bayesian Sampling* for our model's decision-making process to categorize objects. As we will discuss later in the discussion section of this chapter and in Chapter 5, this is an uncertain assumption for us to make. However, because there is response variability present in our participants' responses, the MAP estimate does not work with this analysis. Our analysis discussed here could be valid with a MAP estimator observer under the presence of *Measurement Noise*. However, in its current state, our model is unable to account for this factor. We will discuss the presence of *Measurement Noise* further in Chapter 4.

Model Comparison Analysis

The analysis described above attempted to iterate upon the analysis shown in Figures 3.2 and 3.8 in a way that was not dependent on the frequency of responses for each object. However, we can expand on this idea even further to answer another important question in our data set: are participants ignoring either the Sides or Dots features or combining them in their categorical judgements? Rather than assuming a linear or sigmoidal relationship between participant response data and our model's posterior $P(A|\mathbf{f}_g)$ and estimating a slope parameter, we can instead compare different models for p_t. We can then use model comparison to determine which model best fits each participant's data.

We will assume four hypothetical models: an *Expert* model which includes both features, a *Sides-only* model that ignores the Dots feature, a *Dots-only* model that ignores the Sides feature, and a *Guessing* model which responds randomly with an equal probability of $p_t = 0.5$. The three other models will use different configurations of our Bayesian Categorization Model applying the assumptions of each respectively.

This results in a different posterior $P(A|n_m, d_m)$ for each and thus a different $p_t = P(A|n_m, d_m)_t$ dependent on the model. We again use Equations 3.3 and 3.4 to calculate the likelihood of each model, where p_t is dependent on the model being used. Our Bayesian model comparison analysis (using an equal initial prior) thus follows the following equation, where d refers to the collected empirical data:

$$P(Expert|d) = \frac{P(d|Expert)}{P(d|Expert) + P(d|Sides) + P(d|Dots) + P(d|Guessing)} \tag{3.6}$$

We applied this model comparison analysis on participants across Experiments 1 and 2 for the last block of the experiment. For each participant, the final posteriors at the last trial of this analysis were used to determine which model was the best fitting model. The histograms of how many participants were fit into each model are shown in Figure 3.9. We found that overall, more participants were fit to an *Expert* model than the other models in Experiment 1, with the *Dots only* model being the next most fitted model (Figure 3.9 a). We also found variability comparing against the three Groups, with Group 3 having the most participants who fit the *Expert* model. In Experiment 2, most participants were fit to the *Dots only* model, with the *Expert* model being the next most fitted model (Figure 3.9 b). Overall, this again suggests that while many participants are able to use both features when performing the categorization task, some participants are potentially over-relying on the Dots feature and possibly ignoring the Sides feature.

While compelling, this model comparison analysis still relies on *Bayesian Sampling* as a core assumption in Equation 3.4. We will return to this model comparison analysis later in Chapter 4 and greatly improve upon it.

3.5 Discussion

In the present study, we tested human participants in a category learning task using novel objects and categories that participants learned using only the sense of touch. We further compared participant performance to that of our Bayesian Categorization Model. In Experiment 1, we tested three different training regimens and found that participants reached maximum performance when they learned by being repeatedly tested with corrective feedback. Participants in Group 3 had higher overall performance throughout the experiment and reached a greater maximum performance than participants in either Groups 1 or 2. Participants in Group 3 also showed a response behaviour that was most closely accounted for by our computational model, suggesting that participants learned the categories the best using this method.

In Experiment 2, we ran new participants on repeated testing with corrective feedback training regimen over five days and found that performance did not significantly improve despite the additional time to learn the categories. This suggested to us that participants were reaching a performance ceiling in our task that was lower than the performance of our Bayesian Categorization Model.

Overall, while the model we explored in this study seemed to account for a large proportion of participant performance, there is still a discrepancy between the observer model's performance and human performance.

Testing with corrective feedback leads to better category learning

Participants in Group 3 outperformed the participants in both Groups 1 and 2. This was clear when comparing raw and normalized performance across Groups and when comparing best-fit linear regressions. More surprising is that participants in Group 1 did not seem any better than those in Group 2. The only difference between these two groups was that Group 1 was given corrective feedback during testing blocks. Given that average performance in Group 3 was overall better than Groups 1 and 2, it was expected that Group 1 would also outperform 2. Participants in Group 2 are provided less information to learn from than participants in Groups 1 and 3, as they cannot learn from the testing blocks without corrective feedback.

There are a few possible explanations for these results. The first is that the benefit of testing blocks with corrective feedback does not make a large enough impact with only a few blocks. Group 1 only had a third of the number of corrective feedback testing blocks as Group 3, which might not be sufficient to affect performance. If this were the case, the benefit provided by corrective feedback may be exponential. This would lead to a small benefit with a small number of blocks that increases dramatically as more blocks are introduced. Further experimentation manipulating the amount of corrective feedback testing blocks would be required to determine this. It is also possible that mixing types of learning, that is training blocks versus testing blocks with corrective feedback, leads to a detriment that negates the benefit of corrective feedback. It is possible that participants can't learn a category through both types of learning in the same time frame. A third possibility is that participants in Group 2 are reinforcing their knowledge of the categories by their responses as if they were given corrective feedback, even though their responses could have been wrong. On average they are getting more objects correct than incorrect, so there would be an overall benefit to this strategy, and this might be enough to close the gap between Groups 1 and 2, making it difficult to differentiate the two groups.

Human observers are poor at self-reporting confidence

By far the worst-performing measure in this study was participants' confidence ratings. We found a positive linear relationship between confidence and posterior probability of

Category A, but with a fairly flat average slope of 0.33. This is not entirely unexpected; human observers are generally poor estimators of their own cognitive state. Even an observer who claims to be guessing can still be performing well (Dienes, 2007). Here we used a very simple confidence rating scale that relied on the self-report of participants. This is a commonly used method of extracting confidence but it is not without problems (Massoni et al., 2014).

Self-reported confidence ratings do not map directly onto performance. Often participants will be overconfident and report higher confidence than their actual success rate. In the present study, we instead observed confidence ratings below average performance. Given that participants seemed to find the task difficult and did not reach optimal performance, this low confidence rating could be an accurate reflection of their internal state.

It is possible that an alternative confidence rating method could have garnered better results. One promising method is a no-loss gambling game developed by Dienes & Seth (2010). In this method, the self-reported confidence of a participant is put into a two-stage lottery that can potentially reward them. First, a random number is drawn between 0 and 100. If their confidence rating is higher than this number, the participant is rewarded based on their accuracy in the experiment. If their confidence rating is lower, they get placed into a second lottery where a second random number is drawn. If the first random number is greater than the second, the participant is rewarded. While a little complicated, this method provides an incentive to answer truthfully and match the probability of success and discourages being over- or under-confident (Massoni et al., 2014).

A pilot study using this method instead of the simple confidence rating, unfortunately, proved to be incompatible with our categorization task. Participants either ignored the lottery completely or confused the reward of the lottery with the corrective feedback of the task. Thus, after much deliberation, the simple, inaccurate method was chosen over the no-loss gambling method.

Response variability

Our object-wise analysis of participant response behaviour revealed variability in how human observers categorize each object within our stimulus set. Data variability in an experiment with human participants in itself is not surprising. However, this is not the response behaviour we expected given our Bayesian Categorization Model. Our observer model categorizes stimuli using a MAP estimate. This means that our observer model's responses are *deterministic*; it will respond the same way to the same stimulus every time it is presented determined by the calculated posterior probability. This difference between the expected responses of our model and the observed response behaviour of our participants marks a shortcoming in our Bayesian Categorization Model.

What are the possible sources of this response variability in our human participants? One explanation could be that participant's understanding of the category distributions are changing throughout the experiment. Participants are naïve to the category distributions before the start of the experimental task and are required to learn the categories throughout the experiment. This learning could be understood and modelled as learning the *Within-Category Feature Mean* and *Within-Category Feature Variability*. As participants are learning the categories, we might expect their estimates of these parameters to shift and change throughout the experiment. However, it is expected that participants will be finished learning by the end of the first day of the experiment. Our results from Experiment 2 suggested that participants had learned the categories to the best of their abilities by the end of the first day of testing and did not significantly improve through further days of testing. From this, we don't suspect that participant understanding of the underlying category distributions is significantly changing once learning is complete by the final block.

Participants could also be applying unintended complex decision rules on their response outside of a MAP estimate. One such rule could be "if I responded '*Elyk*' three trials in a row, then next trial answer '*Noek*'" or "respond '*Elyk*' every other trial". We scanned for simple patterns in response behaviour throughout our participants and did not find any indication of such rule sets being used. However, if a participant was inconsistent with their decision rules and employed one or several different rules at different times throughout the experiment we would be unable to accurately detect this.

One source of response variability is the presence of a lapse rate. In some percentage of trials, participants may not respond based on their measurements but instead randomly choose between the two categories. This could be driven by a lapse in attention to the task, either from fatigue or other factors outside of our control. We don't expect the lapse rate present in participants to be very large, perhaps only contaminating 5% of trials at most. Furthermore, we'd also expect that on half of those lapse rate trials, the participant would answer correctly by chance, meaning lapses might account for a roughly 2.5% performance cost. In comparison, our calculated percent match to the observer averaged across participants was around 83% by Block 9, meaning that on roughly 20% of trials, participants' responses were different from those of our Bayesian Categorization Model. So while this does provide some explanation for the variability seen, it fails to explain the full extent of variability.

Another possibility is that rather than using a MAP estimate to drive response decisions, observers are instead employing *Bayesian Sampling*, sometimes called *Posterior Sampling* or *Probability Matching*. This theory posits that, rather than responding based on the posterior probability directly, an observer instead samples from the posterior and responds based on this sampling. Under this scenario, the frequency of responses should match the posterior probability over many trials. It is very difficult experimentally to distinguish between *Bayesian Sampling* and other sources of response variability. An observer who is learning the underlying category distributions would have their decision cutoff point shift, potentially resulting in behaviour that appears as though it follows *Bayesian Sampling* (Kubovy et al., 1971; Healy & Kubovy, 1981). Likewise, the presence of any amount of *Measurement Noise*, which would create variability in responses by making every measurement a stochastic process, would also result in behaviour that appears to follow *Bayesian Sampling*.

It is also possible that while we expected *Enclosure* to be the haptic exploratory procedure participants would use to extract the Number of Sides feature, they instead employed a *Contour Following* strategy. *Contour Following* is a much slower exploratory procedure taking significantly more time than *Enclosure*, requiring the participant to slide their hand or finger along the entire edge of the object (Lederman & Klatzky, 1993). If participants were using *Contour Following*, then the 5-second trial time window might not have been sufficient to accurately extract the Number of Sides feature measurement. This could result in significant variability in the measurement of this feature compared to the Dot Spacing feature.

A major potential source of this response variability could be internal sensory *Measurement Noise*. Our modelling has ignored *Measurement Noise* up to this point. It was assumed that with a long exploration time, participants would be able to extract sufficient information about the features to acquire an accurate measure of the *distal stimulus*. However, this may not be the case, and the presence of significant *Measurement Noise* could potentially have a large impact on participant performance. If this variability is large enough, it could result in categorization errors by an observer using a MAP estimator. We extend our model to incorporate measurement noise in Chapter 4.

3.6 Conclusion

The two experiments presented here investigated the ability of human participants to categorize objects via haptic exploration and assessed human performance in light of a Bayesian Categorization Model. Participants were able to learn novel categories of physical objects when presented with two overlapping category distributions. The Bayesian Categorization Model developed in Chapter 2 proved to be moderately accurate in predicting human response behaviour throughout both Experiments. However, human performance reached at most 80% accuracy in comparison to the model. Further work will attempt to explore if the presence of *Measurement Noise* is a significant source of this response variability.

3.7 Figures and Captions

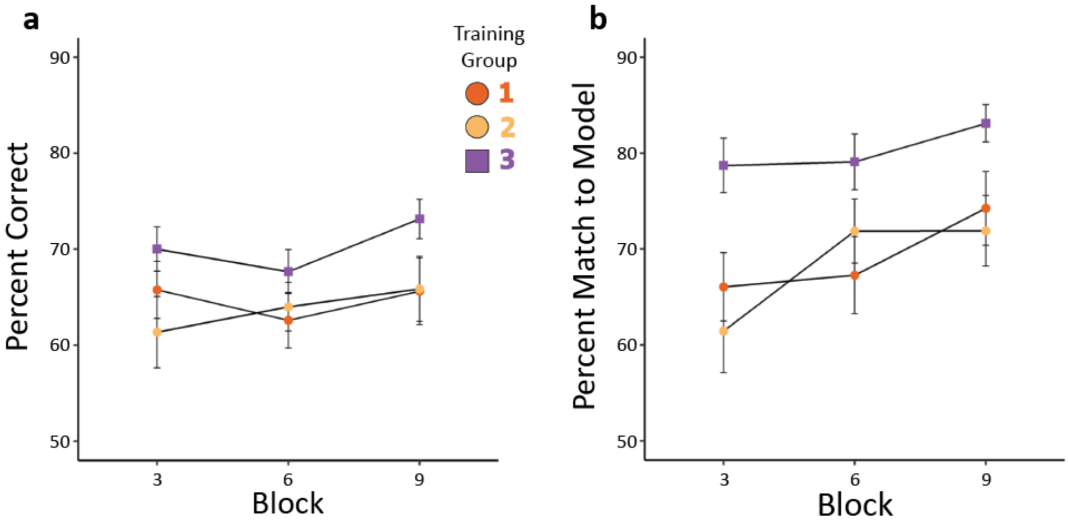


FIGURE 3.1: Performance across the three Groups over the course of Experiment 1. Group 1 (orange) and Group 2 (yellow) completed a mixture of training Blocks and testing Blocks, and Group 3 (purple) only completed testing Blocks for the entirety of the experiment. The comparison points where all three groups completed a testing Block were Blocks 3, 6, and 9. Trials where we expected participants to guess (where $P(A|n, d) = 0.5$) were removed from the analysis. A) Percent correct across the three Groups. By Block 9, participants in both Groups 1 and 2 reached a final percent correct of 65%, while participants in Group 3 reached 73%. B) Percent match to predicted responses according to the Posterior Probability of our Bayesian Categorization Model across the three Groups over the course of the experiment. By Block 9, participants in Group 1 reached a final matched percent of 74%, Group 2 reached 72%, and Group 3 reached 83%.

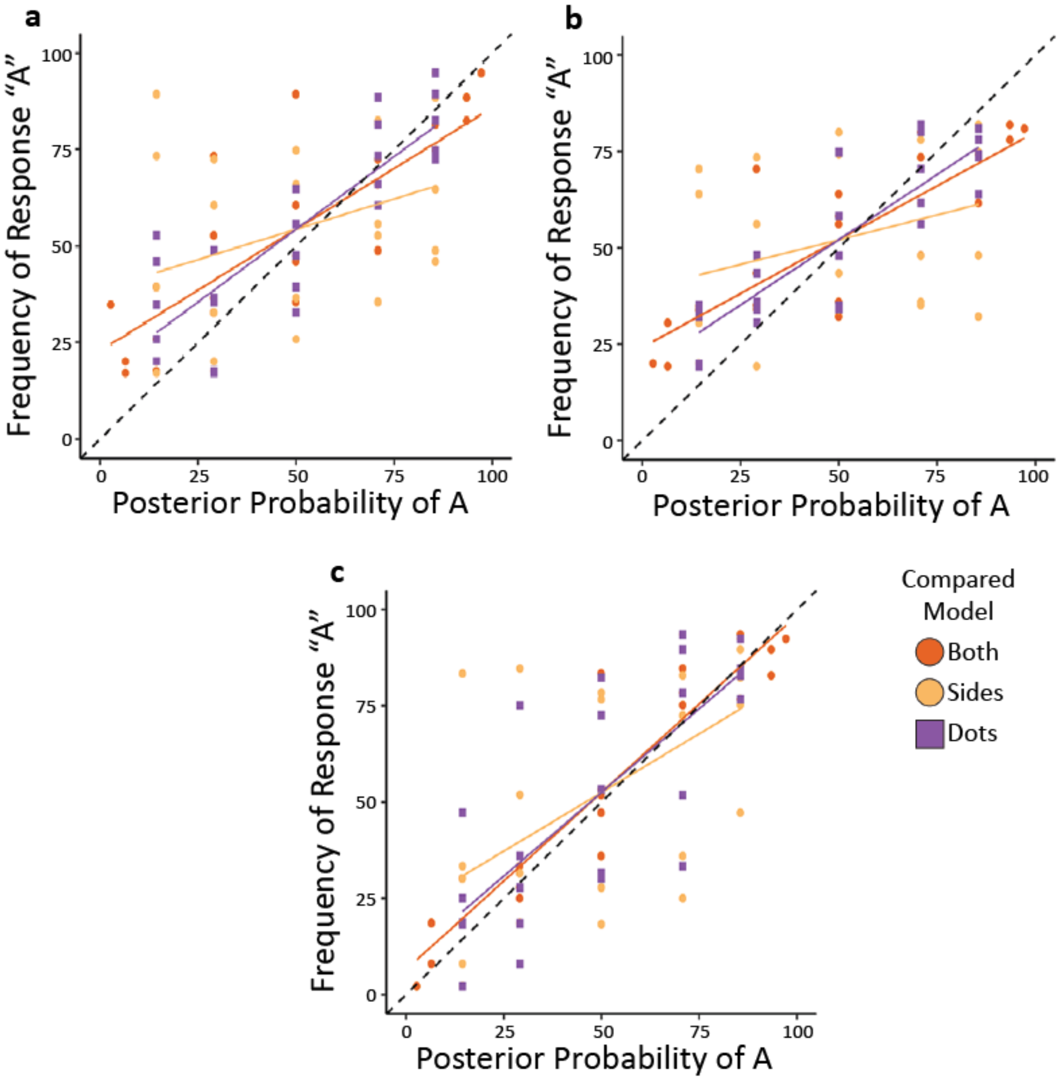


FIGURE 3.2: Frequency of Category A responses compared against the Posterior Probability of each object belonging to Category A according to three variations of our Bayesian Categorization Model: an expert model which used both Sides and Dots (orange), a Sides-only model (yellow), and a Dots-only model (purple). A slope closer to 1 indicates a better fit for the model. A) displays results in Group 1, with the best fitting model being the Dots-only model. B) displays results for Group 2, with the best fitting model being the Dots-only model. C) displays Group 3, with the best fitting model being either the expert model or the Dots-only model.

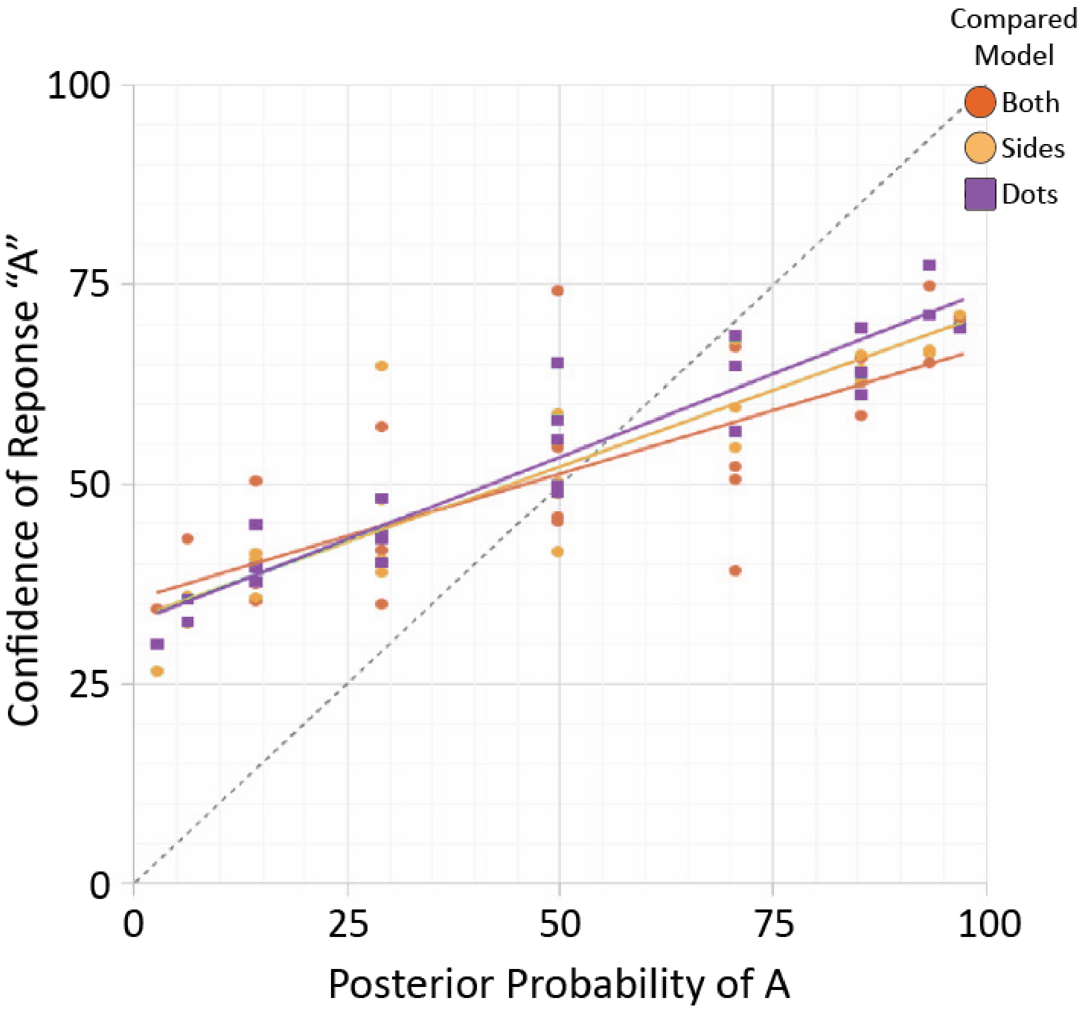


FIGURE 3.3: Confidence that each object belonged to Category A compared against the Posterior Probability of each object belonging to the Category A category across our three Groups. A slope closer to 1 indicates a better fit for the model predicting confidence. The slope of all three Groups was found to be about 0.33.

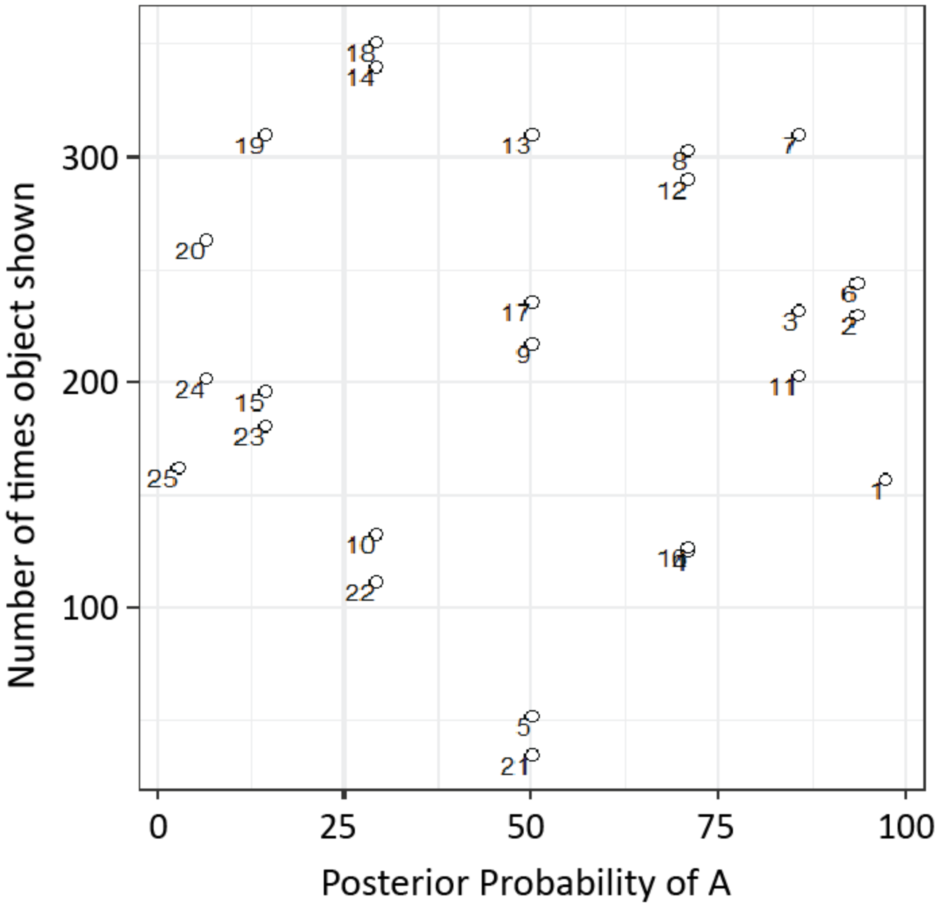


FIGURE 3.4: Histogram of the number of trials each object was presented across all participants and testing Blocks combined during Experiment 1. Each numbered point corresponds to one of the 25 objects. Objects were arranged according to their Posterior Probability of Category A $P(A|n,d)$ as calculated using Equation 2.8.

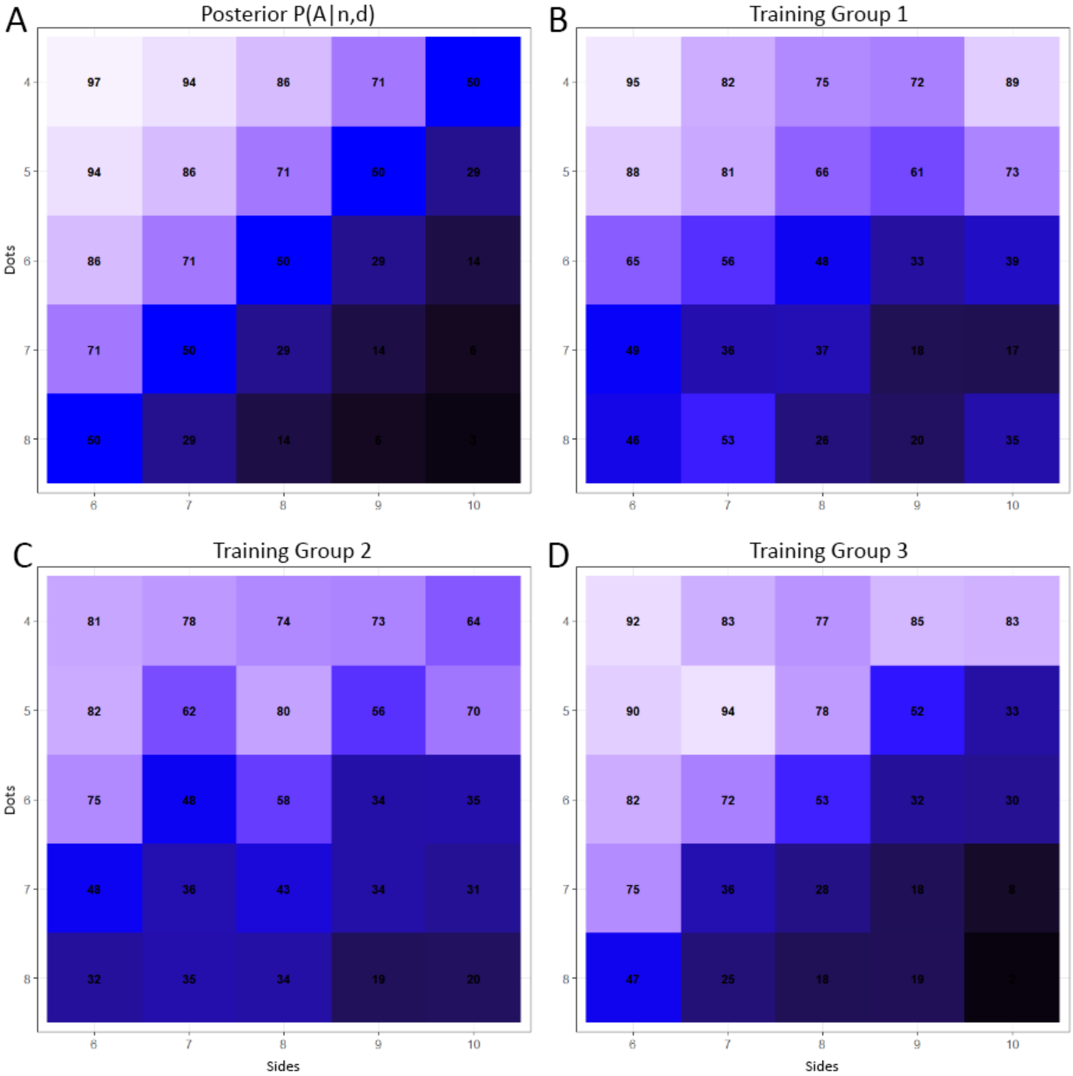


FIGURE 3.5: Response frequencies of responding "A" for each of the 25 objects across participants. A) displays the Posterior Probability P(A|n,d) calculated using Equation 2.8 for visual reference. B) displays the frequency of "A" responses for participants in Group 1. C) displays the frequency of "A" responses for participants in Group 2. D) displays the frequency of "A" responses for participants in Group 3.

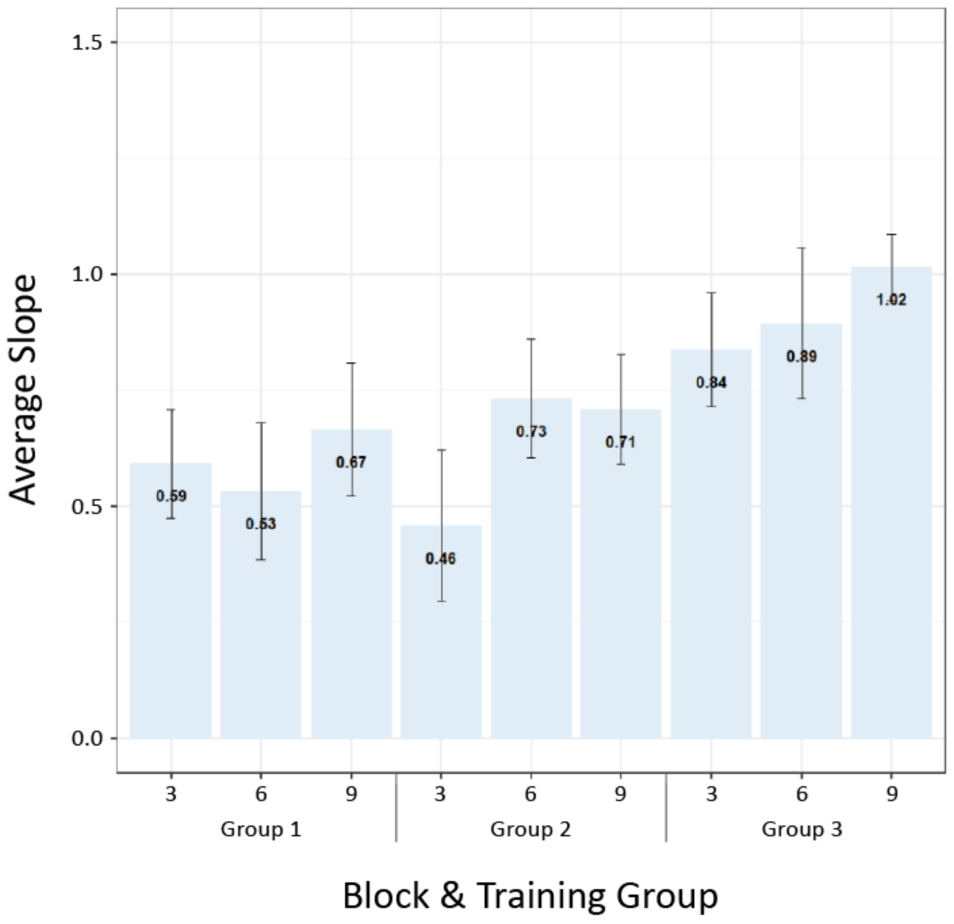


FIGURE 3.6: Average linear regression slopes calculated across Groups and Blocks using a more advanced parameter estimation analysis. The x-axis is split for both Blocks and Training Groups. This approach considered each trial as a coin flip where the probability of the participant responding Category A followed directly from the Posterior Probability $P(A|n,d)$. Slopes were calculated for each individual participant and averaged within each Group and Block. Error bars shown are standard error.

Table 3.1: Results for Experiment 2 across the five days of testing. The average percent correct, percent match to the model, and confidence were calculated across participants on the last block of each day.

Day	Percent Correct	SD	Percent Match	SD	Confidence	SD
1	71.8%	7.4	81.7%	7.8	76.6%	8.0
2	74.7%	8.2	85.3%	7.5	79.5%	5.7
3	72.5%	7.6	79.4%	11.2	78.2%	10.1
4	74.0%	9.6	81.2%	11.3	79.2%	9.7
5	70.2%	9.2	77.0%	9.4	78.7%	10.1

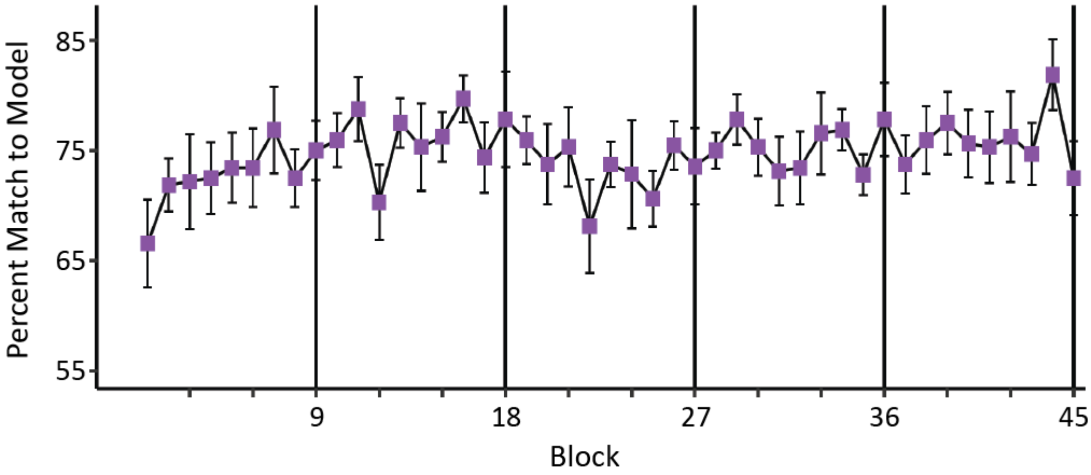


Figure 3.7: Performance for participants over the course of five days in Experiment 2. Trials where we expected participants to need to guess (where $P(A|n, d) = 0.5$) were removed from the analysis. Participant performance was calculated in terms of their proportion match to the Bayesian observer model

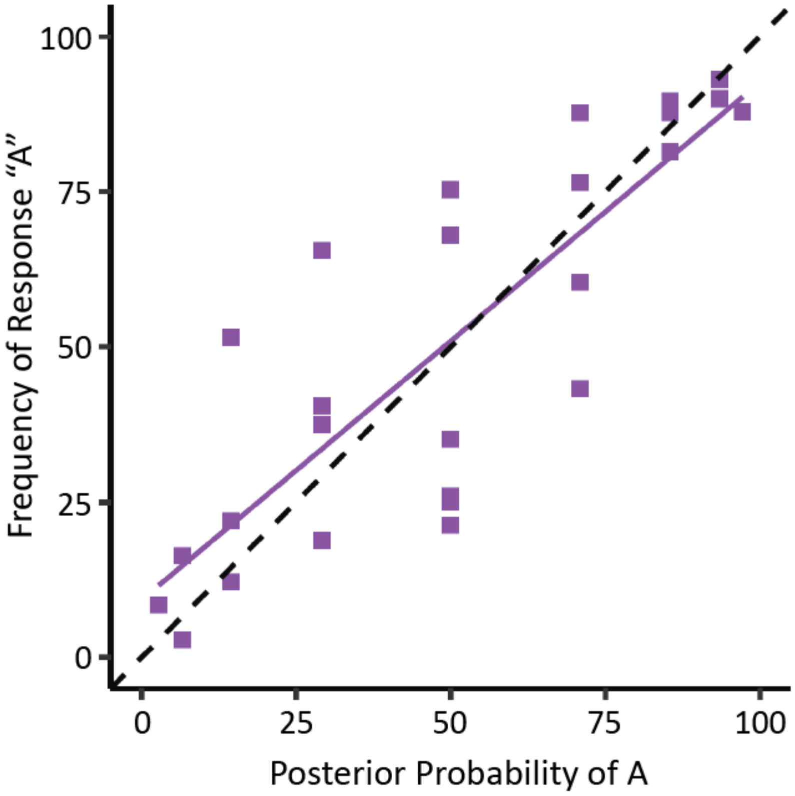


FIGURE 3.8: Frequency of Category A responses compared to the Posterior Probability of each object belonging to Category A using the "expert" Bayesian Categorization model which considered both Sides and Dots.

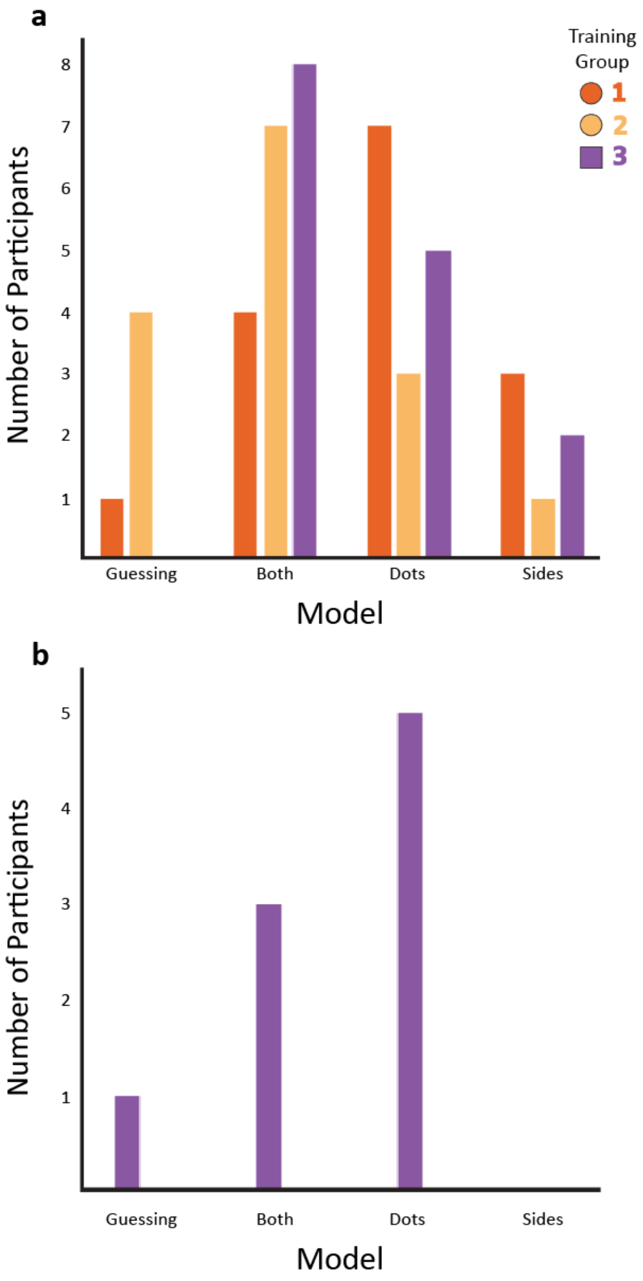


FIGURE 3.9: Histogram of the number of participants fitted to either a *Guessing*, *Both Sides and Dots*, *Dots-only*, or *Sides-only* model following our model comparison analysis. a) Experiment 1 participants. b) Experiment 2 participants.

Chapter 4

Measurement Noise and Computational Modelling

4.1 Introduction

In the previous chapter, we described two experiments that explored the application of the Bayesian categorization model introduced in Chapter 2 in a human categorization experiment. In Experiment 1, we found that participants were able to learn novel categories over time and that our model predicted a good proportion of response behaviour. However, there were still discrepancies between our model's performance and the average participant's performance. At best, our Bayesian Categorization Model only accounted for roughly 80% of the categorization trials in participants trained through corrective feedback. Experiment 2 investigated whether this discrepancy was due to a lack of training on the categories. The discrepancy remained even after five days of continuous testing, during which participants had reached asymptotic performance.

Another explanation for this discrepancy is that the categorization model we used is too simplistic. As mentioned in Chapter 3, the model did not account for the sensory noise that might be present in human participants. The presence of such noise would reduce the reliability of measured features and lead to trial-by-trial inaccuracies. This would presumably lower the maximum performance achievable by an observer. In this chapter, we explore a robust computational model that accounts for this sensory noise.

Throughout the previous chapter, we assumed that the impact of sensory noise in an observer categorizing our stimuli was negligible. This assumption was justified by the exploration period of 5 seconds given to our participants, a relatively long duration in comparison to many previous visual categorization, cue combination, and tactile studies. For example, Bankieris et al. (2017) presented audio and visual stimuli over 500ms to participants making categorical judgements. Ashby & Gott (1988) presented visual stimuli comprised of two lines of varying lengths at a 90° angle to participants for up to 5 seconds, allowing participants to terminate the trial and respond anytime within that time frame. Brooks & Hannah (2006) visually displayed pairs of visual stimuli at a time to participants for up to 10 seconds each. Ariely (2001) displayed sets of circles for 500ms intervals to participants when exploring how human participants extract statistical information from stimuli displays. Among tactile-focused experiments, the duration of stimuli presentation is typically on the order of 1 second (Peters et al., 2015; Li et al., 2017; Lederman & Klatzky, 1993; Hillis et al., 2002). It was believed that with the advantage of a relatively large exploration time window, our participants would be able to extract highly accurate feature measurements. This assumption, however, was likely naive.

The consideration of *Measurement Noise* in perceptual tasks has long been a central component of cue-combination paradigms (Alais & Burr, 2004; Ernst & Banks, 2002; Ernst & Bülthoff, 2004; Rosas et al., 2005; Knill & Saunders, 2003; Jacobs, 2002). Sensory *Measurement Noise* drives cue or feature reliability and can lead to shifts in perception as the measurements involved are combined into a percept. However, few studies to our knowledge have explored the impact of *Measurement Noise* in categorical tasks. Bejjanki et al. (2011) explored this question by looking at audio-visual speech perception of phonemes and found that feature uncertainty driven by sensory noise was taken into account by human participants. Bankieris et al. (2017) trained and tested participants on novel audiovisual categories comprised of visual dot arrays and audio labels and found that participant response behaviour matched that of a model which accounted for *Measurement Noise*. Both of these studies expanded on the general cue-combination formula with the inclusion of category variability or environmental noise alongside sensory noise.

How might an observer's performance and response behaviour in our categorization experiment change if their sensory data were degraded by the presence of sensory *Measurement Noise*? The presence of sensory noise would result in a degree of variance in the feature measurements. Specifically, the observer's feature measurements (e.g., dot spacing and number of sides) would no longer match the object's feature values exactly but would rather vary stochastically around the true feature values from trial to trial. If this variance were large enough, it could potentially lead to a higher likelihood of perceiving objects of one category as the other. Overall, we would expect performance in our categorization task to worsen in proportion to the amount of *Measurement Noise* present in the observer. This would also impact the learning rates of the category distributions. Since the observer is forming their understanding of the categories based on the measurement information, training on noisier data would reduce the accuracy of the observer's internal category distributions, and increase the time required to learn the categories sufficiently.

As described in detail in Chapter 2, our Bayesian Categorization Model has two main components: the Encoder model and the Decoder model. The Encoder model is the hypothesized process by which measurement data are generated. It describes the way by which the observer obtains proximal sensory measurements from the distal stimulus object. The Decoder model by comparison is the hypothesized process by which the category a particular object belongs to is inferred based on the proximal sensory measurements.

The Bayesian Categorization Model used up to this point has been *noiseless*. This means that there was no inclusion of *Measurement Noise* in either the Encoder or Decoder components of the model. Here we will begin to explore the effects of a *noisy* Bayesian Categorization Model model. This will include the presence of *Measurement Noise* in the Encoder model which generates the measurements used by the observer to make the categorical judgement. In a truly optimal observer model, the Encoder and Decoder models agree with each other. In other words, the observer is aware of the generative process by which the measurements were generated and uses that knowledge in its inference. However, this is not always the case, and sub-optimal observers may be using incomplete Decoder models which can lead to erroneous inferences. Two sub-optimal Decoder models were described in Chapter 2 as Decoder models 1 and 2, which were unaware of the presence of sensory *Measurement Noise* or the *Within-Category Feature Variability* respectively.

To explore the full theoretical impact of the presence of *Measurement Noise* in our categorization experiment, we tested the performance of simulated observers (which will be referred to henceforth as simulants) through a replication of our categorization experiment described in Chapters 2 and 3. The term "observer" will be reserved as a generalization that can apply to either simulants or participants. Simulants can be quickly generated using any configuration of Decoder models and amounts of *Measurement Noise* and run through the simulated experiment. This allows us to explore whether the use of sub-optimal Decoder models could significantly impact the performance of an observer in a categorization task. By further manipulating the *Measurement Noise* for each simulant we can determine the theoretical impact of having "poor" measurement acuity for either the Sides or Dots measurement on categorization performance, providing a possible explanation for the discrepancies between participant performance and our *noiseless* Bayesian Categorization Model. This will justify a further Experiment 3 aimed at empirically measuring the *Measurement Noise* present in human participants in our experiments using our object stimuli.

The use of simulants will also let us explore how accurately we can infer which Decoder model participants might be using in our categorization experiment. The generated simulants provide expected trial-by-trial response behaviour dependent on their set *Measurement Noise* and Decoder model. By comparing different simulants using different Decoder models, we can determine if the resulting response behaviours are different enough to be accurately differentiated into either of the Decoder models post-hoc. This will justify re-analyzing Experiments 1 and 2 as we attempt to infer which Decoder model our participants may have been using.

In the following sections, we will describe in detail the design of our simulation, how simulants were generated and how their results were analyzed. We will explore the outcomes of our simulations and what they inform us about the hypothetical presence of *Measurement Noise* in our categorization task used in Experiments 1 and 2. From there, we will describe a Two-Interval Forced-Choice (2IFC) experiment run to determine the *Measurement Noise* of the Sides and Dots features of our object stimuli present in participants. Finally, we will re-explore the results of Experiments 1 and 2 using a more complete Bayesian Categorization Model which takes into account *Measurement Noise*.

4.2 Categorization Experiment Simulation

The categorization task simulated was similar to what human participants ran in Experiments 1 and 2. The categories and stimulus set used were the same ones previously described; 25 objects ranging from 6 to 10 sides and 4mm to 8mm in dot spacing. Category A objects were drawn from a 2D Gaussian distribution with a mean of 7 sides, 5mm dot spacing, and 1.5 standard deviation. Category B objects were drawn from a 2D Gaussian distribution with a mean of 9 sides, 7mm dot spacing, and 1.5 standard deviation.

The stimulus set being constrained to 25 objects was a limitation present in the previous experiments due to the physical nature of the experimental design. This limitation was not technically a problem for our simulations and we could have presented our simulants with a denser feature space. However, in order to maintain parity between our simulation results and the results of our human experiments, we decided to impose the same restrictions.

Simulants were aware of the category means and sigmas and used the actual values described above when calculating *Within-Category Feature Variability*. They were also aware of their true *Measurement Noise* and used those values when calculating their likelihoods. Whether or not a simulant considered these values when calculating the posterior probability for either Category given a stimulus was dependent on which Decoder model they used. As described in Chapter 2, Decoder model 1 simulants would only consider the *Within-Category Feature Variability* when calculating the posterior and ignore (or were unaware of) the presence of *Measurement Noise*. These simulants used Equation 2.20 to calculate the likelihood $P(f_m|C)$. Decoder model 2 simulants would only consider *Measurement Noise* and ignore (or were unaware of) the presence of *Within-Category Feature Variability*, using Equation 2.21 in their likelihood calculations. Only Decoder model 3 simulants considered both sources of noise in their calculations, using Equation 2.22.

Although the simulants varied in which Decoder model they used when calculating the posterior probability for either category, all simulants used the same Encoder model (Equation 2.18) when presented with each object and encoding the measurements. This model combines both *Within-Category Feature Variability* and *Measurement Noise* and is congruent only with Decoder model 3. This means that measurement noise and category noise were always present factors, and the difference was whether or not the simulant considered these sources of noise.

Simulant Design

In each simulation trial, data was first generated by the Encoder model and then decoded using one of the Decoder models to respond to which category the simulant thought they were presented with. The first step, the Encoder model (described in Equation 2.18), described both how the object was drawn from the categories with their *Within-Category Feature Variability*, and then perceived by the simulant with their *Measurement Noise*. The categorization task simulation first randomly chose between Category A or B with equal probability (corresponding to the Prior $P(A) = P(B) = 0.5$ throughout our experiments). Then, it drew an object from the category, restricted to the set of 25 objects. The features of that chosen object were "presented" to the simulant (denoted by vector **f**, a set of features f). A noisy measurement (f_m) was then drawn from a Gaussian distribution centred on **f** with a sigma equal to the simulant's *Measurement Noise* for that feature.

Once the Encoder model process was complete, the simulant used either Decoder models 1, 2, or 3 to calculate its posterior probability that the object belonged to Category A using the encoded features. Finally, the simulant made a decision based on this probability of whether to respond $''A''$ or $''B''$ using a *maximum a posteriori* (MAP) estimator. If $P(A|\mathbf{f}) > 0.5$, the simulant responded with A. If $P(A|\mathbf{f}) < 0.5$, the simulant responded with B. In the case where $P(A|\mathbf{f}) = P(B|\mathbf{f}) = 0.5$, the simulant would randomly respond $''A''$ or $''B''$ with equal probability.

We can further explore the expected response behaviours for each of the 25 objects (each object denoted by vector $\mathbf{f}_g$, with g referring to one of the 25 objects) by looking at the proportion of times for each object the MAP estimate favours Category A over Category B. The features present in each object $\mathbf{f}_g$ give rise to a wide range of noisy feature measurements $\mathbf{f}_m$. For instance, a simulant with some *Measurement Noise*, when presented with $d = 6$, might encode $d_m = 6.2$ or $d_m = 5.8$. We labelled the discretized space measurements as $\mathbf{f}_{mk}$, where k denotes each measurement bin. We chose a bin size of $k = 0.25$ and a range of ±5 around the upper and lower limits of each feature range (note that this would technically give rise to physically impossible objects, however, we are treating these feature dimensions as unit-less for the purpose of our simulations). We then calculate $P(\mathbf{f}_{mk}|\mathbf{f}_g)$, the probability of a particular set of measurements given the presented object g. We can then calculate the simulant's categorization frequency for each of the 25 objects.

$$P(''A''|\mathbf{f}_g) = \sum_k \begin{cases} P(\mathbf{f}_{mk}|\mathbf{f}_g) & P(A|\mathbf{f}_{mk}) > 0.5 \\ 0 & P(A|\mathbf{f}_{mk}) < 0.5 \\ \frac{1}{2}P(\mathbf{f}_{mk}|\mathbf{f}_g) & P(A|\mathbf{f}_{mk}) = 0.5 \end{cases} \tag{4.1}$$

Equation 4.1 gives us the proportion of times the MAP estimate categorizes an object as Category A, $P(''A''|\mathbf{f}_g)$, where $''A''$ refers to the actual response given by the simulant, and vector $\mathbf{f}_g$ refers to the given object. The calculation for the posterior probability of Category A for a particular set of feature measurements $P(A|\mathbf{f}_{mk})$ was laid out in Chapter 2 (Equation 2.9). By taking the sum across each discretized perceived measurement k *only* when the simulant's posterior was $P(A|\mathbf{f}_{mk}) > 0.5$, we can find the total probability that an object $\mathbf{f}_g$ gives rise to measurements that yield a response $''A''$. When the simulant's $P(A|\mathbf{f}_{mk}) = 0.5$, they should be purely guessing as there is an equal probability of the object belonging to Category A or B. Thus, half the times that they are presented with these objects they should respond $''A''$.

By following the calculation in Equation 4.1 for each of the 25 objects g, we can generate a table listing the probabilities that when presented with each object, a simulant will respond with $''A''$. This table is dependent on the Decoder model used throughout these calculations. It is also dependent on the *Measurement Noise* present in the simulant, regardless of which Decoder model they used. Each simulant, with a particular Decoder model and set of *Measurement Noise*, will have a response table unique to them.

4.2.1 Simulant Response Table Results

Figure 4.1 displays these calculated $P(''A''|\mathbf{f}_g)$ response tables at various combinations of feature measurement noise levels. This was done for Decoder models 1, 2, and 3. The same *Within-Category Feature Variability* and *Measurement Noise* were used for both the Encoder model and Decoder models. For example, in the response tables where the Sides measurement noise of $\sigma_{nm} = 2$ was used, this value was used for both the generative Encoding of the data and the Decoder model's understanding of their own *Measurement Noise*. We did not run a scenario where the Encoder and Decoder model's *Measurement Noise* values differed.

When there is no *Measurement Noise* present ($\sigma_{nm} = 0$ and $\sigma_{dm} = 0$), Decoder models 1 and 3 generate the same response patterns. This is expected as with no measurement noise present, Decoder model 3 reduces to Decoder model 1. Decoder model 2 exhibits a very different response pattern with every object other than the two located at the respective category means, leading to a guessing response. This highlights a major flaw in the assumptions of Decoder model 2. If an observer is ignorant of the presence of category noise, then their understanding of the categories is that there is only one object representing the category positioned at the mean. If there is no *Measurement Noise* present in their observations, and they are aware of this lack of measurement noise, then it is impossible for the simulant to consider objects outside of the two respective category means. Therefore, when presented with errant objects, they have no information to utilize and must guess evenly between the two categories.

When *Measurement Noise* was equal between the two features ($\sigma_{nm} = 1$ and $\sigma_{dm} = 1$), Decoder models 1, 2, and 3 all produced the same response table. This result at first seemed puzzling as it suggests that an observer would not respond differently to our categorization task regardless of which Decoder model they used. Such a result is counter-intuitive to the concept of an "optimal" observer model. Decoder model 3 is the model that matches the Encoder model used when generating the data. Decoder models 1 and 2 both lack some information, whether it be the *Within-Category Feature Variability* or *Measurement Noise* respectively, and thus we might expect them to perform worse than Decoder model 3. This does not appear to be the case at first glance. However, there is a key simplification being made in the calculation of $P(''A''|\mathbf{f}_g)$ in Equation 4.1 which explains why the three Decoder models reduce to the same response tables. We employed a deterministic decision strategy for our simulants where they only needed to consider the equality $P(A|\mathbf{f}_{mk}) > 0.5$ to determine whether they would respond "A" or "B". This rule means it no longer matters what the precise value of the posterior $P(A|\mathbf{f}_{mk})$ is, only whether it is above 0.5. We can consider the line along the 2D feature space where $P(A|\mathbf{f}_{mk}) = 0.5$ to be the Decision criterion, where any measurement $\mathbf{f}_{mk}$ past this criterion should result in response $P(''A''|\mathbf{f}_g)$.

When the two distributions are equally symmetrical, the variances themselves don't impact where this Decision criterion is located. Two category distributions with very large sigmas will have the same criterion line as two category distributions with very small sigmas. When we increase the *Measurement Noise* of the Encoder model, this increases the variation in the encoded measurements each object can produce and thus results in some objects being miscategorized, leading to the response table's values converging to 0.5 as *Measurement Noise* increases. We can see this convergence to 0.5 in Figure 4.1 as we increase the *Measurement Noise* values to ($\sigma_{nm} = 2$ and $\sigma_{dm} = 2$). However, as far as the Decoder models are concerned, the criterion line across the feature measurement space has not changed, and thus all three Decoder models with a particular set of *Measurement Noise* produce the same response table.

Where this becomes interesting however is when the *Measurement Noise* sigmas are unequal between the two features. When there is a mismatch, the resulting distributions become unsymmetrical. We can see why this occurs by realizing that the *Within-Category Feature Distributions* and *Measurement Noise* distributions give rise to a new Gaussian Distribution whose variance is the sum of the two variances. As we derived in Chapter 2 (Equation 2.28) we can consider the Encoder model as a two-step process that first draws an object from the *Within-Category Feature Distribution*, and then draws a measurement from that object using the *Measurement Noise* distribution. The resulting Category A and Category B Gaussian Distributions therefore have a new standard deviation ${\sigma^*}_{Cf}$, which can be calculated with Equation 2.27:

$$ {\sigma^*}_{Cf} = \sqrt{\sigma^2_{Cf} + \sigma^2_{f_m}} \tag{4.2} $$

We can see that if we were to increase the *Measurement Noise* sigma σ_{f_m} by the same amount for each feature, the resulting ${\sigma^*}_{Cn}$ and ${\sigma^*}_{Cd}$ would remain symmetrical and the Decision criterion line would not be changed. However, when $\sigma_{nm} \neq \sigma_{dm}$ the distributions become unsymmetrical, and the criterion will skew towards the feature dimension with the smaller *Measurement Noise*. This makes sense as an observer with much greater *Measurement Noise* in one feature than the other will naturally rely less on that noisier feature in their decision-making since it is unreliable. In other words, an observer with an extremely high Sides sigma will mostly ignore the Sides feature and rely on the Dots feature.

It is in this unsymmetrical scenario in which the Decoder model's response tables differentiate. Since the three Decoder models consider the *Within-Category Feature Variability* and *Measurement Noise* in different ways, the variances of the resulting unsymmetrical distributions will differ leading to variation in their criterion placements. We can see this in the last column of Figure 4.1 where ($\sigma_{nm} = 2$ and $\sigma_{dm} = 1$). Decoder model 1, which only considers *Within-Category Feature Variability*, is not impacted by the unsymmetrical *Measurement Noise* sigma at all as it doesn't consider that in its calculation. The resulting criterion is based on ${\sigma^*}_{Cn} = 1.5$ and ${\sigma^*}_{Cd} = 1.5$ (which comes directly from the *Within-Category Feature Variability*). As the distributions from Decoder model 1 are still symmetrical, the criterion does not change from before. Decoder model 2, which only considers *Measurement Noise*, is directly impacted by the unsymmetrical sigma as that is the only noise it is considering. Its resulting criterion is based on ${\sigma^*}_{Cn} = 2$ and ${\sigma^*}_{Cd} = 1$, and it becomes very skewed towards favouring the Dot's feature. Decoder model 3, which considers both *Within-Category Feature Variability* and *Measurement Noise*, is marginally impacted by the unsymmetrical sigma values. We can calculate the new ${\sigma^*}_{Cf}$ values that the resulting criterion is based on using Equation 4.2, which come to ${\sigma^*}_{Cn} = 2.5$ and ${\sigma^*}_{Cd} = 1.8$. The resulting criterion is skewed, but not as much as Decoder model 2.

It is important to note that Decoder model 3 matches the Encoder model used to generate the data originally. That means that the Decoder model 3's response table is the optimal way to respond to reach the highest performance. We can expect that when the Sides and Dots feature sigma is symmetrical, all three Decoder models should respond identically and achieve the same percent correct. However, when the feature sigma is unsymmetrical, Decoder model 3 represents the optimal response behaviour and should perform the best out of the three models. Note that the presence of *Measurement Noise* in the Encoder model results in worse performance overall for all three models, as it increases the chance that the simulant might be presented with a Category A object yet respond $''B''$.

4.2.2 Varying Measurement Noise Between Decoder models

With the $P(''A''|\mathbf{f}_g)$ response tables, we can greatly reduce the computational cost of the categorization task simulation. Rather than repeatedly calculating the posterior probability $P(A|\mathbf{f}_{mk})$ for every object presented, we can calculate this once and then use the generated table of $P(''A''|\mathbf{f}_g)$ and take advantage of the binomial nature of the task. Each trial can be treated as a binomial coin flip with two possible outcomes, either an $''A''$ response or a $''B''$ response from an observer. The probability of either response is given in the table. Each Decoder model gives rise to a unique table, providing a different set of probabilities depending on which model the simulant is using.

Another way to conceptualize this scenario is to imagine there is a single coin that has some probability of landing heads P_g. The probability of the coin landing heads is dependent on g, which changes each time the coin is flipped. This value g is drawn from the category distributions for A and B we outlined above. The coin landing heads correspond with an observer responding $''A''$, and tails for $''B''$.

We first explored the impact of *Measurement Noise* on the categorization performance of the three main Decoder models outlined in Chapter 2. We generated a set of simulants varying their internal *Measurement Noise* sigmas from 0 to 4 with a step size of 0.25 for both the Sides and Dots features resulting in 289 unique simulants. Each simulant was run through 10,000 trials of a categorization task. This was done for each of the three Decoder models, resulting in a total of 867 simulant participants each running 10,000 trials. Each simulant had a corresponding response table generated based on their Decoder model. This response table was used to randomly choose an $''A''$ or $''B''$ response in every trial given the object the simulant was presented. We could then calculate the overall accuracy of the model by calculating the frequency with which the simulant categorized the given object correctly.

There were no free parameters in our simulation. The *Within-Feature Category Variability* of the two categories was known and available to the simulants, and the *Measurement Noise* for the two features was generated and varied across simulants. Learning these parameters was not a part of this simulation; simulants were experts at the task and performed at the same skill level from the first trial onward.

Decoder model 1 simulants are only aware of the *Within-Category Feature Variability* and ignorant of their internal sensory *Measurement Noise*. Figure 4.2 shows the results from our simulation of Model 1. The lowest percent correct was recorded with the worst noise sigmas at ≈ 56%, while the highest percent correct was recorded with the best set of sigmas (perfect measurements with virtually no noise) at ≈ 77%. There was a roughly linear increase going from the worst sigmas to the best sigmas while decreasing both sigma values by equal amounts. When we increased one feature's sigma, but not the other, performance also suffered, with a percent correct of ≈ 60% when $\sigma_n \approx 0$ and $\sigma_d = 4$. The same occurred when these values were flipped. From these results, we can draw a fairly standard conclusion: performance in a categorization task is dependent on the internal *Measurement Noise* of the observer, even if they are unaware of this noise.

Decoder model 2 simulants are only aware of the *Measurement Noise* and ignorant of the *Within-Category Feature Variability* underlying the categories. Effectively, these simulants treat the situation as though there are only two objects, a single *Elyk* and a single *Noek*, each situated at their respective means. Figure 4.3 shows the results from our simulation of Model 2. The worst noise sigmas again yielded poor performance at ≈ 56%, but the worst performance was recorded when both $\sigma_{nm} = 0$ and $\sigma_{dm} = 0$ at ≈ 54%. The best performance was recorded with the "next best" sigmas where $\sigma_{nm} = 0.25$ and $\sigma_{dm} = 0.25$ at ≈ 77%. If only one of σ_{nm} or σ_{dm} was set to zero, performance was ≈ 60%.

We can see why this occurred by examining the Decoder model 2 response table in Figure 4.1 when both *Measurement Noise* sigmas were set to zero. Under these circumstances, the simulant guesses for all objects other than the two representing the *Within-Category Feature Means* as it cannot consider any other objects belonging to either category. The theoretical performance here can be calculated by combining the probability of drawing an *Elyk* object and responding "*Elyk* " correctly or the probability of drawing a *Noek* object and responding "*Noek* " correctly, which in this case comes to 54%. Note that the calculation for the probability of drawing each object in either the *Elyk* or *Noek* boxes requires us to either estimate through brute force simulating random draws in our categorization experiment or account for the discrete feature range by using the Truncated Gaussian probability density formula covered in Chapter 2.

Decoder model 3 simulants represent ideal observers who are aware of both the *Within-Category Feature Variability* and their internal sensory *Measurement Noise*. Figure 4.4 shows the results from our simulation of Model 3. Similar to Model 1, the lowest percent correct was recorded with the worst noise sigmas at around ≈ 56%, while the highest percent correct was recorded with the best sigmas (virtually perfect noise) at ≈ 76%. The slope between the worst sigmas and best sigmas performance was roughly linear.

Where Model 1 simulants differ greatly from Model 2 and Model 3 simulants is when $\sigma_n \neq \sigma_d$. In the extreme cases with the best *Measurement Noise* in one feature and the worst in the other, we found an improved performance from Model 3 with percent correct at either $\sigma_{n_m} = 4$ or $\sigma_{d_m} = 4$ to be around $\approx 70\%$. Model 2 simulants also demonstrated this improvement at the "next best" sigmas $\sigma_{fm} = 0.25$. This follows from what we would expect when examining the response tables (Figure 4.1). In these mismatch *Measurement Noise* conditions, simulants using Decoder model 1 are unable to account for the unsymmetrical distributions resulting from combining the *Within-Category Feature Variability* and uneven *Measurement Noise*. This results in Model 1 being heavily impacted by the worst measurement present in its inference. However, Models 2 and 3 can account for this uneven *Measurement Noise*. As it turns out, a percent correct of $\approx 70\%$ is the expected theoretical maximum percent correct if an observer is only using one feature rather than two. This indicates that as the *Measurement Noise* of one feature increases more than the other feature, Models 2 and 3 will put less weight on that feature until eventually ignoring it completely.

We can empirically determine if overall one Model performed better than another Model in our simulations by calculating a "distance" value between the resulting planes. To determine this distance we calculated the sum of the point-by-point difference between two planes. We excluded the simulation results where either $\sigma_{n_m} = 0$ or $\sigma_{d_m} = 0$ as simulants using Decoder model 2 had the different response behaviour previously discussed. The distance between Models 1 and 2 was 494% favouring Model 2, which corresponds with the higher percent correct expected in the mismatch *Measurement Noise* conditions. The distance between Models 1 and 3 was 562% favouring Model 3, again corresponding with the mismatch *Measurement Noise* conditions but also indicating that Model 3 performed even better than Model 2. We confirmed this by calculating the distance between Models 2 and 3 to be 68% favouring Model 3.

Sub Models and Hypotheses

We have so far described and focused on only three Decoder models— one which only considers *Within-Feature Category Variability*, one which only considers *Measurement Noise* and one which considers both. This third model which we have been referring to as Decoder model 3 is considered to be our Optimal Decoder model. This model uses all of the information available to it and should theoretically perform the best in our categorization task. The other two models, Decoder models 1 and 2, can be considered to be sub-optimal variants of Model 3, with each lacking some piece of information in its consideration of the given objects.

We can outline further sub-optimal Models by considering simulants who completely ignore either of the features present in the objects. This addresses the possibility of participants ignoring the Sides feature that arose out of Experiments 1 and 2. It is difficult to account for this in our analyses from participant self-reports alone as a participant's self-evaluation of their strategy might not align with their actual in-experiment behaviour. However, by defining sub-models that account for ignoring the different features, we can determine post-hoc if participant behaviour matches a Decoder model that ignores either the Sides or Dots feature.

We can create further sub-models of each Decoder model for either "Sides-Only" observers, "Dots-Only" observers, or "Sides and Dots" observers who account for both. This gives us a total of 9 sub-models, with one of them being our Optimal "Expert" model. We calculated these sub-models by removing the feature-relevant likelihood term from our calculation of the posterior probability $P(A|\mathbf{f}_{mk})$ so that they would only be considering one of the features and not the other.

Figure 4.6 shows the simulation results for the 9 sub-models over the same range of *Measurement Noise* sigmas as before. Column A shows the results of the "Sides and Dots" Decoder models, which is the same data we have seen and discussed before. Columns B and C show the *Sides-Only* and *Dots-Only* Decoder models respectively. We can observe immediately that there are little to no differences between any of the Decoder models, besides Decoder model 2 having difficulty when the *Measurement Noise* sigma was set to 0. Calculating distance values between the *Sides-Only* Decoder models confirms this with there being a distance of 7.4% between Models 1 and 2, and a distance of 7.2% between Models 1 and 3.

These results make sense when considering the insights made from analyzing the response tables in Figure 4.1. The three Decoder models were most distinguishable in the presence of uneven dots and sides *Measurement Noise*. This led to different criterion placements driven by our MAP estimator. When only one feature is being considered, the change in *Measurement Noise* of the attended feature no longer matters as it changes equally between the two categories.

4.2.3 Post-Hoc Inference of Observer Decoder models

We have up to this point explored in detail the response behaviour of the three Decoder models under varying circumstances of *Measurement Noise*, as well as their expected performance in the categorization task. One goal of having well-defined opposing Decoder models is to be able to infer post-hoc which Decoder model an observer may have used to generate the responses they gave. This will be the focus of a later section in this chapter where we will re-analyze Experiments 1 and 2 with our updated Decoder models using empirically driven feature *Measurement Noise* values. However, before we can perform this post-hoc inference on participant response data, we must first determine if it is viable to make this inference with our simulants.

To determine if this analysis is possible, we generated new Model 1, 2 and 3 simulants using *Measurement Noise* sigmas $\sigma_{n_m} = 1.645$ and $\sigma_{d_m} = 1.183$ (the justification for these values will be provided in the following section detailing Experiment 3). Critically, these *Measurement Noise* values are uneven and lead to different response tables for each

Decoder model. Each simulant was run through 360 trials, the number of trials of a single experimental day in Experiments 1 and 2. A model comparison analysis between Decoder models 1, 2, and 3 was performed over the 360 trials, where the Model with the highest posterior probability was recorded. This was done 50 times per simulant response table for a total of 18,000 trials.

To compare responses between each model we can take advantage of the Binomial "coin flip" nature of the two-category categorization task which we previously discussed. Each trial of the experiment can be considered as a signal Binomial trial with two outcomes, a response of $''A''$ (corresponding with "heads") or a response of $''B''$ (corresponding with "tails"). In general, the probability of a simulant response (whether $''A''$ or $''B''$) is given by the following Binomial relationship:

$$P(Responses|\mathbf{f}_g)_{Dn} = \prod_t P(''A''|\mathbf{f}_g)_{Dn_t}{}^k (1 - P(''A''|\mathbf{f}_g)_{Dn_t})^{(1-k)} \tag{4.3}$$

where $k = 1$ when the simulant responded $''A''$, and $k = 0$ when the simulant responded $''B''$. The probability $P(''A''|\mathbf{f}_g)_D$ comes from our response tables discussed previously and is dependent on which Decoder model was used. We will denote the three Decoder models as $D1$, $D2$, and $D3$. A fourth control "guessing" model (where the simulant simply answers randomly between the two categories) was included and denoted as Dr. Finally, the *Sides-Only* and *Dots-Only* sub-models were also included (denoted Ds and Dd). Since we determined from analyzing the simulation results in Figure 4.6 that the three Decoder models give the same output when only one feature is considered, we arbitrarily used Decoder model 3's *Sides-Only* and *Dots-Only* sub-models for Ds and Dd.

We can calculate the probability of observing *all* the responses by multiplying across all trials t. We will use subscript n as Dn to denote the Model used by simulants in generating responses (also referred to as the generative Model), and superscript $*$ as Dn^* to denote the Model our analysis matched to those responses (also referred to as the matched Model). The model comparison analysis followed a Bayesian comparison between the six models according to the following equation:

$$P(\mathbf{f}_g|Responses)_{D1} = \frac{P(Responses|\mathbf{f}_g)_{D1}}{\sum_{Dn} P(Responses|\mathbf{f}_g)_{Dn}} \tag{4.4}$$

To assess the accuracy of our model comparison analysis and its ability to recover the correct generative Decoder model, we ran multiple simulations with each generative Model. Each Model was run for 50 sets of 360 trials. This number of trials was chosen as it represents 9 blocks of our categorization experiment in Experiments 1 and 2. At the end of each 360 trial set, the Model with the highest posterior probability was chosen as the matched Model. The frequency with which our analysis matched each generative

Model to each of the six Models was calculated and organized into confusion matrices (Figure 4.7). We did this for four configurations of *Measurement Noise* for the *Sides* and *Dots* features.

In general, our model comparison analysis was able to recover the correct generative Model used. If we consider $D1$, $D2$, and $D3$ together against the other alternative generative models (Dr, Ds, and Dd), we can see that our analysis was able to almost perfectly determine if the simulant data was generated using either the three Decoder models, the Guessing model, the *Sides-Only* model, or the *Dots-Only* model. This accuracy was irrespective of the configuration of *Measurement Noise* used to generate the data. From this, we can maintain a high degree of confidence that if our model comparison matches behavioural data to either the Guessing, *Sides-Only*, or *Dots-Only* models, this is likely an accurate matching to the true generative model. This was true both for when $\sigma_{n_m} = \sigma_{d_m} = 1$ and $\sigma_{n_m} = \sigma_{d_m} = 0.5$. Unfortunately, when the feature measurement sigmas are symmetrical, we are unable to distinguish between $D1$, $D2$, and $D3$. This is because, as discussed previously, the response tables between these three models are all identical when $\sigma_{n_m} = \sigma_{d_m}$. As such, the expected behaviours are also identical, and our model comparison analysis comes to an equal posterior probability that either model could have generated the results.

When $\sigma_{n_m} = 2$ and $\sigma_{d_m} = 1$ (Figure 4.7 c), our analysis became very accurate in correctly matching $D1$ 94% of the time, $D2$ 86% of the time, and $D3$ 92% of the time. This indicates that having an asymmetrical combination of *Sides* and *Dots Measurement Noise* is necessary for our model comparison analysis to be able to match behavioural data to its generative model accurately. When this asymmetry was reduced to $\sigma_{n_m} = 1.645$ and $\sigma_{d_m} = 1.183$ (Figure 4.7 d), which will be the sigma values we will use later in this chapter, this model matching accuracy lowered to $D1$ 84% of the time, $D2$ 82% of the time, and $D3$ 64% of the time. This reinforces that having a larger separation in *Measurement Noise* between our two features leads to higher accuracy of our model comparison analysis. This also gives us an idea of how trustworthy we should find our analysis.

When we reduced the number of trials in our analysis set to 50 sets of 120 representing 3 blocks of the experiment (Figure 4.7 e and f), we generally found a decrease in the accuracy of our analysis. When $\sigma_{n_m} = 1.645$ and $\sigma_{d_m} = 1.183$, we found model-matching accuracy of $D1$ 68% of the time, $D2$ 74% of the time, and $D3$ 42% of the time. The analysis had the most difficulty with $D3$, confusing it with $D1^*$ 28% of the time and with $D2^*$ 42% of the time. This tells us that model comparison analysis over 120 trials, while informative, should be treated as less trustworthy as our analysis has a higher likelihood of confusing the three generative Decoder models together.

To test the sensitivity of our analysis against incorrect assumptions about *Measurement Noise*, we simulated two scenarios where the generative model's noise sigmas were different from that assumed by our analysis. Figure 4.7 (g) shows the resulting confusion matrix when the generative model's noise sigmas were $\sigma_{n_m} = 2$ and $\sigma_{d_m} = 1$, but the analysis assumed $\sigma_{n_m} = 1.645$ and $\sigma_{d_m} = 1.183$. This difference can be quantified by the ratio between the two sigmas, which was 2 for the generative model, and 1.39 for the matched models. The analysis in this case had a lot more difficulty. $D1$ was matched correctly 84% of the time but was incorrectly matched to $D3^*$ 16% of the time. $D2$ was matched correctly 48% of the time but was incorrectly matched to $D3^*$ and $Dd^* \approx 17\%$ of the time. $D3$ was matched correctly only 20% of the time while being confused with $D2^*$ 78% of the time. Dr, Ds and Dd were still generally matched correctly.

Figure 4.7 (h) shows the opposite case when the generative model's noise sigmas were $\sigma_{n_m} = 1$ and $\sigma_{d_m} = 2$, but the analysis assumed $\sigma_{n_m} = 1.645$ and $\sigma_{d_m} = 1.183$. This is a ratio of 0.5 for the generative model and 1.39 for the matched models. Our analysis in this scenario was again very accurate when matching Dr, Ds and Dd, and matched $D1$ correctly 88% of the time. However, it confused $D2$ as D^* 90% of the time. It also confused $D3$ with $D1^*$ 92%, and rarely any model to $D2^*$ or $D3*$ at all.

This model comparison analysis will be the basis for how we approach part of our re-analysis of Experiments 1 and 2 where we will attempt to determine which generative Model human participants might have been using throughout our experiments. As an overall takeaway of our evaluation of our model comparison analysis, we can place high confidence when our analysis matches either a Guessing model, a *Sides-Only* model, or a *Dots-Only* regardless of the assumed *Measurement Noise* of our models, or whether we analyze across 120 or 360 trials of the experiment. Likewise, we can place high confidence that when our analysis matches any of the three Decoder models, the behavioural data was truly generated with either of the three Decoder models. We are a bit less confident in the ability of our analysis to accurately infer between the three Decoder models when analyzing across 360 trials (using $\sigma_{n_m} = 1.645$ and $\sigma_{d_m} = 1.183$), and place even less confidence when analyzing across 120 trials.

4.2.4 Discussion

The simulations detailed here demonstrate the potential impact of an observer's *Measurement Noise* on their ability to accurately categorize object stimuli in a categorization task. There was a direct correlation between *Measurement Noise* in either the Sides or Dots feature dimensions and maximum performance. When simulants were generated with lots of noise ($\sigma = 4$), performance was at its lowest around $\approx 56\%$ correct. When simulants were generated with no noise ($\sigma = 0$), performance was at its best at around $\approx 76\%$. This was true for both Model 1 and Model 3 simulations. Had we continued to increase the simulants' *Measurement Noise*, their performance would approach $\approx 50\%$ where the simulant would essentially be guessing every trial. Even with moderate

Measurement Noise between $\sigma = 1$ and $\sigma = 1.5$, the performance decrease was still significant by $\approx 4\%$ and $\approx 6\%$.

Performance for Model 1 simulants was greatly impacted by the presence of just a single noisy feature measurement. When these simulants had high noise in one feature measurement and low noise in the other, performance was not much better than when both features had high noise with expected performance to be between $\approx 60\%$ and $\approx 63\%$. This means that the performance of an observer using Decoder model 1 will be largely dependent on their worst-performing feature measurement. However, this was not the case for Model 3 simulants. When Model 3 simulants had high noise in one feature measurement and low noise in the other, performance suffered not nearly as much as with Model 1 simulants, with performance dropping only to $\approx 70\%$. This means that an observer using decoder Model 3 is protected to some degree from the poor performance of their worst feature measurement. This makes intuitive sense as an observer aware of the unreliability of a feature measurement would be able to put less weight on those measurements and rely more on the other feature. Observers using Decoder model 1 on the other hand would be unable to take this into consideration as they are unaware that there is any internal measurement noise at all.

A concern that arose from the use of the stimuli set created for Experiments 1 and 2 was that the Sides feature and Dots feature were not measured by participants to the same acuity. This was observed through self-reports given by participants after running through the experiment, where many participants stated that they only attended to the Dots feature and had difficulty with the Sides feature. We also found evidence for this through a model comparison analysis attempting to determine whether participants were only attending to either the Sides or Dots features, using both features or guessing. Results indicated that most participants were either using both features or relying on only the Dots feature. This could occur if the participant's *Measurement Noise* for the Sides and Dots feature was unequal, with the Sides feature being worse than the Dots feature. The noisier feature measurement could have reduced performance significantly depending on which Decoder model applies to the participants in our experiments.

This concern is what drove the expanded 9 sub-models for *Sides-Only* observers, *Dots-Only* observers, and "Expert" observers who considered both features. One possible strategy for an observer who has high *Measurement Noise* for one feature is to ignore that feature entirely. This strategy might make sense considering that participants were attempting to learn the novel categories in our Experiments. The participants running through the categorization experiment were naive to the objects, the categories, and the *Within-Feature Category Means* and *Variability* before starting the experiment. Throughout repeated exposure to the categories, they were required to learn these parameters. The presence of *Measurement Noise* increases the difficulty of this task; If one feature's *Measurement Noise* is too high, the participant may be unable to gather any relevant information about the *Within-Feature Category Variability* of that feature and may opt to ignore it, leading to an incorrect understanding of the categories or treating the experiment as a single-feature categorization task. This would be especially true if participants were using Decoder model 1 as they would be greatly negatively impacted by the presence of a poor feature measurement. In those cases, ignoring the feature completely would lead to better performance.

The model comparison we outlined here is powerful, but limited by the range of hypothesized sub-models that we included in our analysis. Unfortunately, we are unable to account for all possible hypothetical sub-optimal models here. When an observer is "optimal", there is generally just one single optimal model that can be derived and compared against. However, when an observer is not optimal at a task, the problem balloons in complexity. There are an endless number of ways that an observer can be sub-optimal. With a study of this kind, we cannot exhaustively test all possible sub-optimal models. Even with our best attempt to do so, we would still be limited by the context of what we thought to consider and account for in the creation of our models, and there could still be many others that we failed to consider. What we have chosen to focus on here are interesting cases that are relevant to the overall focus of this thesis. The 9 models we included in our model comparison analysis attempt to explore the impact of the presence of *Measurement Noise* in our categorization task and roughly account for some of the observed possibilities for participant behaviour in Experiments 1 and 2.

The theoretical work we have covered thus far lays the groundwork for how our Bayesian Categorization Model can be improved upon. The model comparison analysis is a powerful approach to analyzing participant data that solves the issues faced with trying to develop more advanced analytical techniques in Chapter 3. However, to better apply these analysis methods to our data set, the *Measurement Noise* present in human participants needs to be constrained. These values can be empirically determined through further experimentation. In the following section, we will detail Experiment 3 in which we sought to empirically determine the *Measurement Noise* of the Sides and Dots features present amongst participants with our stimulus set.

4.3 Experiment 3

The presence of internal *Measurement Noise* in an observer has the potential to have a high impact on performance in a categorization task. The simulation experiment we just described reveals the extent of this theoretical impact. If an observer's *Measurement Noise* is low, their categorization performance should be unimpeded and they should be able to reach the highest expected percent correct in a categorization task. However, an observer with high *Measurement Noise* may perform significantly worse, potentially approaching near-guessing behaviour.

These results could offer insight into Experiments 1 and 2, where we found that performance against the Bayesian Categorization Model was high but not able to reach 100% accuracy. Taking into account participants' *Measurement Noise* might explain this discrepancy and lead to a more human-accurate model. This motivated the development of Experiment 3 in which we set out to determine the *Measurement Noise* present in participants exploring the stimulus object set used through our previous experiments. To this end, we employed a 2IFC experimental design procedure to determine this sensory noise.

In designing this experiment, a couple of considerations were kept in mind to expand upon the analysis of Experiments 1 and 2. The first consideration was maintaining parity with the previous experiments. To this end, the same 25 stimuli objects were used for this experiment as were used in the previous experiments. It is important to note that this object set is not ideal for use in a 2IFC experimental design. If we were solely interested in determining the sensory measurement noise of observers when measuring the number of sides or spacing of dots on objects similar to ours, we would ideally create a newly expanded stimulus set with a far higher resolution ceiling, allowing for a finer tuning in step size to get a precise and accurate estimation of each participant's noise.

There are only 5 levels per feature with the current 25 stimuli object set (that is, there are only 5 objects within the "sides" or "dots" feature dimensions). Designing a 2IFC experiment with this stimulus set greatly limits the resolution of our sensory measurement noise estimate. However, it was important to run this experiment with the pre-existing stimulus set because we were primarily interested in participant's measurement noise within the context of the previous experiments. We were not interested in the generalized accuracy of measuring dots on an object. Instead, we were interested in the accuracy of measuring dots specifically in our categorization experiment.

The second consideration was to explore the impact of the task load placed on participants during the categorization experiment. The design of Experiments 1 and 2 required participants to make a categorical judgement using both the sides and dots features after only 5 seconds of exploration. This presents a potential limit on participant performance in their ability to accurately measure the features, which could increase the measurement noise present. That is, it is possible that requiring participants to make two separate measurements in a 5-second time window might adversely impact their measurement noise. To explore this aspect of the previous experiments, we again only allowed participants 5 seconds per trial to explore the objects in making their measurements. Additionally, we compared blocks of trials where participants were told to focus on just one feature against blocks of trials where participants were told to focus on both.

Experiment 3 does not represent a thorough exploration into the sensory measurement noise of human participants when physically exploring objects to determine the number of sides or spacing of a dot matrix texture. Instead, Experiment 3 acts as a companion to the previous Experiments 1 and 2 to expand on the findings observed there and offer further insight into the discrepancy found between participant performance and our Bayesian model estimated performance in the categorization task.

4.3.1 Methods

Participants

Fifteen participants (17-23 yr old, median age 18 yr) were recruited from the McMaster University community for this experiment. Participants were screened for the following conditions known to adversely affect tactile acuity that may have impacted their performance in the experiment: diabetes, nervous system disorders, cognitive impairment, learning disabilities, dyslexia, attention deficit disorder, carpal tunnel syndrome, and arthritis of the hands. All participants were naïve to the experiment and the unique stimuli used throughout. Participants were compensated for their participation by either course credit through the university or 20$. Participants completed a short demographic and handedness questionnaire prior to participation, and an exit interview after the experiment had concluded.

Materials, Stimuli, and Software

The same twenty-five stimuli objects from Experiment 1 and 2 were used in Experiment 3. These objects were created using the Ultimaker 2 Go 3D printer and printed using PLA plastic. The objects were differentiated by two features, the number of sides and the spacing of a dot-matrix texture that covered one of the faces. All objects were otherwise identical with a radius of 25mm and a thickness of 5mm.

The two feature dimensions had 5 levels per dimension. The dot-matrix texture consisted of evenly spaced 0.5mm radius hemispheres protruding from the surface. We established 5 different levels for dot spacing and 5 levels for sides, creating 25 unique objects. The space between each dot on an object ranged from 4mm to 8mm in 1mm increments. The number of sides of each object ranged from 6 to 10 sides in steps of 1 side.

The experiment was run on a Windows 10 PC system using specialized software developed in LabVIEW 2017 (National Instruments). Like the previous experiments, a human experimenter was required to present and retrieve the physical stimuli and input participant responses into the software. However, the software itself instructed the experimenter on which objects to present and managed the progress of the experiment.

Experiment instructions were pre-recorded on audio files and played for the participant, and instructions were verbally reminded by the experimenter each block.

Procedure

The experiment consisted of 6 blocks of 48 trials each, taking a maximum of 3 hours to complete. The blocks were separated into three types, where either the participants' perception of the sides feature, the dots feature or both the sides and dots features in conjunction were under investigation. The two blocks where the participant was told to focus on only a single feature were considered *Low Task Load* blocks, and the block where the participant was told to focus on both features was considered *High Task Load* blocks. These three types of blocks were repeated twice, leading to a total of 6 experimental blocks. The blocks in the first half of the experiment were referred to as *First-Half* blocks, and the blocks in the latter half of the experiment were referred to as *Second-Half* blocks.

During a testing block, participants sat at a desk behind an opaque screen which blocked the participants' vision. Participants were handed two stimulus objects per trial, one after the other, and were given 5 seconds to explore each one. At the end of each 5-second interval, participants immediately placed the object down. After exploration of the second object, participants were asked to respond with either "first" or "second" which stimulus object had more of the feature being investigated. In the sides block, they were asked to respond to which object had more sides. In the dots block, they were asked to respond to which object had more dots. In the both sides-and-dots block, they were asked to respond first which object had more sides, then which object had more dots.

The presentation of objects followed a method of constant stimuli procedure. A reference object was chosen with Sides of $n = 8$ and Dot Spacing of $d = 6$ (corresponding to feature level 3 for both Sides and Dots). To avoid any possible confound of having the same physical object presented repeatedly in succession during the experiment, this reference object was duplicated with an identical newly 3D printed object. Comparison objects were selected from our set of 25 objects, with five feature levels for both the Sides and Dots features possible. During each block, the comparison objects were pseudo-randomly chosen in such a way that the feature levels were presented to the participant 10 times. The order of presentation of the reference object and comparison object was randomized in every trial.

During the Sides *Low Task Load* block, participants completed 40 trials comparing the reference object against four comparison objects that varied only by Sides ($n = 6, 7, 9, 10; d = 8$). During the Dots *Low Task Load* block, participants completed 40 trials comparing the reference object against four comparison objects that varied only by Dots ($n = 8; d = 4, 5, 7, 8$). During the *High Task Load* block where attention was required toward both features, participants completed 48 trials comparing the reference object against the four feature levels of both the Sides and Dots features, resulting in 24 comparison objects in total. Each combination of Sides and Dots feature level was presented twice. In this way, 10 comparisons were made between the reference object and a comparison object for each feature level within either the Sides or Dots feature. The comparison between the reference object and its identical feature level object was excluded leading to 48 total trials that allowed us to compare the 10 comparisons per feature level with the *Low Task Load* blocks in this *High Task Load* condition where participants were attending to both.

4.3.2 Results

To calculate participant measurement sigmas, we fit cumulative Gaussian psychometric functions with $\mu = 0.5$ and σ as a free parameter to the proportion of trials within a block in which the comparison object was reported to be larger than the reference object. We did this separately for the Sides and Dots features. We also split measurement sigma calculations between the first and second halves of the experiment to determine if there was any training effect in measurement sigma as participants had continued exposure to the object set. Finally, we split measurement sigma calculations between *Low Task Load*, where participants only needed to focus on either the sides or dots feature alone, and *High Task Load*, where participants needed to focus on both the sides and dots features at the same time.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} dt \tag{4.5}$$

where x is the difference in feature value between the standard stimulus and the comparison stimulus. Figure 4.8 displays the main results found from this experiment. Measurement sigmas for the dots feature were lower than those for the sides feature. This was consistent across the *First-Half* blocks and *Second-Half* blocks, as well as the *Low* and *High Task Load* blocks. Increasing the task load also consistently increased measurement sigma for both the Sides and Dots features. Participants on average decreased their measurement sigma for sides throughout the experiment, but not for dots. A Repeated Measures ANOVA revealed significant within subjects effects of Feature $F(1,15) = 8.049$, $p = .012$, Training, $F(1,15) = 14.399$, $p = .002$, and task load, $F(1,15) = 20.086$, $p < .001$, and a significant interaction between Feature and Training, $F(1,15) = 6.978$, $p = .018$.

The *Second-Half*, Heavy task load experiment condition represents the most relevant context in comparison to the previous Experiments 1 and 2. The average best-fitting measurement sigmas across our participant pool here were $\sigma_n = 1.645$ (SD=1.231) for the sides feature and $\sigma_d = 1.183$ (SD=0.368) for the dots feature. Figure 4.9 plots the psychometric functions corresponding to these two measurement sigmas.

Figure 4.10 displays the full range of best-fitting measurement sigmas for the Sides feature and Dots feature across all 15 participants. Error bars displayed represent

95% confidence intervals. This shows the full range of variation observed across our participant pool, with some participants reaching a fairly low measurement noise floor and others reaching very high measurement noise.

4.3.3 Discussion

Experiment 3 provides greater context in understanding the results found in Experiments 1 and 2. Through this 2IFC experiment, we aimed to explore three main questions: is there a difference between participant measurement sigmas for the Sides feature and Dots feature, does increasing task load impact these measurement sigmas, and can this measurement sigma change and improve throughout the experiment?

Measurement sigmas were lower for Dots versus Sides feature measurements. This is in line with the results and observations found in Experiments 1 and 2, where it appeared that participants were worse at measuring the Sides feature than the Dots feature. We suspected this to be the case from the model comparison analysis where most participants were estimated to either be using both features or only the Dots feature, but not the Sides feature. This also came out through participant self-report when we asked participants to verbally describe any strategy they used. Many participants reported ignoring the Sides feature. The worse measurement noise for Sides than Dots found in Experiment 3 supports the idea that participants were having difficulty with this feature in Experiments 1 and 2.

In terms of task load, we found that measurement sigmas became worse when we asked participants to attend to both features in a trial than when they were only required to attend to one. This effect was present for both the Sides feature and the Dots feature. Interestingly, the impact of task load was uniform for both feature measurements, meaning that both the Sides and Dots measurement sigmas increased by the same amount when participants attended. This *High Task Load* condition matched the context of the previous categorization experiment, where participants were required to pay attention to both features when making their categorical judgments. The effect of task load provides interesting insight for designing future experiments and considerations for analysis and interpretation.

We observed an experiment half effect throughout the experiment, with measurement sigma for the Sides feature decreasing significantly by the latter half of the experiment, but not as much for the Dots feature. We also observed an interaction between experiment half and task load, with the effect of task load diminishing in the *Second-Half* blocks.

Overall, through the 2IFC experiment detailed here, we have empirically determined an average *Measurement Noise* value for both the Sides feature and Dots feature in the context of our experimental categorization task. These values can be applied to our analytical methods developed earlier in this chapter to constrain the *Measurement Noise* parameters to those realistically found among our human participants.

Estimating Measurement Thresholds with limited exploration time and a limited stimulus set

For Experiment 3, we set out to get an idea of participant measurement thresholds with the features that we were using throughout our other experiments. One of the most important aspects of our methodology was to match as closely as possible the experimental conditions of our previous categorization experiments. There were a few aspects we identified as being critical to our unique design: the stimulus objects, the trial-by-trial task requirement, and the trial length were all constrained to match our categorization task closely. Importantly, the goal of our methodology was not to measure participants' absolute best ability to extract the number of sides feature or dots feature, but to measure their ability within the context of our experiment.

In regards to the trial time, we were not interested in determining the amount of time needed to measure the number of sides or dots as accurately as possible. We expect that imposing a time limit lowers performance as the participant might not have enough time to make an accurate measurement. Given an infinite amount of time to explore the objects, a participant should be able to extract near-perfect measurements of the object. Likewise, if we gave the participants less time their performance would reasonably decrease. Naturally, then there should be an optimal amount of time for each participant that they would need to explore the object to extract the information accurately, and that time likely is not the 5 seconds we used. Our 5-second trial time was decided upon for the categorization task as a way to constrain the length of the experiment. Increasing the trial length would increase the total experiment time, and to keep within a reasonable experiment length we would have instead had to lower the number of experimental trials. We decided on our 5-second trial length to balance out having enough trials within a 2-hour experiment.

In regards to the stimulus objects, a major limitation of our threshold task is the poor stimulus resolution offered by our stimulus set. Out of our 25 objects, we only have five stimulus levels for each feature dimension (6 sides to 10 sides, and 4mm to 8mm with a 1mm step size). This means that the fixed reference object can only be compared against four other comparison objects. Had we been interested purely in measuring a participant's ability to discern differences in the number of sides or the spacing of dots, we would ideally have had many more stimulus levels to compare against. However, it was important for us to stay within the context of our categorization experiment, and that meant using the same objects for that task. There are some physical difficulties as well that led us to use our categorization stimulus set for this experiment. A higher resolution would have required us to 3D print many more objects which would have not only been time-consuming, but difficult both to store and work with during the experiment. A higher resolution would have also only been possible for the Dot Spacing feature, as the Number of Sides feature is physically constrained to integer values.

A major issue arises when analyzing measurement sigmas with a low-resolution stimulus set. With only four levels of comparison, we have poor precision in determining the actual measurement sigma of the observer. This becomes especially problematic when a participant exhibits very high performance. In these cases, their measurement thresholds result in a step function with an infinitesimally small measurement sigma. Essentially, the participant's sigma is smaller than our stimulus set allows us to measure. This creates an interesting analysis problem; how do we report a participant's measurement sigma for a feature when the threshold we extract bottoms out at our minimum estimate?

One approach is to consider the smallest threshold we can measure before this issue of bottoming out (what is the threshold when someone makes a single mistake?). Another approach is to consider how often we reasonably expect people to have a perfect performance at different sigmas. We can find this by simulating many blocks of trials at different sigma levels. For each sigma level, we can calculate how many times, when two objects are drawn with a difference of 1, the participant perfectly identifies which object is larger. For instance, an observer with a measurement sigma of 0.6 doing 10 trial blocks 100 times is expected to have a perfect block about 20% of the time.

Despite the similarities from the previous categorization experiment in terms of procedure and the use of the same set of stimuli, this is still a fundamentally different task being asked of the observers participating in the experiment. The main difference between the categorization task and the 2IFC task is the sources of variability present that the observer must account for. In the categorization task, both the *Within-Feature Category Variability* and *Measurement Noise* are present and impact the judgement being made by the observer. Categorization requires the observer to first form a percept of the feature properties that they perceived and then form a categorical decision based on that percept. The 2IFC task in contrast does not involve this *Within-Feature Category Variability* or this extra layer of decision making. All that is required of the participant is to accurately distinguish the feature differences.

Why does increasing the demands of the task decrease Sides and Dots measurement acuity?

There are a few possible explanations for why *High Task Load* decreased measurement acuity with our stimuli object set. For instance, increasing the demands during the experiment may make the task physically more difficult. The nature of our experimental design involving physical objects means that extraction of feature information from the stimuli requires physical exploration of the objects. The exploration strategies for extracting haptic information for different types of features are different; requiring participants to perform both exploration strategies in a short time window may have impeded their performance. As we outlined previously in Chapter 3, we expected participants to employ *Lateral Motion* to explore the Dots feature and either *Contour Following* or *Enclosure* to explore the Sides feature (Klatzky et al., 1993). It was expected that these exploration motions could be performed simultaneously; however, this may have posed a greater challenge for participants than originally thought.

Additionally, participants may not have used these optimal exploration strategies and instead relied on other means to explore the objects. Especially with the Sides feature, it was observed that participants often attempted to count the sides rather than feel for the overall shape as we had originally hypothesized. This may be because the number of sides was a low integer that feasibly *could* be counted in the 5-second trial time. However, when paired with the physical exploration of the dots, counting 6 to 10 sides accurately may not have been possible in that time frame. This could explain the worse measurement acuity for the Sides feature than the Dots feature.

The performance decrease with the increased task load may also be an effect of splitting attention. Requiring observers to attend to both features in the 5-second time window means that both feature measurements must be held in memory. This memory must remain intact while the observer measures both objects, as a comparison and reference object was observed sequentially in each trial (each with a 5-second exploratory period). This means that by the time a participant is providing their response, some of this memory will need to have persisted for more than 10 seconds. The additional information required to keep in memory may make this task more difficult, leading to the performance decrease.

4.4 Reanalyzing Experiments 1 & 2

As we have highlighted throughout this chapter, the Bayesian model used when analyzing Experiments 1 and 2 was limited by the exclusion of the presence of internal *Measurement Noise*. From our simulations, we found that *Measurement Noise* has a strong influence on the response behaviour of our models, regardless of whether the decoder model ignores it (represented in Decoder model 1) or accounts for it (represented in Decoder models 2 and 3). This impacts the expected performance of our models on our categorization task as well as the expected trial-by-trial behaviour in how the model categorizes each object.

We can now return to our collected data sets from Experiments 1 and 2 and reanalyze them using the model comparison analysis methods established earlier in this chapter. Through our 2IFC Experiment 3, we found that human participants had an average *Measurement Noise* sigma value of $\sigma_n = 1.645$ for the Sides feature and $\sigma_d = 1.183$ for the Dots feature. By accounting for the presence of this *Measurement Noise* in our updated Bayesian Categorization Model, we can more accurately match our participants' data to our Bayesian models to better determine how well our model accounts for that human behaviour.

Here we will use the empirical *Measurement Noise* sigmas found in Experiment 3 as a rough estimate for the measurement noise present and apply them to the model to see if this yields us a closer fit. We will run our model comparison analysis to fit each participant into their closest matched model, and then use that matched model to compare against the participant's performance. In the future, a more complete experiment may include the pairing of a 2IFC task alongside the categorization experiment to have within-participant sigma noise values to apply to the Bayesian Categorization Model.

Measurement Noise Ratios

The average *Measurement Noise* sigmas found in Experiment 3 across 15 participants was $\sigma_n = 1.645$ for the Sides feature and $\sigma_d = 1.183$ for the Dots feature. However, it is important to acknowledge that using these average noise sigma values will lack a degree of accuracy as we found there was variability in individual participant *Measurement Noise* values for both the Sides feature and the Dots feature. Ideally, we would need to know an individual participant's true noise sigma values to most accurately match them to our models. Unfortunately, we cannot retroactively know the actual *Measurement Noise* sigmas for our participants in Experiments 1 and 2.

As we illustrated by examining the response tables generated from different configurations of *Measurement Noise* sigmas (Figure 4.1), and the accuracy of our model comparison analysis (Figure 4.7), the precise sigma values used are not as important for our analysis to correctly classify the generative model behind a set of behavioural data. What matters most is the ratio between the Sides feature sigma and the Dots feature sigma, and how close our assumed values ratio is to the true generative model's ratio. This assumed ratio must be in the same direction as the true ratio of our participants.

The ratio of Sides:Dots for the average sigmas we found across the 15 participants in Experiment 3 was 1.39, with the Sides noise sigma being larger than the Dots noise sigma. Of these participants, 10 exhibited a ratio in this direction. Of these participants, their average *Measurement Noise* sigma values for the Sides and Dots features were $\sigma_n = 2.13$ and $\sigma_d = 1.05$, resulting in an average ratio of 2.03. The 5 remaining participants exhibiting a ratio in the reverse direction had average *Measurement Noise* sigma values of $\sigma_n = 0.59$ and $\sigma_d = 1.48$, resulting in an average ratio of 0.39.

Since we found that 1/3 of our participants in Experiment 3 had this reversed direction between their *Measurement Noise* sigmas, we might expect that a significant number of participants in Experiments 1 and 2 could also have this pattern of *Measurement Noise*. While we cannot verify this directly with our previously collected data set, we can attempt to account for this possibility with our model comparison analysis. We can instead consider two possible "types" of observers: one whose Sides feature sigma is worse than their Dots feature sigma, and one whose Sides feature sigma is better than their Dots feature sigma. We will label this second type of observer *reverse* observers, signifying the reversal of their Sides:Dots ratio.

Rather than using the average sigma values of $\sigma_n = 1.645$ and $\sigma_d = 1.183$ across our models, we can split our models further into the two types of observers. We will label these additional *reverse* models as $D1R$, $D2R$, $D3R$, DsR, and DdR. The original 6 models used in the model comparison analysis will use the average sigmas of the participants whose Sides noise sigma was larger than their Dots noise sigma, $\sigma_n = 2.13$ and $\sigma_d = 1.05$. The *reverse* models will use the average sigmas of the participants whose Sides noise sigma was smaller than their Dots noise sigma, $\sigma_n = 0.59$ and $\sigma_d = 1.48$. By including both types of models in our model comparison analysis, we can account for the possible range in participant variation and lead to better fits per participant.

To test the efficacy of our expanded model comparison analysis, we re-ran our confusion matrix simulation analysis comparing the frequency of times our analysis matched each model to every possible generative model. We ran these simulations for 50 sets of 120 trials of our categorization task and calculated the frequency of times the data from each set of trials was matched to each model. The results of this confusion analysis are shown in Figure 4.12. We found that the analysis remained generally very accurate at correctly matching behavioural data to the true generative models. The analysis was worst at matching to model $D2$, correctly matching $D2$ 60% of the time. It remained best at matching Dr 100% of the time. All of the other models were correctly matched within 70% to 94% of the time.

The expanded model comparison analysis remains fairly powerful at accurately matching behavioural data to the generative models behind that data. While this comparison is still not exhaustive of all possible generative models, it represents a fairly decent representation of the range of generative models behind what we might expect from our participant pool. In the following analyses, we will first run our participants through the original formation of our model comparison analysis with 6 comparison models and sigma values of $\sigma_n = 1.645$ and $\sigma_d = 1.183$. Afterwards, we will move to our expanded model comparison analysis using sigma values of $\sigma_n = 2.13$ and $\sigma_d = 1.05$ for the main set of models and $\sigma_n = 0.59$ and $\sigma_d = 1.48$ for the *reverse* set of models.

4.4.1 Model Comparison Analysis

The participant pool for our model comparison analysis contained the 15 participants from Experiment 1 who were in training regimen Group 3 and the 10 participants from Experiment 2. Of the Experiment 2 participants, only the first 360 trials (corresponding to the first 9 blocks, or the first day, of the experiment) were included. The first 9 blocks of Experiment 2 were designed as a replication of Experiment 1, allowing us to combine these participant pools for a total of 25 participants.

First, we ran our model comparison analysis using $\sigma_n = 1.645$ for the Sides feature *Measurement Noise* sigma and $\sigma_d = 1.183$ for the Dots feature *Measurement Noise* sigma. We previously established through our simulations that we can make post-hoc inferences as to which Decoder model an observer may have been using throughout our categorization experiment. Following the model comparison method we described through Equations 4.3 and 4.4, we ran our model comparison analysis for 3 sets of 120 trials (each set corresponding to 3 blocks of the experiment). Only the results from the last 3 blocks of the experiment were considered for deciding overall which Model fit the participant the best, as this corresponded to when the participant had presumably learned the novel categories to the best of their abilities within the time frame of the experiment. The models matched against in our analysis will be the same ones used in our previous simulations, with $D1$, $D2$, and $D3$ corresponding to Decoder models 1, 2, and 3 respectively, Dr corresponding to a control "guessing" model, and Ds and Dd corresponding to a *Sides-Only* and *Dots-Only* model respectively. We ran these models with equal priors of $1/6$ initialized in our trial-by-trial analysis for each set of 120 trials. We first examined the frequencies of matches to models $D1$, $D2$, and $D3$ combined against Dr, Ds, and Dd, as we established through our sensitivity analysis that our analysis is generally very accurate at distinguishing between these comparisons. Afterwards, compared against models $D1$, $D2$, and $D3$ to see how our analysis matched between these three models.

Figure 4.13 (a) displays the frequency that either $D3$, Ds, Dd or Dr was matched to participants. 8 participants were matched to $D3$, and 11 participants were matched to the *Dots-Only* model Dd. 4 were matched to the *Sides-Only* model Ds, and 2 were matched to the guessing model Dr. Expanding this frequency analysis to separate $D1$, $D2$, and $D3$ (Figure 4.13 b) revealed that 2 participants were matched to $D1$, 3 were matched to $D2$, and 5 were matched to $D3$. Overall, the Decoder models that accounted for both the Sides and Dots features as well as the *Dots-Only* model fit the majority of our participants. Finding a large number of participants matched to Dd corresponds with our previous findings in Experiments 1 and 2 where participants seemed to focus more on the Dots feature than the Sides feature.

Next, we ran our expanded model comparison analysis using sigma values of $\sigma_n = 2.13$ and $\sigma_d = 1.05$ for the main set of models and $\sigma_n = 0.59$ and $\sigma_d = 1.48$ for the *reverse* set of models ($D3R$, DsR, and DdR). We again ran these models with equal priors of $1/7$ initialized in our trial-by-trial analysis for each set of 120 trials. Figure 4.14 (a) again reveals that the majority of our participant pool was matched to $D3$ or Dd. This is illustrated more clearly in Figure 4.14 (b) where we combine the data of each model with the corresponding *reverse* model.

Figures 4.14 (c) and (d) display the fully separated histogram of matched models, including all three decoder models and their *reverse* models in our analysis. Here we see again that Dd was the most frequently matched model with 11 participants matched to it. 3 participants were matched to $D1$, 4 participants to $D2$, 4 participants to $D3$, 2 participants to Ds, and 1 participant to Dr. If we take all the Decoder models together, we find that our analysis matched 11 participants to either $D1$, $D2$ or $D3$.

Investigating the two participants who originally were matched to Dr further, we found that participant 9 from Experiment 1 was reclassified to DdR, and participant 1 from Experiment 2 remained classified to Dr. This indicates that participant 9 from Experiment 1 was likely erroneously matched to Dr as there were no better fitting models, and that participant 1 from Experiment 2 truly fits Dg the best. When we expanded the range to include the three Decoder models (Figures 4.14 d), we found that 5 participants were matched to $D2$, 3 participants were matched to $D1$, and 4 participants were matched to $D3$. Dd remains the most matched model but with now 9 participants matched to it, indicating, that two participants were switched to being matched to one of the Decoder models.

4.4.2 Best Matched Model Fitting

To test how well our matched models fit their respective participant data, we ran each participant's matched model through simulated trials of the participant's experiment. The participant's matched model was simulated running through the same trials as the participant did for each block of 40 trials using its generated response table $P(''A''|\mathbf{f}_g)$. We ran these simulations 10,000 times per block and calculated an average percent correct expected from the model given that block of trials. Additionally, since we ran our model comparison analysis for sets of 3 blocks (120 trials), we used the matched model within these sets of blocks. For instance, if a participant was matched to Dd in the first 6 blocks but then $D3$ in the last 3 blocks, our expected performance simulations would use the respective matched models to each block.

This performance simulation analysis explores how well our Bayesian Categorization Model fits the behavioural data of our human participants. It also controls for trial-by-trial variability in difficulty, since the models are running through the same trials and presented the same objects as our participants. Unfortunately, we cannot do the same "Percent match to model" analysis as we did in Chapter 3 (Figure 3.1) as once our models involve the inclusion of *Measurement Noise* they become a stochastic process and respond with some variability to the same stimuli. By simulating through a block of trials many times and getting a distribution of the percent correct calculated for those simulations, we can determine if the participant's actual recorded percent correct is within a reasonable range of the distribution of expected percent correct.

We also calculated the expected performance of our previous Bayesian Categorization Model used in Chapter 3 which did not include *Measurement Noise* at either the Encoder or Decoder stage. We will hereafter refer to this previous model as the *noiseless* Bayesian Categorization Model, as it was unaware of and did not account for the presence of *Measurement Noise* entirely (neither at the Encoder nor Decoder stage). This model was similarly run through the same trials as the participant did, with a percent correct calculated for each block. This allows us to directly compare the performance of our previous iteration of our model to our new iterations which account for *Measurement Noise*.

As we did when analyzing our data in Experiments 1 and 2, we excluded trials where our "Expert" observer model would calculate a posterior probability of $P(A|D) = P(B|D) = 0.5$ for a given object. This posterior probability calculation is taken from our original *noiseless* Bayesian Categorization Model which does not include Measurement Noise in either the Encoder or Decoder portions of the model. As such, the performance of this *noiseless* model is deterministic and does not include any variability.

Figure 4.15 displays three example participants from Experiment 1 of this analysis. For each participant, we calculated their percent correct per block, the distribution of our matched model's percent correct per block (with the solid line indicating the mode of the distribution and the dotted lines indicating 1 standard deviation), and the *noiseless* model's percent correct per block. Participant 3 was matched to the *Sides-Only* model Ds for blocks 1-6 (although our analysis also competed against Dr for these blocks), and Decoder model 3 $D3$ for blocks 7-9. Participant 15 was matched to the *reverse Dots-Only* DdR model for blocks 1-6, and Decoder model 2 $D2$ for blocks 7-9. Participant 33 was matched to the Decoder model 2 $D2$ for blocks 1-6, and the *Dots-Only* Dd for blocks 7-9. For these example participants, the performance of our matched models fit more closely to the participant data than the *noiseless* Bayesian Categorization Model.

We expect that our analysis is the least trustworthy at the beginning blocks of the experiment for each participant as they were required to learn the novel categories throughout the experiment. As such, they did not initially know the distributions underlying the categories and would be learning these parameters as the experiment continued. Our model comparison analysis does not account for learning. As such, it is most informative to focus on the data from the last 3 blocks of the experiment (blocks 7-9) when examining the closeness of fit, as this should represent when participant learning of the category distributions is complete.

We can describe the fit of our matched models to participant data further by running a one-sample z-test per testing block. We compared the calculated percent correct of each participant against the simulated percent correct of their matched model, treating the matched model as the population. We simulated the population distribution of the matched model by running it through 10,000 simulated blocks following what the participant was presented in their respective block. The calculated mode and standard deviation of these simulated trials were used as a population mean and standard deviation for our z-test. The null hypothesis in this context is that the participant's percent correct equals that of the matched model's mean percent correct.

We calculated the z-score per block and compared it against a significance level of 0.05. Examining the results of the last 3 blocks across all participants, we failed to reject the null hypothesis in almost every instance. Only participants 9, 21, and 30 from Experiment 1 and participant 8 from Experiment 2 had one of their last 3 blocks reject the null hypothesis. This null result helps describe the comparison between our participant results and matched model results, where the population of our matched model's results could have generated the observed participant's data.

This z-score test requires us to assume that percent correct is a normally distributed variable, which is potentially an incorrect assumption as the relationship between percent correct and the *Measurement Noise* which drives it is not necessarily linear. A better way to measure the goodness of fit of our matched models to participant data is to perform a Posterior Predictive Model Check. This method directly compares the performance predicted by our matched models and the actual performance of our participants. To do this, we took the matched model for each participant block of data and ran 10,000 simulations of that block, calculating the percent correct for each run. We then compared the average model percent correct of the simulated blocks against the participant's percent correct for that block. If the model's average percent correct was lower than the participant's, we then compared each simulated percent correct against the participant's and calculated the frequency of times the model's percent correct was larger. If the model was higher, we made the same comparison but calculated the frequency of times the model was lower than the participant's. This method resulted in a one-tailed p-value of how many times the model's prediction was further than the participant's.

We performed this analysis across participant data from Block 9. We defined four bins signifying different levels of fit of our model. A p-value between 0 – 0.05 was defined as a very poor fit since this would indicate that the participant percent correct is very far from the model's mean percent correct. A p-value between 0.05 – 0.1 was defined as a moderate fit. A p-value between $0.1 - 0.25$ was defined as a good fit. Finally, a p-value between 0.25 – 0.5 was defined as a very good fit. The theoretical maximum for a best-fitting model was 0.5 as this would mean that half of all model calculated percent correct were above the participant's and half were below, which would occur when the participant's percent correct was equal to the model's average percent correct.

The histogram of our Posterior Predictive Model Check results is displayed in Figure

4.16. This corresponds with the expanded matched model data in Figure 4.14 (c) and (d). For most of our participants, their matched models proved to be either a good fit or a very good fit. 5 participants were classified as having only a moderate fit to their matched model. Only 2 participants (participant 30 from Experiment 1 and participant 5 from Experiment 2) were classified as having a poor fit to their matched model, one of which was the same participant who matched poorly with our z-score test. Overall, this supports the idea that the use of our matched models through our model comparison analysis led to a better-fitting model with high predictive power.

Further comparing the average percent correct of our participants against the average percent correct of the matched models per block reveals a strong fit between our distributions (Figure 4.17). The average percent correct of the matched models follows participant data very closely, especially in the latter half of the experiment. Further comparing against the performance of the *noiseless* Bayesian Categorization Model used previously through Chapter 3 demonstrates that the addition of *Measurement Noise* in our models has decreased expected performance to match more closely to our human participants.

In order to more empirically describe the comparison between our average participant performance and average matched model performance, we ran a Repeated Measures ANOVA with experiment block as the repeated measure factor, and group (participant vs matched model) as the between-subjects factor. We limited this analysis to only the last 3 blocks of the experiment as this is when participant learning of the category distributions would theoretically be completed. We failed to find a significant difference between our participants' performance and their matched model's performance, $F(2, 1) = 0.291$, $p = .592$, with Mauchly's Test not revealing sphericity violation, $X^2 = 0.479$, $p = .787$. This lends further support to our hypothesis that the Bayesian Categorization Model which accounts for *Measurement Noise* better represents participants' behavioural data in our categorization task.

4.5 Conclusion

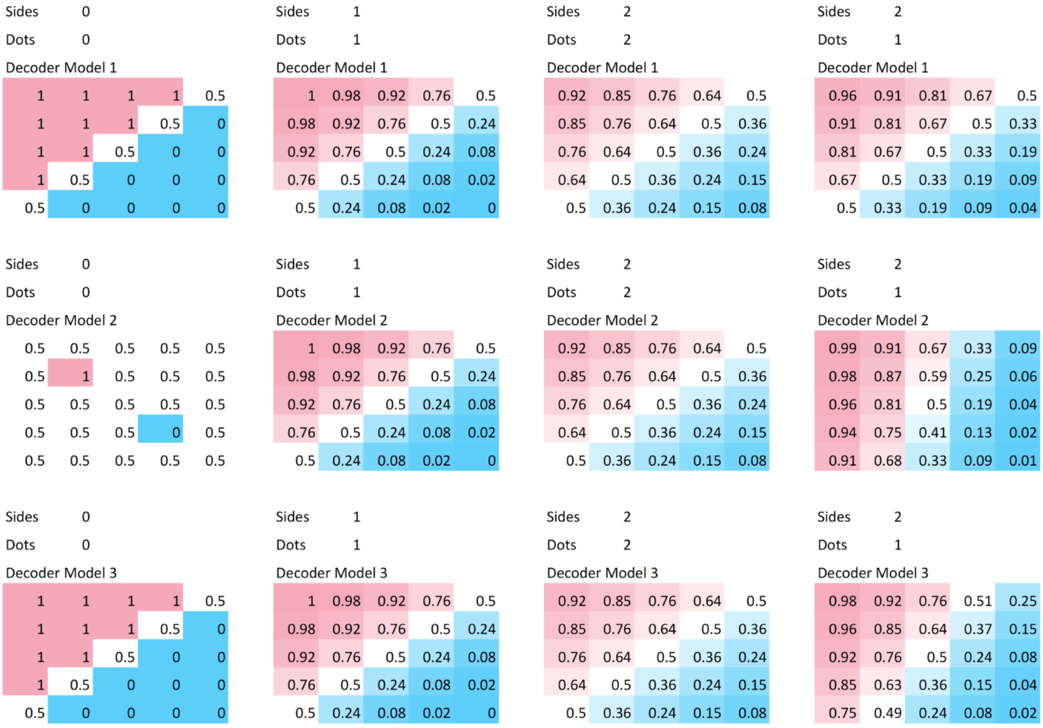
The experiments and simulations detailed here provide a continued exploration into the object categorization paradigm explored throughout this thesis and a thorough justification of the importance of considering internal *Measurement Noise* in observer models. The simulation of our categorization task with simulant observers generated with various noise sigmas revealed that the presence of *Measurement Noise* can greatly impact the performance of observers, whether the observers are aware of this internal noise or not. This justified the need for consideration of *Measurement Noise* in the context of our experiments, whereas it had previously been ignored. Experiment 3 attempted to find an empirical estimation of the *Measurement Noise* present within Experiments 1 and 2 across participants by employing a 2IFC experimental design. These average best-fitting noise sigma values could then be used to re-analyze the results of Experiments 1 and 2 with better-fitting models which accounted for the presence of *Measurement Noise*.

We also provided empirical evidence that the features chosen for our stimuli object set, the Number of Sides and Dot matrix texture, likely had differing *Measurement Noise* with the Sides measurement being worse than the Dots measurement. We also determined that both these noise sigmas were on average at a level that would have a significant performance impact on our categorization task. The nature of the *Measurement Noise* present helps to explain the results seen in Experiments 1 and 2.

Returning to our data collected throughout Experiments 1 and 2, we applied our updated Bayesian Categorization Model to examine if the inclusion of *Measurement Noise* led to a better fitting model. We found that the inclusion of *Measurement Noise* lowered the expected performance of our models which better matched the performance of our human participants. Further matching different configurations of our model to each participant allowed us to fit each participant to a best-fitting model through our model comparison analysis. We found that these matched models fit participant data very well.

One of the major questions left at the end of Chapter 3 was the source of the performance gap between our human participants and our Bayesian Categorization Model. Why did our model consistently perform much better than human participants on our categorization task? Here have demonstrated that the inclusion of *Measurement Noise* in our Bayesian Categorization Models led to a better-fitting model for each participant. Overall, this chapter has provided evidence and justification for the importance of considering the presence of *Measurement Noise* in a categorization task.

4.6 Figures and Captions



Sides 0
Dots 0
Decoder Model 1

1	1	1	1	0.5
1	1	1	0.5	0
1	1	0.5	0	0
1	0.5	0	0	0
0.5	0	0	0	0

Sides 1
Dots 1
Decoder Model 1

1	0.98	0.92	0.76	0.5
0.98	0.92	0.76	0.5	0.24
0.92	0.76	0.5	0.24	0.08
0.76	0.5	0.24	0.08	0.02
0.5	0.24	0.08	0.02	0

Sides 2
Dots 2
Decoder Model 1

0.92	0.85	0.76	0.64	0.5
0.85	0.76	0.64	0.5	0.36
0.76	0.64	0.5	0.36	0.24
0.64	0.5	0.36	0.24	0.15
0.5	0.36	0.24	0.15	0.08

Sides 2
Dots 1
Decoder Model 1

0.96	0.91	0.81	0.67	0.5
0.91	0.81	0.67	0.5	0.33
0.81	0.67	0.5	0.33	0.19
0.67	0.5	0.33	0.19	0.09
0.5	0.33	0.19	0.09	0.04

Sides 0
Dots 0
Decoder Model 2

0.5	0.5	0.5	0.5	0.5
0.5	1	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0	0.5
0.5	0.5	0.5	0.5	0.5

Sides 1
Dots 1
Decoder Model 2

1	0.98	0.92	0.76	0.5
0.98	0.92	0.76	0.5	0.24
0.92	0.76	0.5	0.24	0.08
0.76	0.5	0.24	0.08	0.02
0.5	0.24	0.08	0.02	0

Sides 2
Dots 2
Decoder Model 2

0.92	0.85	0.76	0.64	0.5
0.85	0.76	0.64	0.5	0.36
0.76	0.64	0.5	0.36	0.24
0.64	0.5	0.36	0.24	0.15
0.5	0.36	0.24	0.15	0.08

Sides 2
Dots 1
Decoder Model 2

0.99	0.91	0.67	0.33	0.09
0.98	0.87	0.59	0.25	0.06
0.96	0.81	0.5	0.19	0.04
0.94	0.75	0.41	0.13	0.02
0.91	0.68	0.33	0.09	0.01

Sides 0
Dots 0
Decoder Model 3

1	1	1	1	0.5
1	1	1	0.5	0
1	1	0.5	0	0
1	0.5	0	0	0
0.5	0	0	0	0

Sides 1
Dots 1
Decoder Model 3

1	0.98	0.92	0.76	0.5
0.98	0.92	0.76	0.5	0.24
0.92	0.76	0.5	0.24	0.08
0.76	0.5	0.24	0.08	0.02
0.5	0.24	0.08	0.02	0

Sides 2
Dots 2
Decoder Model 3

0.92	0.85	0.76	0.64	0.5
0.85	0.76	0.64	0.5	0.36
0.76	0.64	0.5	0.36	0.24
0.64	0.5	0.36	0.24	0.15
0.5	0.36	0.24	0.15	0.08

Sides 2
Dots 1
Decoder Model 3

0.98	0.92	0.76	0.51	0.25
0.96	0.85	0.64	0.37	0.15
0.92	0.76	0.5	0.24	0.08
0.85	0.63	0.36	0.15	0.04
0.75	0.49	0.24	0.08	0.02

FIGURE 4.1: Intensity plots displaying the generated "coin-flip" response tables $P(''A''|\mathbf{f}_g)_g$ for Decoder Models 1, 2, and 3. These were generated for varying levels of measurement noise σs. Column one shows Decoder Models 1, 2, and 3 response tables when $\sigma_{nm} = 0$ and $\sigma_{dm} = 0$. Column two shows Decoder Models 1 2 and 3 response tables when $\sigma_{nm} = 1$ and $\sigma_{dm} = 1$. Column three shows Decoder Models 1, 2, and 3 response tables when $\sigma_{nm} = 2$ and $\sigma_{dm} = 2$. Column four shows a mismatch condition for Decoder Models 1, 2, and 3 response tables when $\sigma_{nm} = 2$ and $\sigma_{dm} = 1$.

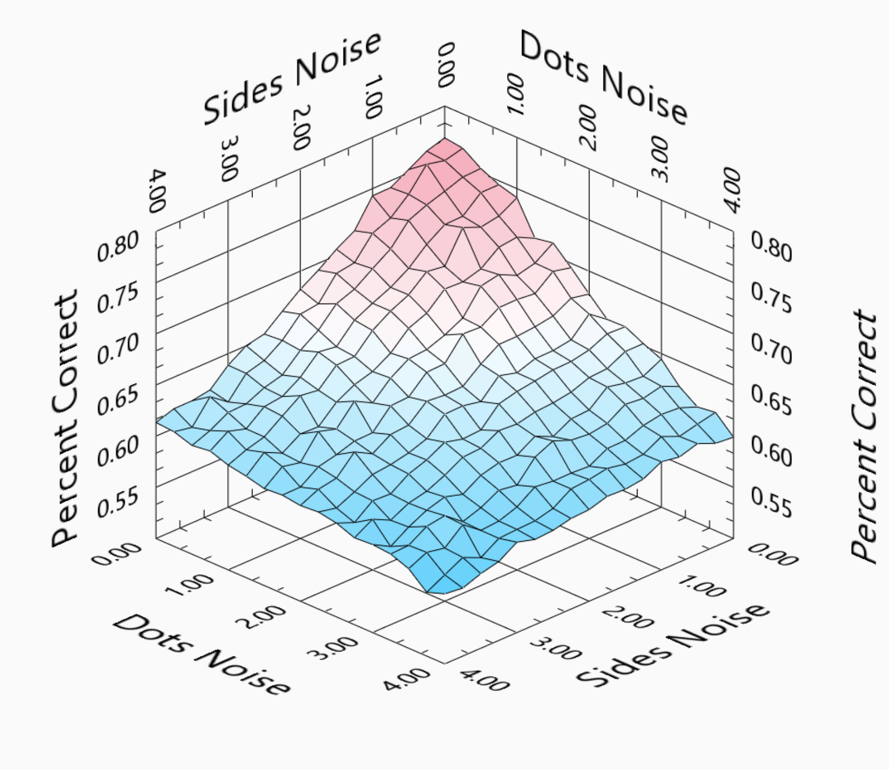


FIGURE 4.2: Results for experiment simulation of observer Decoder model 1. We generated 17 unique simulants per feature dimension for 289 total simulants with noise sigmas ranging from $\sigma = 0$ to $\sigma = 4$. Each simulant was run through 10,000 trials of a virtual categorization task identical to Experiment 1. Plotted is performance percent correct against the dot spacing feature noise and sides feature noise. The lowest percent correct was found with the simulant with the highest measurement noise at $\approx$ 56%, and the highest percent correct was found with the simulant with the lowest measurement noise at $\approx$ 77%. In the mismatch cases where one feature noise was high and the other was low, performance was greatly impacted with the percent correct recorded at $\approx$ 60%.

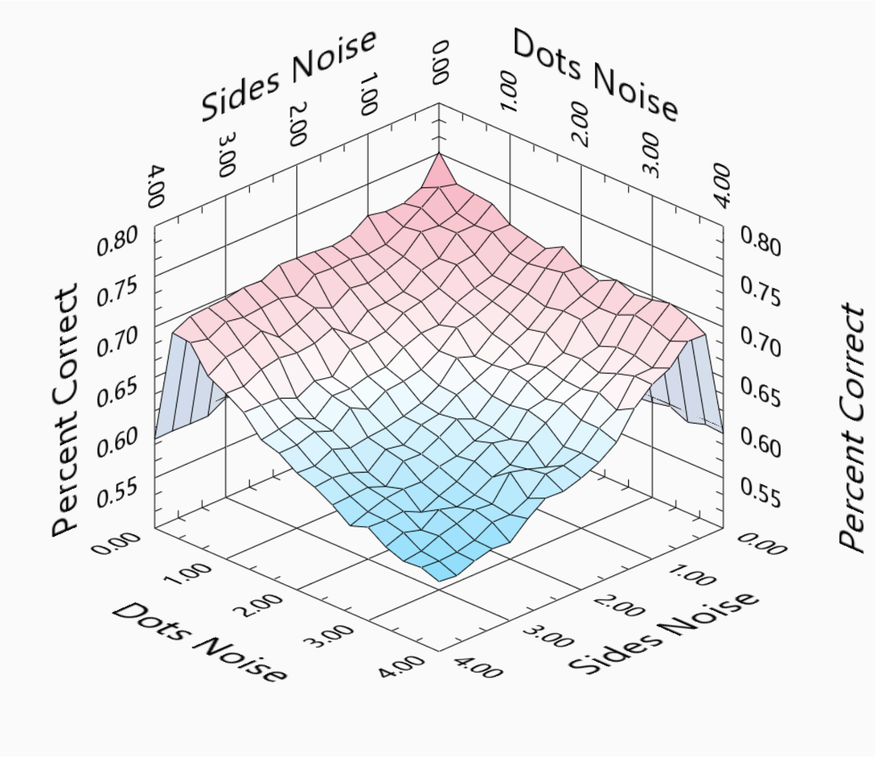


FIGURE 4.3: Results for experiment simulation of observer Decoder model 2. We generated 17 unique simulants per feature dimension for 289 total simulants with noise sigmas ranging from $\sigma = 0$ to $\sigma = 4$. Each simulant was run through 10,000 trials of a virtual categorization task identical to Experiment 1. Plotted is performance percent correct against the dot spacing feature noise and sides feature noise. The lowest percent correct was found with the simulant with $\sigma_{nm} = 0$ and $\sigma_{dm} = 0$ at $\approx$ 54%, and the highest percent correct was found with the simulant with $\sigma_{nm} = 0.25$ and $\sigma_{dm} = 0.25$ at $\approx$ 77%. In the mismatch cases where one feature noise was high and the other was low, performance was recorded at $\approx$ 60%.

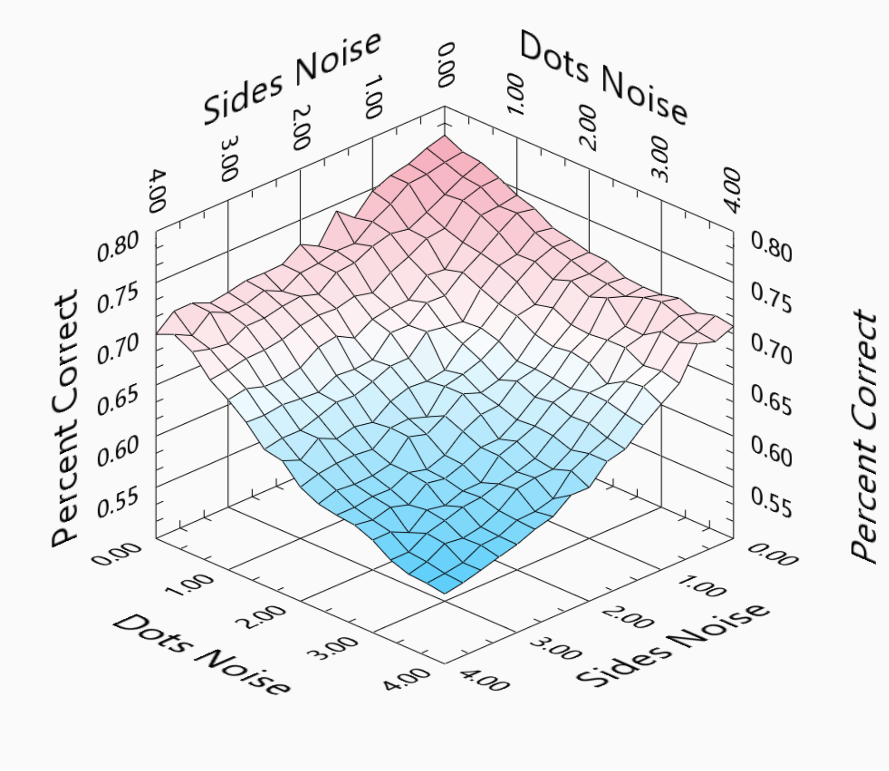


Figure 4.4: Results for experiment simulation of observer Decoder model 3. We generated 17 unique simulants per feature dimension for 289 total simulants with noise sigmas ranging from $\sigma = 0$ to $\sigma = 4$. Each simulant was run through 10,000 trials of a virtual categorization task identical to Experiment 1. Plotted is performance percent correct against the dot spacing feature noise and sides feature noise. The lowest percent correct was found with the simulant with the highest measurement noise at ≈ 56%, and the highest percent correct was found with the simulant with the lowest measurement noise at ≈ 77%. In the mismatch cases where one feature noise was high and the other was low, performance was somewhat impacted with the percent correct recorded at ≈ 70%.

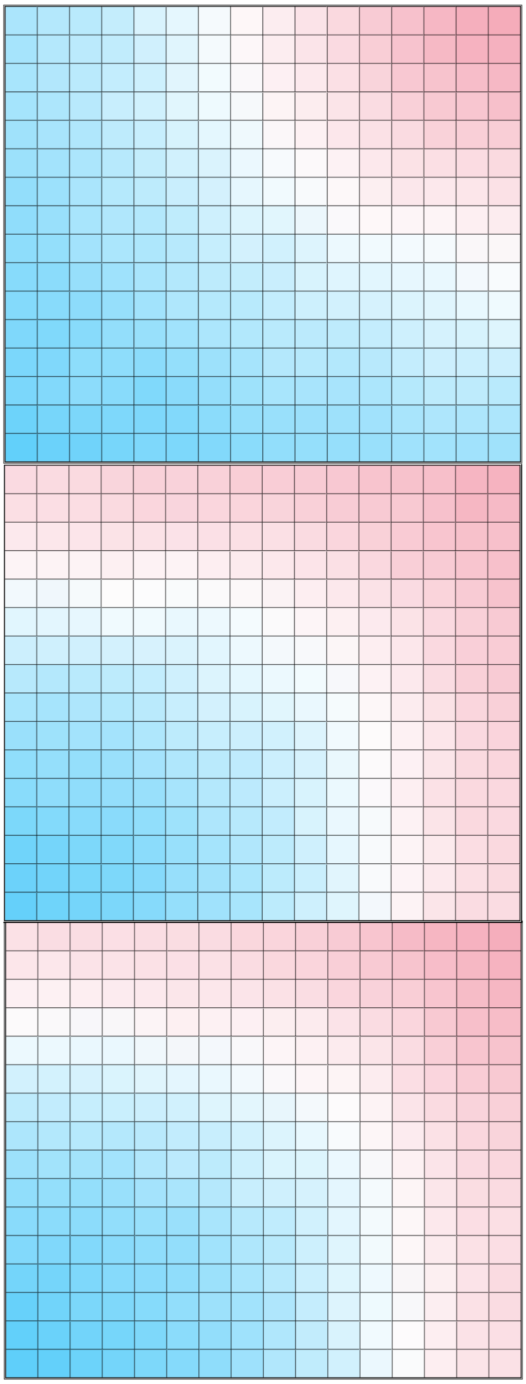


FIGURE 4.5: Intensity plots of the results for the experiment simulation of observers using Decoder models 1, 2, and 3 respectively. Pink represents the highest percent correct recorded, and blue represents the lowest percent correct recorded.

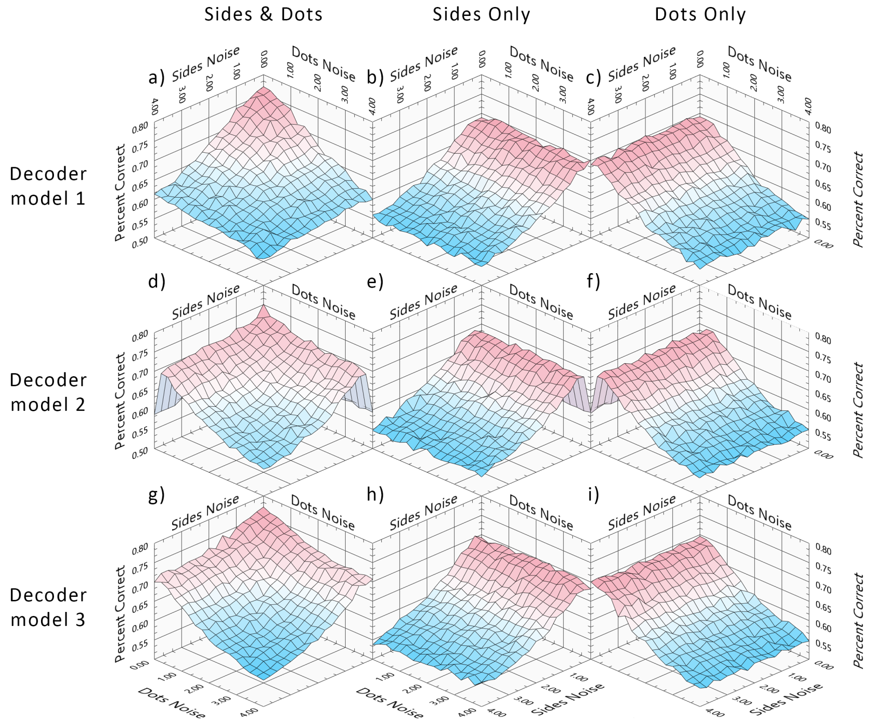
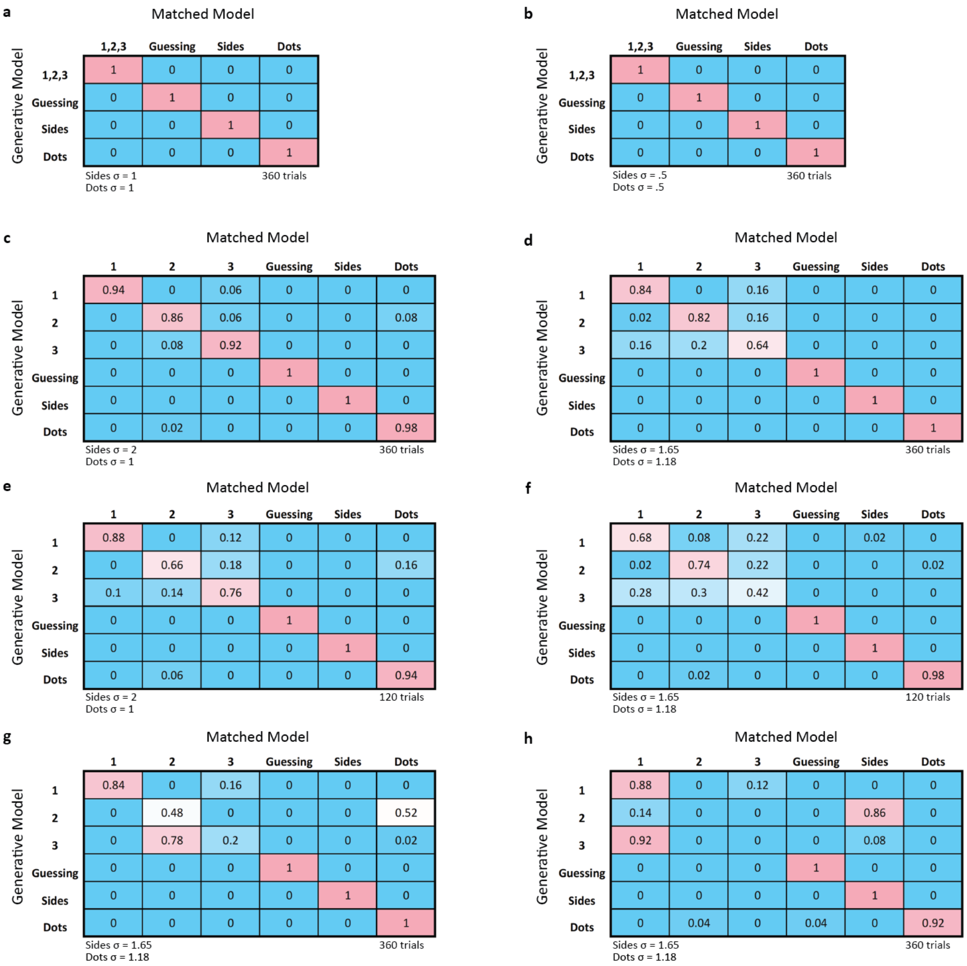


FIGURE 4.6: Results of the experiment simulation of 9 Decoder submodels. The three rows show simulation results for Decoder models 1, 2, and 3 respectively. Column 1 shows simulation results for the three models where both the Sides and Dots features were attended to. Column 2 shows simulation results for the "Sides Only" Decoder models, where the Dot's feature was ignored. Column 3 shows simulation results for the "Dots Only" Decoder models, where the Side's feature was ignored.



a

Matched Model

Generative Model	1,2,3	Guessing	Sides	Dots
1,2,3	1	0	0	0
Guessing	0	1	0	0
Sides	0	0	1	0
Dots	0	0	0	1

Sides σ = 1
Dots σ = 1
360 trials

b

Matched Model

Generative Model	1,2,3	Guessing	Sides	Dots
1,2,3	1	0	0	0
Guessing	0	1	0	0
Sides	0	0	1	0
Dots	0	0	0	1

Sides σ = .5
Dots σ = .5
360 trials

c

Matched Model

Generative Model	1	2	3	Guessing	Sides	Dots
1	0.94	0	0.06	0	0	0
2	0	0.86	0.06	0	0	0.08
3	0	0.08	0.92	0	0	0
Guessing	0	0	0	1	0	0
Sides	0	0	0	0	1	0
Dots	0	0.02	0	0	0	0.98

Sides σ = 2
Dots σ = 1
360 trials

d

Matched Model

Generative Model	1	2	3	Guessing	Sides	Dots
1	0.84	0	0.16	0	0	0
2	0.02	0.82	0.16	0	0	0
3	0.16	0.2	0.64	0	0	0
Guessing	0	0	0	1	0	0
Sides	0	0	0	0	1	0
Dots	0	0	0	0	0	1

Sides σ = 1.65
Dots σ = 1.18
360 trials

e

Matched Model

Generative Model	1	2	3	Guessing	Sides	Dots
1	0.88	0	0.12	0	0	0
2	0	0.66	0.18	0	0	0.16
3	0.1	0.14	0.76	0	0	0
Guessing	0	0	0	1	0	0
Sides	0	0	0	0	1	0
Dots	0	0.06	0	0	0	0.94

Sides σ = 2
Dots σ = 1
120 trials

f

Matched Model

Generative Model	1	2	3	Guessing	Sides	Dots
1	0.68	0.08	0.22	0	0.02	0
2	0.02	0.74	0.22	0	0	0.02
3	0.28	0.3	0.42	0	0	0
Guessing	0	0	0	1	0	0
Sides	0	0	0	0	1	0
Dots	0	0.02	0	0	0	0.98

Sides σ = 1.65
Dots σ = 1.18
120 trials

g

Matched Model

Generative Model	1	2	3	Guessing	Sides	Dots
1	0.84	0	0.16	0	0	0
2	0	0.48	0	0	0	0.52
3	0	0.78	0.2	0	0	0.02
Guessing	0	0	0	1	0	0
Sides	0	0	0	0	1	0
Dots	0	0	0	0	0	1

Sides σ = 1.65
Dots σ = 1.18
360 trials

h

Matched Model

Generative Model	1	2	3	Guessing	Sides	Dots
1	0.88	0	0.12	0	0	0
2	0.14	0	0	0	0.86	0
3	0.92	0	0	0	0.08	0
Guessing	0	0	0	1	0	0
Sides	0	0	0	0	1	0
Dots	0	0.04	0	0.04	0	0.92

Sides σ = 1.65
Dots σ = 1.18
360 trials

FIGURE 4.7: Confusion matrices detailing the accuracy of our model comparison analysis inferring the correct generative model given simulated behavioural categorization data. Simulations were run for 50 sets of 360 trials. Bayesian model comparison was used to determine the best-matched model per set of trials. Rows show the generative model used to generate behavioural data, and the columns show the frequency that our analysis matched that data to each model. Listed below each trial are the *Measurement Noise* sigmas used for the generative models and matched models. a-d) displays four different combinations of feature *Measurement Noise* and analysis after 360 trials. e-f) display two different combinations of feature *Measurement Noise* and analysis after 120 trials. g-h) displays two incongruent cases where the generative model's *Measurement Noise* was different than what is displayed and used for the analysis. g) shows the case where the generative model sigmas were $\sigma_n = 2$ and $\sigma_d = 1$. h) shows the case where the generative model sigmas were $\sigma_n = 1$ and $\sigma_d = 2$.

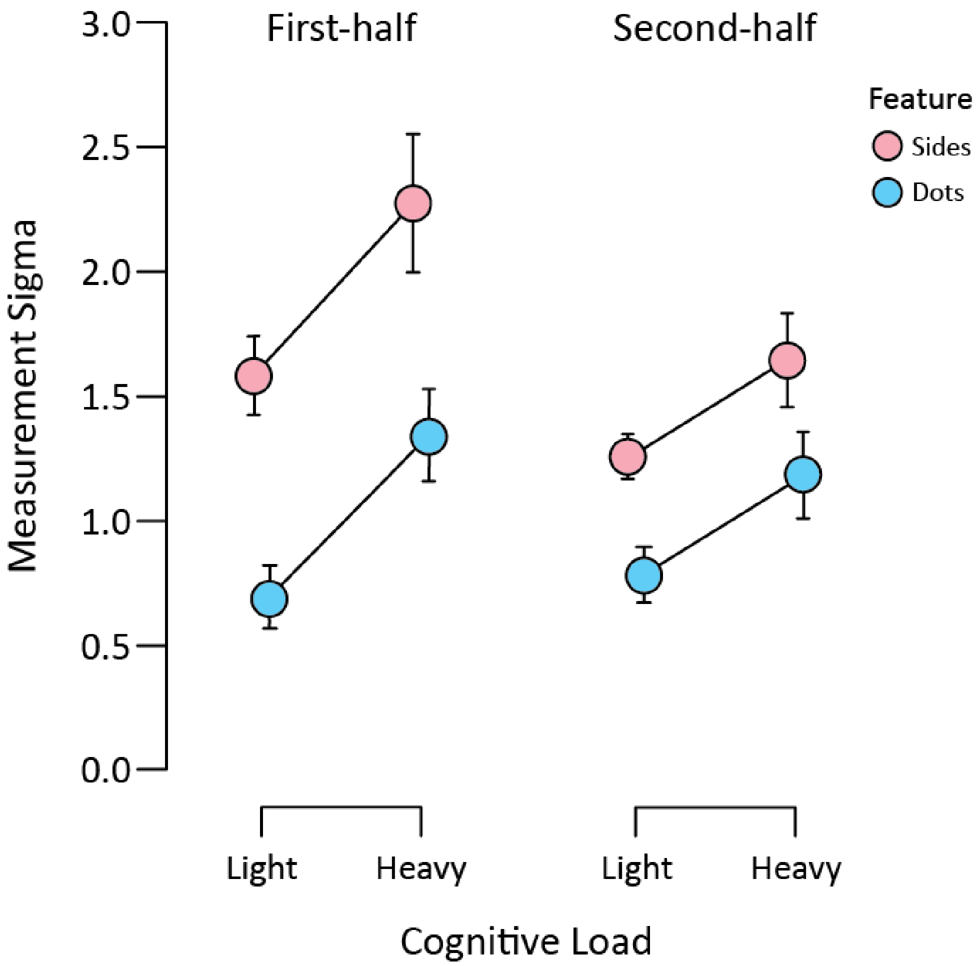


FIGURE 4.8: Measurement sigmas across 16 participants in a 2IFC experiment. Sigmas were calculated for both the Sides and Dots features in the *Low* and *High* task load conditions. Error bars represent Standard Error. *Low Task Load* referred to when participants only attended to either of the two features in a trial, and *High Task Load* referred to when participants attended to both features in a trial. The left side of the graph displays the measurement sigmas calculated in the *First-half* blocks of the experiment, and the right side displays the measurement sigmas calculated in the *Second-half* blocks of the experiment. Generally, measurement sigmas across participants were lower for the dots feature than the sides feature. Increasing the task load also increased measurement sigmas for both features. A repeated measures ANOVA revealed a significant effect of feature, experiment half, and task load, with an interaction between feature and experiment half.

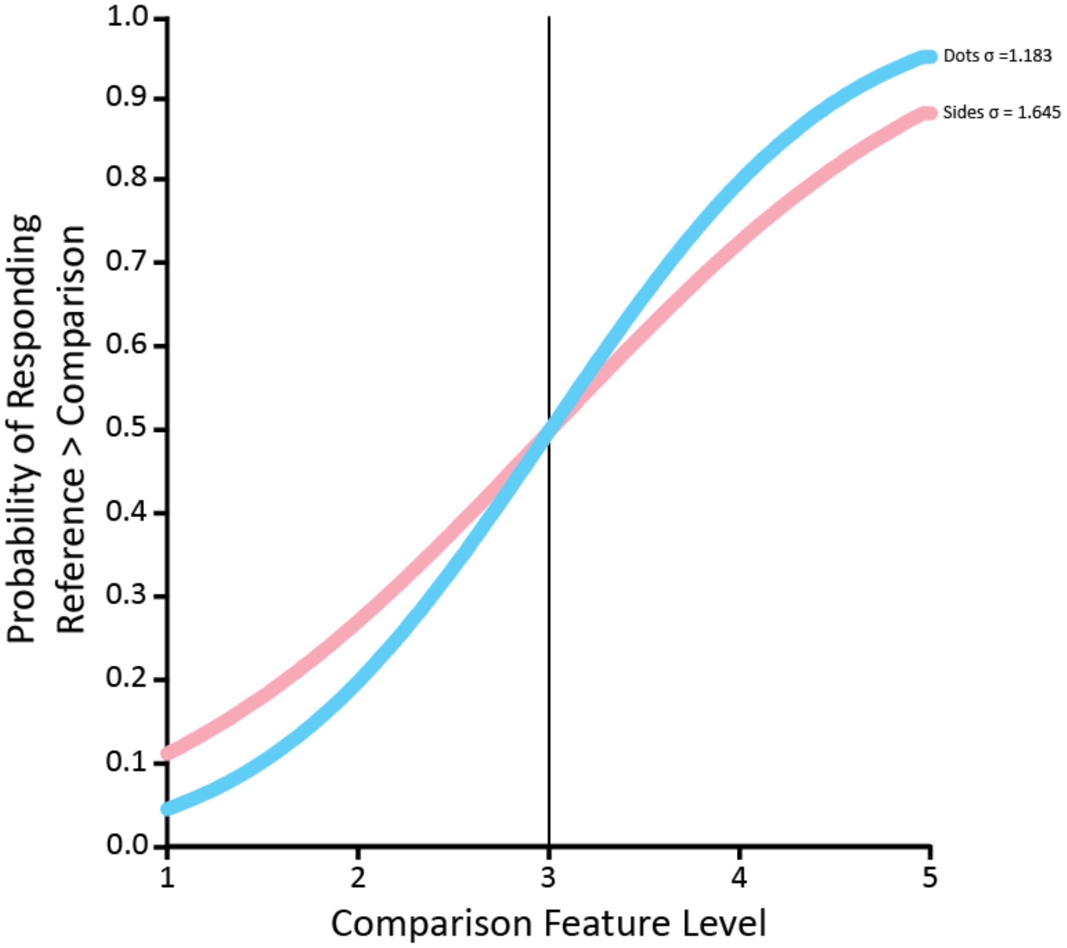


Figure 4.9: Best-fitting psychometric functions for the Sides measurement and the Dots measurement as found in Experiment 3. The x-axis corresponds to the five feature levels of the comparison stimulus compared against the reference stimulus, with level 3 being where the reference stimulus and comparison stimulus were equal. The y-axis plots the probability of a participant responding that the reference stimulus was larger than the comparison stimulus for either the Sides feature with $\sigma_n = 1.645$ or the Dots feature with $\sigma_d = 1.183$. These sigmas represent what was found in the *High Task Load* condition in the *Second-Half* blocks of the experiment.

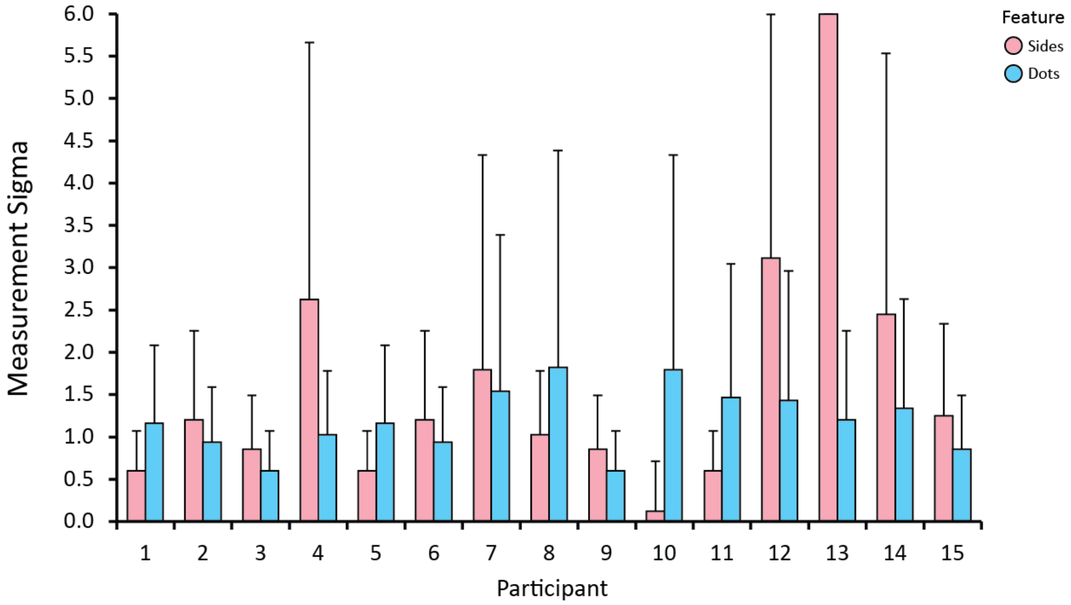
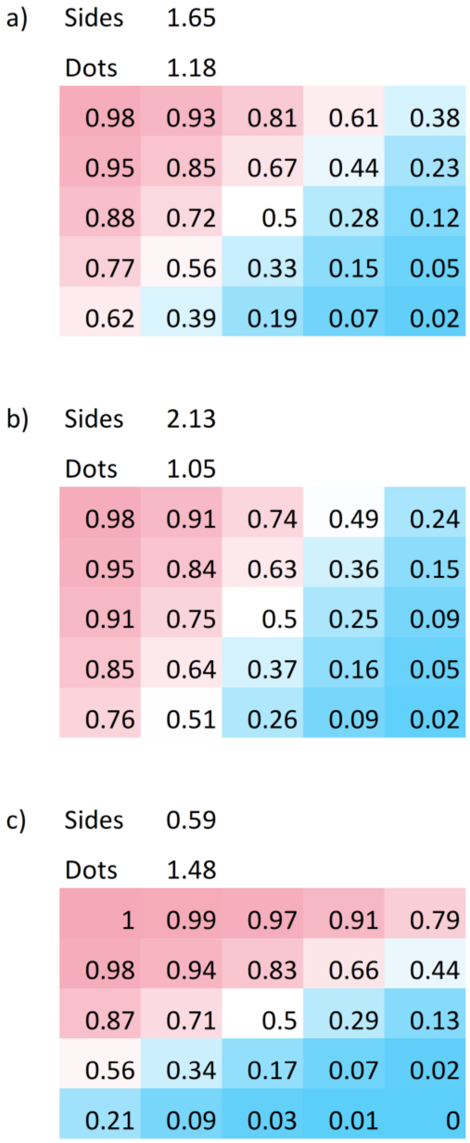


Figure 4.10: Average best-fitting measurement sigmas for the Sides and Dots as found in Experiment 3 across all 15 participants. These sigmas represent what was found in the *High task load* condition in the *Second-half* blocks of the experiment. Error bars represent 95% confidence intervals for each measurement.



a) Sides 1.65

Dots 1.18

0.98	0.93	0.81	0.61	0.38
0.95	0.85	0.67	0.44	0.23
0.88	0.72	0.5	0.28	0.12
0.77	0.56	0.33	0.15	0.05
0.62	0.39	0.19	0.07	0.02

b) Sides 2.13

Dots 1.05

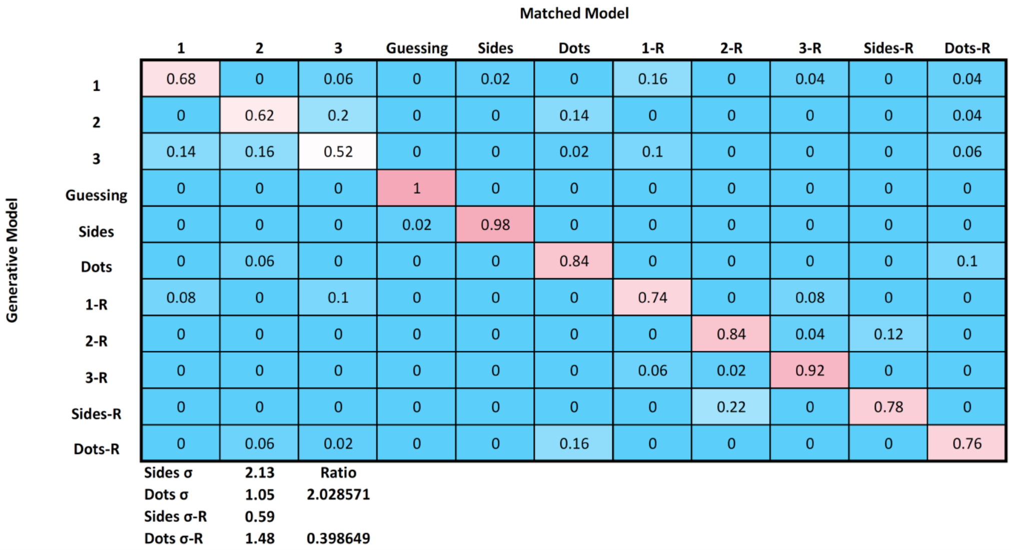
0.98	0.91	0.74	0.49	0.24
0.95	0.84	0.63	0.36	0.15
0.91	0.75	0.5	0.25	0.09
0.85	0.64	0.37	0.16	0.05
0.76	0.51	0.26	0.09	0.02

c) Sides 0.59

Dots 1.48

1	0.99	0.97	0.91	0.79
0.98	0.94	0.83	0.66	0.44
0.87	0.71	0.5	0.29	0.13
0.56	0.34	0.17	0.07	0.02
0.21	0.09	0.03	0.01	0

Figure 4.11: Top-down Intensity plots displaying the generated response tables $P(\text{"}A\text{"}|\mathbf{f}_g)_g$ for Decoder model 3 with three different configurations of *Measurement Noise* sigmas. a) shows a table generated for an observer with a Sides sigma of $\sigma_{nm} = 1.645$ and a Dots sigma of $\sigma_{dm} = 1.183$. b) shows a table generated for an observer with $\sigma_{nm} = 2.13$ and $\sigma_{dm} = 1.05$. c) shows a table generated for an observer with $\sigma_{nm} = 0.59$ and $\sigma_{dm} = 1.48$.



Matched Model

Generative Model	1	2	3	Guessing	Sides	Dots	1-R	2-R	3-R	Sides-R	Dots-R
1	0.68	0	0.06	0	0.02	0	0.16	0	0.04	0	0.04
2	0	0.62	0.2	0	0	0.14	0	0	0	0	0.04
3	0.14	0.16	0.52	0	0	0.02	0.1	0	0	0	0.06
Guessing	0	0	0	1	0	0	0	0	0	0	0
Sides	0	0	0	0.02	0.98	0	0	0	0	0	0
Dots	0	0.06	0	0	0	0.84	0	0	0	0	0.1
1-R	0.08	0	0.1	0	0	0	0.74	0	0.08	0	0
2-R	0	0	0	0	0	0	0	0.84	0.04	0.12	0
3-R	0	0	0	0	0	0	0.06	0.02	0.92	0	0
Sides-R	0	0	0	0	0	0	0	0.22	0	0.78	0
Dots-R	0	0.06	0.02	0	0	0.16	0	0	0	0	0.76

Sides σ	2.13	Ratio
Dots σ	1.05	2.028571
Sides σ-R	0.59	
Dots σ-R	1.48	0.398649

FIGURE 4.12: Confusion Matrix of our expanded model comparison analysis examining the accuracy in inferring the correct generative model given simulated categorization data. Simulations were run for 50 sets of 360 trials. The rows show the generative model used to generate a set of behavioural data, and the columns show the frequency with our analysis matched this generated behavioural data to each model.

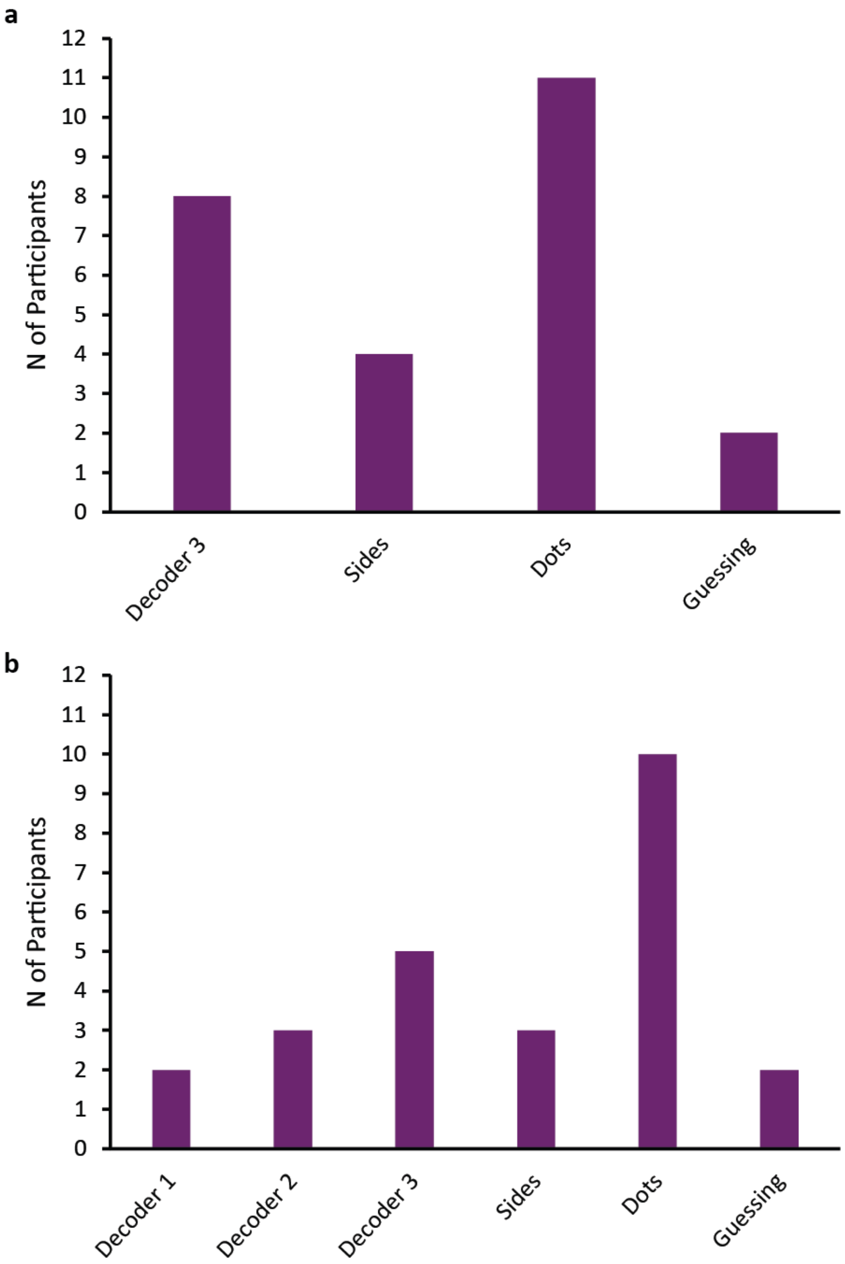


FIGURE 4.13: Histogram of the number of participants matched to a range of Decoder models. The Decoder Models were run with a *Measurement Noise* configuration of $\sigma_{nm} = 1.645$ and $\sigma_{dm} = 1.183$. a) shows the resulting histogram when only $D3$ was run alongside Dn, Dd, and Dg. b) shows the resulting histogram when $D1$, $D2$, and $D3$ were run alongside Dn, Dd, and Dd.

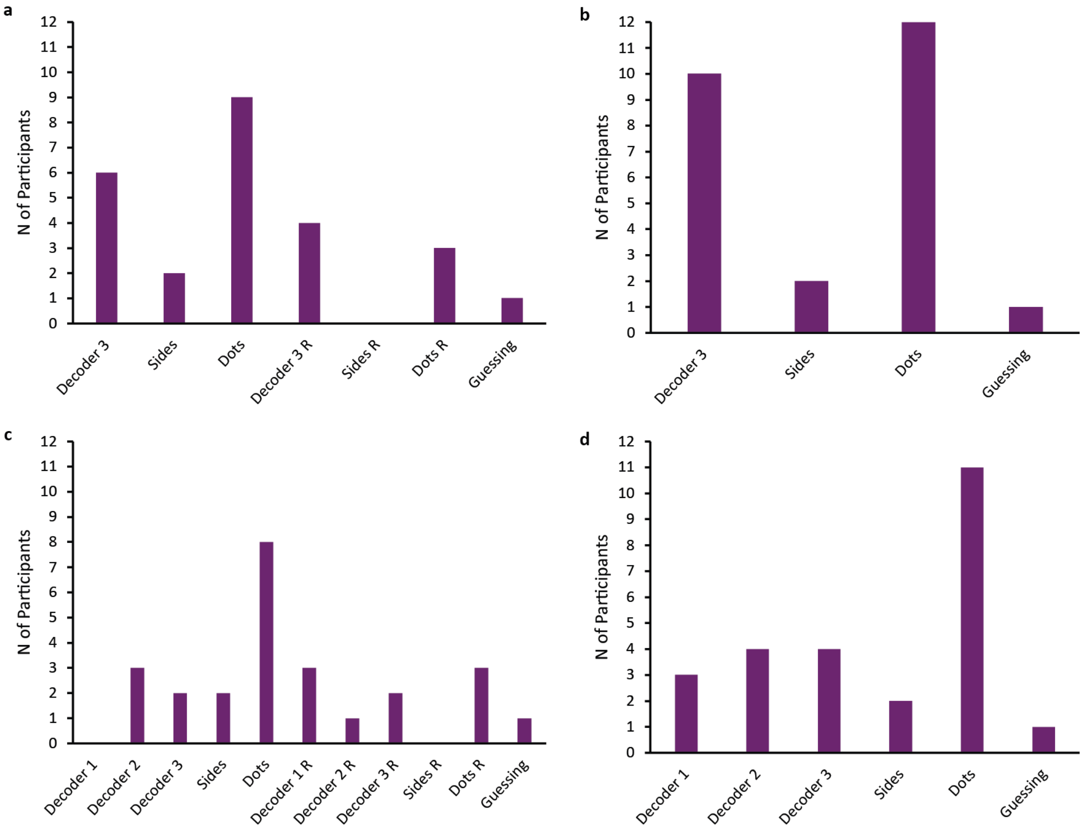
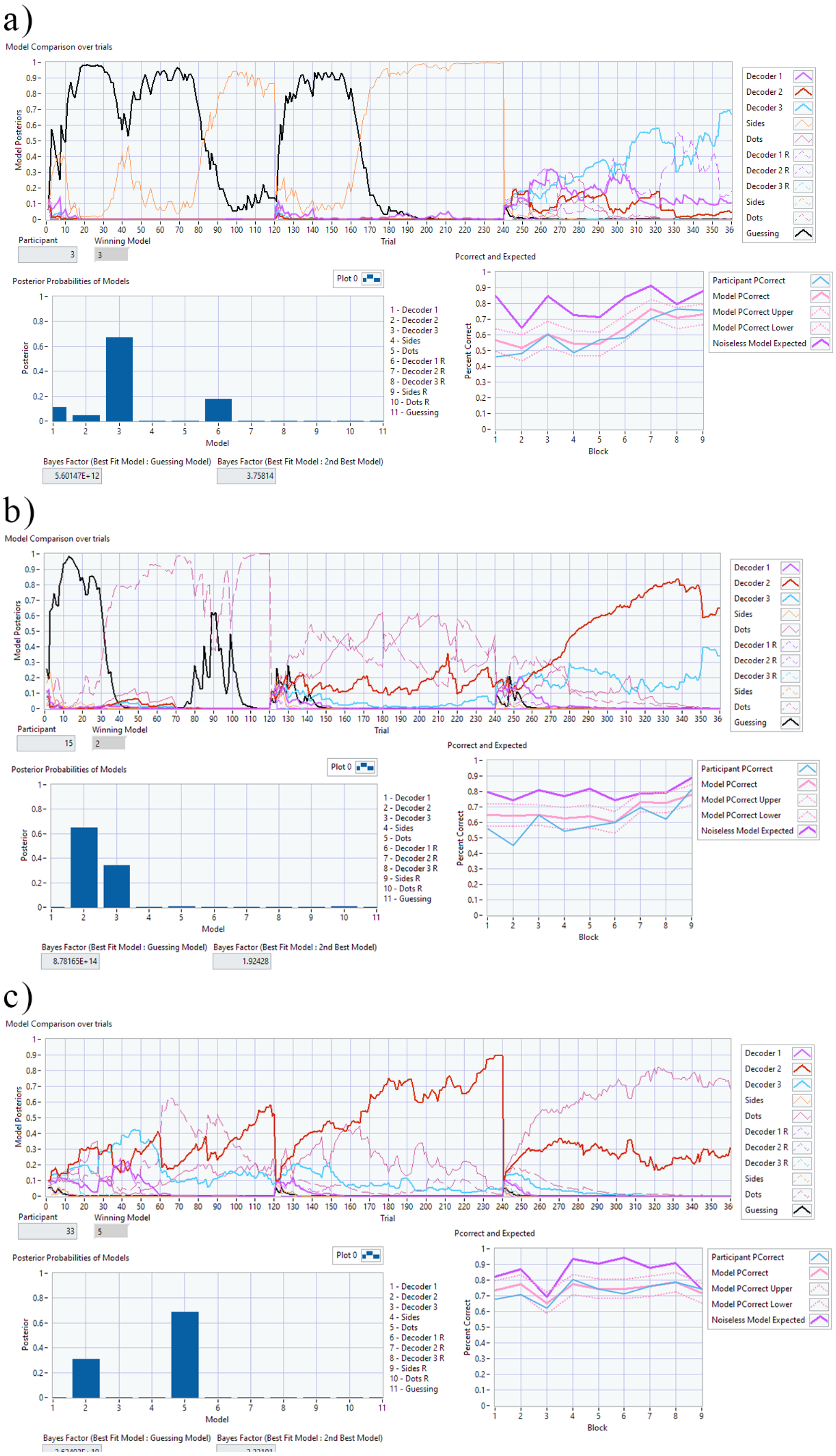


FIGURE 4.14: Histogram of the number of participants matched to an expanded range of Decoder models. The Decoder models were run with a *Measurement Noise* configuration of $\sigma_{nm} = 2.13$ and $\sigma_{dm} = 1.05$. The three *reverse* Decoder models (designated with an R) were run with a *Measurement Noise* configuration of $\sigma_{nm} = 0.59$ and $\sigma_{dm} = 1.48$. a) shows the resulting histogram when only $D3$ was run alongside Dn, Dd, the *reverse* models, and Dg. c) shows the resulting histogram when $D1$, $D2$, and $D3$ were run alongside Dn, Dd, the *reverse* models, and Dg. b) and d) show the results of combining the *reverse* model results with their respective counterparts.



a)

Participant 3

Winning Model 3

Bayes Factor (Best Fit Model : Guessing Model) 5.60147E+12

Bayes Factor (Best Fit Model : 2nd Best Model) 3.75814

b)

Participant 15

Winning Model 2

Bayes Factor (Best Fit Model : Guessing Model) 8.78165E+14

Bayes Factor (Best Fit Model : 2nd Best Model) 1.92428

c)

Participant 33

Winning Model 5

Bayes Factor (Best Fit Model : Guessing Model) 2.62402E+10

Bayes Factor (Best Fit Model : 2nd Best Model) 2.22101

FIGURE 4.15: Results of our expanded model comparison analysis for participants 3, 15, and 33 are shown in A), B), and C) respectively. The "model comparison over trials" plot displays the calculated posterior probabilities for each model in the model comparison analysis across trials. The analysis was reset for every 120 trials (3 blocks). The "winning model" was determined by the model with the highest posterior probability in the last trial of the experiment. A bar chart displaying these final posterior probabilities across all models is also displayed. Two Bayes Factors were calculated, one comparing the posterior of the best-fitting model to the posterior of the Guessing model Dg and one comparing the best-fitting model to the second-best-fitting model. Finally, the participants' raw performance was plotted across blocks (shown in blue) against the average performance predicted by the participant-matched model (shown in pink). The average predicted performance was calculated by running the matched model through 10,000 simulations of the participant's experiment block and calculating the model's performance per simulation. The performance of the noiseless Bayesian Categorization Model was also plotted (shown in purple).

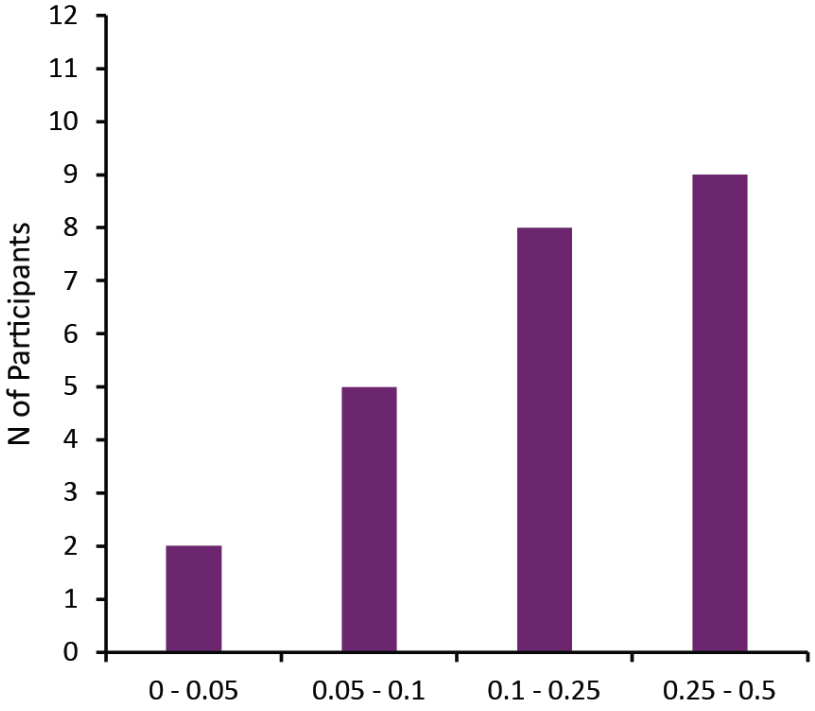


FIGURE 4.16: Histogram of the calculated matched model p-values across our participants in Block 9. This value represents a fitness measure between each participant's percent correct and the expected range of performance produced by their matched model through simulation. The four bins of our histogram are: 0 – 0.05, signifying a poor fit; 0.05 – 0.1, signifying a moderate fit; 0.1 – 0.25, signifying a good fit; 0.25 – 0.5, signifying a very good fit.

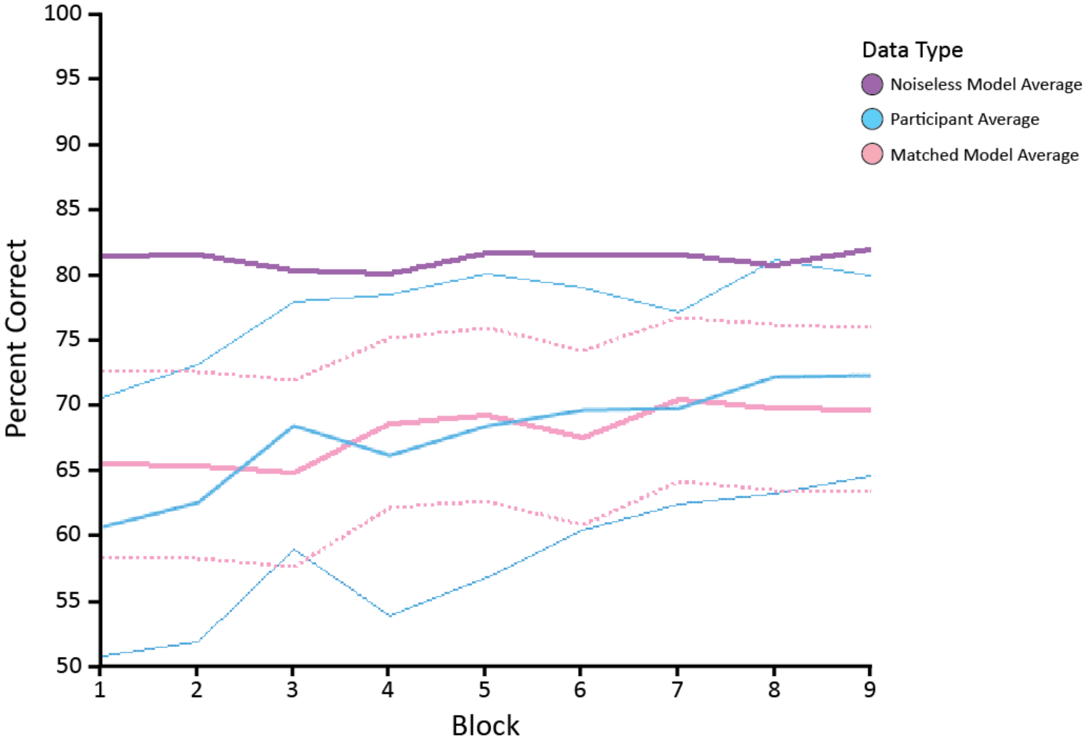


FIGURE 4.17: Total average results of our model comparison analysis comparing average performance as given by our *noiseless* Bayesian Observer Model (in purple), average participant performance per block (in blue), and the average performance of each participant's matched model (in pink). The thinner blue and pink lines represent 95% confidence intervals for average participant performance and average matched model performance respectively.

Chapter 5

General Discussion

5.1 Summary

Throughout this thesis we have presented the exploration of a Bayesian Categorization Model applied to a human categorization paradigm within the sense of touch. We designed a category task using novel 3D printed objects and trained participants on two artificially defined categories, named *Elyk* and *Noek*. Through both empirical and theoretical work, we demonstrated that human participants can learn categories using only touch alone.

In Chapter 2, we detailed the derivation of this Bayesian Categorization Model as well as the experimental design used throughout the rest of the thesis. We explored further complex generalizations of the model which could be extended to many other experiments outside the scope of this thesis. We also outlined a modification of our model to include sensory *Measurement Noise* explored more thoroughly in a later chapter.

In Chapter 3, we investigated human performance on the categorization task compared to our Bayesian Categorization Model. We found that participants were able to reach high performance levels after a single day of experimentation, with only a minor performance increase observed when the experiment extended across multiple days. Furthermore, we demonstrated that our Bayesian Categorization Model was able to be matched to human performance with some degree of accuracy. However, our model still outperformed participants and was not completely predictive of their performance.

In Chapter 4, we investigated the impact of internal sensory *Measurement Noise* on observers in our categorization task through both theoretical simulation and an empirical Two-Interval Forced-Choice (2IFC) experiment. Through simulations of our experiment using a Bayesian Categorization Model which included *Measurement Noise*, we demonstrated that the inclusion of sensory noise had the potential to strongly impact categorization performance. Our 2IFC experiment revealed that *Measurement Noise* was a present factor amongst participants observing our physical stimulus set. Using empirically driven estimates for this *Measurement Noise*, we then re-analyzed the results of Experiments 1 and 2 and found that a more complete *noisy* Bayesian Categorization Model fit our collected participant data even more closely.

Throughout this thesis, we have found evidence supporting the importance of including sensory *Measurement Noise* in human categorization models. Our simulations demonstrated that the amount of *Measurement Noise* found for participants' feature measurements was significant enough to expect a decrease in performance. Incorporating this noise into our Bayesian Categorization Model led to a better-fitting model that accurately matched the performance of our participants in a categorization task.

To our knowledge, the current thesis provides a unique exploration of a categorization task in which participants are limited to only their sense of touch and a Bayesian model like ours applied this type of categorization paradigm. Our main findings are consistent with previous categorization studies which have investigated the other senses or combinations of senses.

5.2 Learning categories using touch

The study of categorization has seen the use of a wide variety of stimuli. Many experiments focus on visual representations, from real-world images or drawings of imaginary animals (Iordan et al., 2015; Brooks & Hannah, 2006; Minda & Ross, 2004; Ahn & Medin, 1992; Yamauchi & Markman, 2000) to more abstract displays including arrays of dots, lines, geometric figures and sine-wave discs (Bankieris et al., 2017; Ashby & Gott, 1988; Smith et al., 2010; Love, 2002). Experiments using stimuli which focus on other senses are not as common, but they can sometimes involve auditory stimuli such as whole novel words or phonemes (Bankieris et al., 2017; Bejjanki et al., 2011). In some experiments, categories are even formed through a combination of audio and visual stimuli. Tactile stimuli have seldom seen exploration in categorization.

One study from Schwarzer et al. (1999) looked at the haptic categorization of stimuli consisting of wooden blocks made from beech wood. These blocks varied by four features: size, shape, surface texture, and weight. Each feature had three levels; for instance, the surface texture feature had fine, medium, and rough instantiations, and the shape feature changed from a cylinder to a mixture of a cylinder and cube, and a cube. Unlike our experimental paradigm, not every possible combination of their stimuli was created. Instead, a select 18 instantiations were chosen to be created. Generally, Category A consisted of objects whose features were mostly *level 1*, and Category B consisted of objects whose features were mostly *level 3*. Their categorization experiment involved a learning phase and a test phase. During the learning phase, participants explored a subset of the objects (8 in total) using only their sense of touch and then attempted to categorize the object, receiving corrective feedback directly from the experimenter. During the testing phase, participants were presented with the full set of 18 stimuli and categorized without receiving corrective feedback. Schwarzer et al. (1999) found that their participants predominantly learned the categories through an analytical learning strategy—one in which participants categorized based on single separated features—rather than a holistic one. Furthermore, participants exhibited an overall preference for the surface texture feature of the objects. The shape feature was preferred second, although the preference difference between surface texture and shape was not as large amongst their adult population as their younger population.

The results found by Schwarzer et al. (1999) seem at odds with what we found throughout our own experiments and model comparison analysis. While we did find that a large proportion of our participants focused on our Dots surface feature, we also found evidence that many participants were using both features to categorize our stimuli. With the inclusion of *Measurement Noise*, the model which combined both the Sides and Dots features was fit to almost as many participants as the model which only looked at the Dots feature. However, the structure of our categorization experiment differs in key ways that make direct comparisons here difficult. Mainly, the formation and consideration of our categories are fundamentally different.

Our categories were formed by underlying Gaussian distributions. To learn these categories, participants were required to learn the underlying parameters. We created our stimuli with these underlying parameters in mind to represent this underlying statistical structure present in our categories. In contrast, the stimuli and categories that Schwarzer et al. (1999) used did not consider underlying statistical distributions along their four features. Instead, features were ternary with arbitrarily distanced levels.

The training of our categories was also done in such a way as to reinforce underlying categorical statistical distributions. This meant that we presented stimuli from our categories by drawing from these distributions. In contrast, Schwarzer et al. (1999) trained participants on a small subset of chosen exemplar stimuli and then tested on a larger set. Each of the stimuli was shown one at a time and categorized by the participant outside of any reinforcement of the categories. Furthermore, the stimuli in their test phase included exemplars outside of the expected range from the learning phase stimuli. The learning phase involved stimuli with a feature being one level away from the Category, while the testing phase involved stimuli with a feature being two levels away from the Category. The closest analogue of this with our experiment would be if we had first trained participants on sharper category distributions and then changed the categories by broadening the distributions for a test phase. However, this would still not make an equivalent experiment as there are too many differences in how our categories are structured. While the study by Schwarzer et al. (1999) is an informative and related work to ours being one of the few experiments we could find involving haptic exploration of physical objects and category learning, its consideration of categories fundamentally differs from ours which makes a direct comparison difficult.

5.3 Opposing paradigms of categorization

The approach taken by Schwarzer et al. (1999) is similar to research like that of Dr. Lee Brooks in which categories are considered a collection of features with often binary integer valued levels (Allen & Brooks, 1991). Stimuli can often involve line drawings of imaginary animals, such as the Bleebs and Ramuses (Brooks & Hannah, 2006), or the Moneks and Plaples (Yamauchi & Markman, 2000), Brunswik faces (Reed, 1972), or sometimes just visual displays of geometric shapes (Medin & Schaffer, 1978). For example, Bleebs and Ramuses were defined by four binary features: 2 or 4 legs, round or angular head, round or angular body, and stripes or spots (Brooks & Hannah, 2006). Categories from these stimuli are typically based on "prototype" representations and variations are made by mixing one or more features from the opposing category. A Bleeb prototype might be coded as 1111 but a 1011 or 1101 would also be Bleebs. This simplification of categories could code for more than two features (Schwarzer et al., 1999) or even more than two categories (Sanborn et al., 2010).

These approaches fundamentally differ from ours and that of other research which considers the features within a category as continuous statistical distributions. Often studies of this type use stimuli that can be parameterized, often as Gaussian distributions, and the stimuli are created by drawing from these distributions. Methodologies tend to be based in *Signal Detection Theory* (*SDT*) (Ashby & Gott, 1988; Smith et al., 2010) or follow similarly from a *cue-combination* approach (Bankieris et al., 2017).

We refer to the first type of experiments as the *Nominal-Ordinal Feature Paradigm*, and the second type of experiments as the *Continuous Feature Paradigm*. The treatment of category features as an ordinal or nominal set, or a continuous measurement, is the major differentiating factor that makes comparisons between the two experimental paradigms difficult. Both ultimately study the same underlying phenomenon, but attempts to understand one set of category formation from the other reveal that the stimuli and underlying assumptions of how they are perceived are too different. This fundamental difference tends to inform the rest of the experimental design in incompatible ways.

Both paradigms can deal with overlap among the individual feature distributions of categories. A *Nominal-Ordinal Feature Paradigm* experiment might have so-called "one away" stimuli with a Category A object that has a feature found normally in Category B. For instance, if we were to follow the *Nominal-Ordinal Feature Paradigm* for our stimulus set, we would have objects with two features of binary levels: fewer sides versus more sides and fewer dots versus more dots. Our categories could be defined as 11 or 00. Thus, there would be only four objects in our set. We could introduce variation similarly to Brooks & Hannah (2006) as different instantiations of these feature levels (different organizations of dots, or more or fewer dots), but they would be considered as different variations of the same feature level. There would no longer be consideration for the distribution of dots or sides seen in a category. The treatment of the feature as either ordinal or nominal also means that the perceptual distance between those levels is not often considered, or the frequency that these overlapping features are seen across the categories. *Continuous Feature Paradigm* experiments in contrast will tend to consider the statistical distributions underlying the features in each category. This extends to how the overlap of the categories overall is considered between the experimental paradigms.

The critical factor important for all categorization research is whether the methodology treats the category and feature distributions probabilistically. Considering categories as probabilistic distributions turns category learning into an exercise in learning the distributions underlying the category—the observer learns the variance they might expect in their measurements of the categories through repeated exposure and exploration. It also allows for models like our Bayesian Categorization Model to be applied which can lead to predictive power and understanding of how observers are learning the categories and inference as to when they are likely to be making mistakes. Experiments under the *Continuous Feature Paradigm* fit well with treating category and feature distributions probabilistically. However, experiments under the *Nominal-Ordinal Feature Paradigm* often do not include a consideration of these probabilistic distributions and thus are incompatible with our model without adjustments to the methodology. The probabilistic treatment of category and feature distributions is also what allowed us to incorporate the presence of *Measurement Noise* in our model. This would not be possible under the *Nominal-Ordinal Feature Paradigm*.

We cannot say with absolute certainty whether the *Measurement Noise* for a particular feature, such as our Sides feature, is the driving factor for why an observer may choose to ignore a feature or focus predominantly on another. To determine this more completely, further experimentation would be required to directly test the hypothesis that observers will choose to ignore a feature with high *Measurement Noise* more often than a feature with low noise. However, the results found throughout this thesis suggest that this may be the case. We believe the data and analysis presented in this thesis provide a compelling argument for the importance of including *Measurement Noise* in human categorization models.

5.4 The analytical power of our model

Throughout Chapter 4 we explored the development of a complex post-hoc analysis method which allowed us to make inferences about how observers might process the information available to them. This resolved many of the issues we encountered when originally analyzing Experiments 1 & 2. Mainly, it allowed us to make a direct analysis of the participants' trial-by-trial response behaviour when our collected data was not immediately suited for this. Through our model comparison analysis, we were able to compare different alternative models explaining how observers might categorize our stimuli and directly test which model best fits individual participants' data. This approach followed naturally from our Bayesian Categorization Model and the way we considered our categories and features under the *Continuous Feature Paradigm*.

Our analysis methodology developed here is a powerful approach for examining categorization by observers. By outlining the Encoder and Decoder models that an observer might use to resolve a categorization problem we can directly explore expected behavioural outcomes from an observer through simulations. Running many tens of thousands of experiment simulations becomes a trivial computational task even under the presence of stochastic *Measurement Noise* and helps better inform further experimentation and analysis. By applying Bayesian model comparison approaches we inferred if participant behaviour fit that of an observer focusing on just one feature (analogous to what's sometimes referred to as *analytical* strategies in other works) or fit an observer who combined both features. Fitting participants to their closest model made it further possible to better estimate an expected performance outcome given the trials they were presented with.

This analysis method can be greatly expanded upon in many ways. The specific models we chose to compare were based on our Bayesian Categorization Model and derivative sub-models. However, any alternative model could be used to fit human participant data. This method can also include parameter estimation to find the best-fitting parameters used in the model. For instance, while we relied on empirically driven average estimates for the sensory *Measurement Noise* of our Sides and Dots features, these values could have been estimated as sub-hypotheses in an attempt to find the best fitting sigma.

This could be further extended into incorporating learning of the categories themselves, something we did not directly explore through our model. Under the *Continuous Feature Paradigm* approach, categories can be encoded as a set of distribution mean's and sigma's. Learning of a category thus would mean learning of these parameters. This learning could be modelled as a complex parameter estimation problem, where different configurations of parameters act as sub-hypotheses in the model comparison. The analysis would then attempt to estimate the best-fitting mean and sigma parameters that would give rise to the participant's behaviour in a given set of trials.

5.5 Possible neural mechanisms underlying Bayesian categorization

Throughout this thesis, we have mainly discussed categorization and our Bayesian model in terms of a theoretical computational model. Following Marr (1982)'s *Levels of*

Analysis, we have been concerned with exploring the *algorithmic* level of processing categorization (McClamrock, 1991). Our focus lay on exploring our computational model—directly tying it into a neural implementation that existed outside the scope of our experimentation. Here we will briefly overview the potential links between computational Bayesian models for categorization and their implementation in the brain.

It has been established that aspects of perception and processes in the brain can be understood as probabilistic inferences, and many studies have found success in applying Bayesian methods and computational models to this end. Many of these studies have shown that the brain can take into account uncertainty in its integration of information to make inferences about its perceptions (Moreno-Bote et al., 2011). Many visual illusions have been demonstrated to be the result of inferences which can be modelled and predicted through Bayesian approaches (Weiss et al., 2002; Knill & Pouget, 2004). Weiss et al. (2002) for instance applied a Bayesian visual motion perception model and found consistent predictive power accounting for many of the illusions human observers experience under various circumstances. Bayesian perception models have also been powerful in explaining tactile illusions, such as the Cutaneous Rabbit Illusion (Tong et al., 2016).

Research into the Bayesian nature of the brain has led to the formation of a *Bayesian Coding Hypothesis*, stating that the brain generally represents sensory information probabilistically (Knill & Pouget, 2004). This follows directly from the population coding hypothesis, which states that the nervous system encodes sensory information through populations of neurons. This has been seen in many neural recording studies. For instance, single-cell recordings of the Middle Temporal Visual Area (MT) visual area in rhesus monkeys revealed that clusters of neurons were responsible for encoding a direct orientation feature when the monkeys perceived visual movement (Saizman et al., 1990; Saizman & Newsome, 1994). Population coding has also been observed in other focus areas of rhesus monkey brains, such as in the neurons in the Lateral Intraparietal Cortex (LIP) (Churchland et al., 2008). Individual neurons within the neuron populations orient to encode specific features, like orientation, forming this tuning curve behaviour to collectively encode the feature (Pouget et al., 2000; Sigala & Logothetis, 2002). In this way, the brain repeatedly organizes itself for feature selection.

The population coding in neurons may follow a gain-encoding schema, where the population of neurons can encode the mean and variance of a Gaussian density function simultaneously. This takes advantage of the fact that neural population coding noise tends to be Poisson-like (Tolhurst et al., 1983). As Saizman & Newsome (1994) have demonstrated, visual cortex neurons coding for orientation follow a Gaussian-like tuning curve, where an individual neuron has a preferred orientation where it responds maximally and its response drops off as that orientation is moved away from. Presenting the entire population of neurons with a particular orientation will elicit a noisy response from all the neurons responding at different rates according to their tuning curve. We can treat noisy Poisson responses of this nature as the posterior distribution for a Gaussian distribution, with the peak encoding the mean of the distribution and the amplitude encoding the variance (Pouget et al., 2003; Knill & Pouget, 2004; Beck et al., 2008; Lakshminarasimhan et al., 2018).

This mechanism of feature encoding through neural population coding has been extended further to categorical representations (O'Connell et al., 2018). Studies have found evidence for an Evidence accumulation model present in the brain, which states that the brain accumulates evidence for multiple possibilities over time until it reaches an internally set threshold (Churchland et al., 2008). This has been extended further into drift-diffusion models, where a calculated decision variable moves towards a threshold or criterion as the nervous system receives sensory information (O'Connell et al., 2018; Rorie et al., 2010). These lower-level feature variables are further combined into higher-level feature encoding when observers are making categorical judgements, which can further be represented through population coding. (Vogels, 1999). Categorical population encoding has been observed in various regions of the brain, such as the left-Dorsolateral Prefrontal Cortex (left-DLPFC), Anterior Cingulate Cortex (ACC), and Temporal-Parietal Cortex (TPC) (Freedman et al., 2003; Heekeren et al., 2004; Koenig et al., 2005; Shadlen & Newsome, 2001). The previously discussed neural areas demonstrating population coding, such as the MT, may encode primitive features that later integrate into abstract categorical representations (Iordan et al., 2015).

5.6 Deterministic vs Probabilistic Sampling

A pervasive question in cognition is the nature of decision strategies under uncertainty. Consider a binary task where an observer must discriminate a stimulus as coming from one of two distributions. One possible way an observer may respond is with a *deterministic* strategy, where an observer should respond according to the comparison between their internal measurement and some criterion. A deterministic observer should respond the same way given the same stimulus measurement. This idea forms a cornerstone to *SDT*, where an observer's response to a stimulus is dependent on the distance of the signal and noise distributions and their set criterion (Macmillan & Creelman, 2005). Under this model, an observer will respond the same way every time they receive a signal greater than their criterion, and their response only changes as either the signal or noise distributions change or the criterion is moved.

An alternative possibility for how an observer may respond is through *probability sampling*, also known as *Bayesian sampling*, where an observer responds with the underlying probability of the stimulus against the distributions (Lee & Janke, 1964; Healy & Kubovy, 1981). For example, if a particular stimulus has a 60% probability of coming from distribution A, an observer employing probability sampling should respond A 60% of the time. There has been some evidence in favour of probability sampling under certain circumstances (Healy & Kubovy, 1981). For example, bi-stable percepts like the Neckercube Illusion are nicely explained by *Bayesian Sampling*. The two percepts of the Neckercube—an inward-facing cube versus an outward-facing cube—result in equal or near-equal posteriors, and an observer who uses *Bayesian Sampling* would experience switching between the two percepts randomly (Moreno-Bote et al., 2011). However, attempts to uncover the presence of *Bayesian Sampling* in perceptual decision-making have not seen much success and the model generally fails to explain human performance in perceptual tasks (Murray et al., 2015). More importantly, it is difficult to experimentally distinguish true probability sampling; data generated by a "noisy" deterministic strategy will match the appearance of probability sampling (Kubovy et al., 1971).

This noise can arise in a multitude of ways. For instance, an observer might shift their criterion during an experiment every trial, meaning even given the same stimulus and the same measurement being made an observer may still respond differently on a trial-by-trial basis (Kubovy et al., 1971; Healy & Kubovy, 1981). More importantly, the presence of internal sensory *Measurement Noise*, caused by the nervous system contaminating the sensation of the stimulus, can lead to different proximal measurements from the same distal stimulus. Even under a *maximum a posteriori* (MAP) estimator the presence of *Measurement Noise* makes the process stochastic by nature. The consequence of stochasticity would lead to empirically the behaviour as an observer using *Bayesian sampling*: variable responses given the same stimulus.

One thing to consider with the empirical challenge in differentiating deterministic from probabilistic decision strategies is that the difficulty arises from the limitation in our empirical observations. The deterministic observer may only *appear* stochastic from an external frame of reference. Their internal measurement is contaminated with their sensory system's internal *Measurement Noise*, a measurement that we do not have direct access to. If we were able to directly observe this internal measurement, we might find that the observer is responding according to a true deterministic MAP estimator.

5.7 Ecological validity of categorization experiments

A common criticism across much of categorization research is the validity of categorization research in real-world settings. Many categorization experiments focus on contrived situations of observers learning a specific, often unnatural set of stimuli in a laboratory setting. Visual categorization experiments may feature stimuli such as the cartoons of fictional animals (Brooks et al., 2007), arrays of dots on the screen (Ariely, 2001), sine-wave disks with varying orientations and bar widths (Smith et al., 2010), or even two lines at a 90-degree angle of varying lengths (Ashby & Gott, 1988). While undoubtedly these bodies of work have made important and informative strides in expanding our understanding of how humans (and non-humans (Smith et al., 2012)) process and form categories, it is worth questioning the ecological validity of these types of work.

Human observers form and use natural categories every day as they explore their environments. These categories are formed through intuitive and goal-driven exploration of encountered objects. Researchers like Markman & Ross (2003) have pointed out the gap between how categories are typically learned in experimental settings and how categories are learned in natural environments (Ross, 2000; Markman & Ross, 2003). This ultimately leads to paradigms such as *indirect learning tasks*, where the observers are not told there are categories nor are tasked with learning said categories (Minda & Ross, 2004).

One of the goals of the methodology we developed and used throughout this thesis was to provide a more naturally adjacent setting for observers to learn the categories of our stimuli. This was the main inspiration to create an experiment where participants are allowed free, unguided exploration of the stimuli. Rather than present the stimuli to the participant in a highly controlled manner, we simply handed the participant objects to explore on their own. We did not instruct them on how best to explore said objects or try to influence the way they might explore them. Instead, we allowed them to physically manipulate the objects however best they saw fit to extract the information they needed.

This does not address the full extent of the ecological validity concerns raised by Markman & Ross (2003). In our methodology, observers were still given the simple goal of classifying the objects without any further goal or context in mind. Future experiments could be designed to further bridge this validity gap by requiring the observers to naturally form categories with the objects, possibly in the form of a game. However, we hoped the considerations we made in allowing for a natural exploration of the objects, on top of the physical nature of our objects, might help observers to perceive the objects and categories holistically and as "real things" rather than abstract concepts.

5.8 Assumption of Equal Priors

Our assumption that participants apply an equal Prior on the two categories follows according to Laplace's *Principle of insufficient reason*: participants in the task have no reason to favour one category over the other, so initially they should assign the two categories equal priors (Feldman, 2012). We took steps to attempt to remove any possible bias in creating unique novel names for the categories. If we had simply named the categories "One and Two" or "A and B", there could potentially have been an ordering bias. Additionally, arbitrarily choosing different existing nouns as the category names like "Chocolate and Peanuts" could have allowed participants to bring in their own pre-existing biases towards the categories based solely on the name.

We further reinforced the equal priors between the two categories by explicitly telling participants that "half of the objects were *Elyk* and half were *Noek* ", and by genuinely drawing from either category according to those priors. Every trial had an equal 50% chance to yield either an *Elyk* or *Noek* object. We can further ensure that participants were trained on this prior by analyzing the objects presented to the participant. The only instance where this was not the case was in Experiment 2, where on a set number of trials (4) per block the participant was handed one of two objects outside of the usual set. However, these trials were also ensured to reinforce the equal priors by not favouring one category over the other. Two of the trials presented an *Elyk* object and the other two presented a *Noek* object. Analyzing these blocks for bias showed that they still reinforced the equal priors to the participant.

As we took these considerations into account in our experimental design and when deciding what information we made available to our participants, we believe that the assumption of equal Priors was valid to make in the context of our experiments.

5.9 Observer Confidence Ratings

We have largely ignored discussing human observer confidence ratings throughout the experiments featured in this thesis. However, measuring confidence was always part of our experimental design; in every trial, after the participant responded with the categorization of their given object stimuli, they then verbally gave a confidence rating from 50% to 100%. This seemed at the time a simple addition to our experimental design with little cost for including (either through adding overall experiment time or adding fatigue to the participant). However, as we briefly mentioned in Chapter 3, this measure proved to be uninformative.

We originally hypothesized that participant confidence ratings for their categorization responses might match the posterior probability as calculated by our observer model. Since this value represented the probability that a given object belonged to the *Elyk* Category, it seemed intuitive that this would also drive the observer's confidence in their response. If an object had a 95% posterior probability $P(Elyk|\mathbf{f}_m)$, then an observer responding "*Elyk* " should also report a high confidence in their response. While we generally found a linear trend that supported this hypothesis, there was a high variance across and within-participant responses that proved to be difficult to interpret. For some participants, confidence response behaviour seemed to follow different strategies, such as only responding within a small range like 60% to 80% rather than the full range available to them. Other participants might stick to only a small subset of responses. While it was possible to scale participant's responses to encapsulate the full range of 50% to 100%, making comparison easier, this also introduced a level of data manipulation that we did not feel justified in making. The only conclusion we could draw with any confidence was that our confidence rating measurement was contaminated by other factors, and was not a direct measure of the observer's actual confidence in their response.

Measurement of confidence is a deceptively complex problem. In *SDT*, the discrimination between response accuracy and the confidence in that response is called a *Type 2 Task*, and attempts to measure confidence judgements are often confounded by the presence of response bias (Galvin et al., 2003). For instance, we could have instead changed our experimental design to ask participants to respond with a binary "high" or "low" confidence rating every trial rather than a percentile from 50% to 100%. This would have conceptually been simpler for the participant. However, this measure would still be confounded by the individual observer's own biases which we cannot otherwise account for.

It is important to consider that confidence is a *retroactive* judgement being made by the observer. A human observer is unlikely to be cognitively aware of the calculations that went into their response—they only have the final output which drove their response. When an experiment asks the participant to report their confidence, they are not necessarily reading out the posterior probability $P(Elyk|\mathbf{f}_m)$ as we might hope they are. Instead, they are evaluating why they came to the judgement they did. The task becomes a meta-cognitive problem, and by asking for the observer's confidence we are instead more likely receiving the observer's ability and biases in observing their own responses.

However, there are indeed ways to measure an observer's confidence ratings despite these meta-cognitive complexities. Our initial hypothesis was not unfounded; Bayes' formula does give us the probability that an observer's response was correct. Methods have been outlined detailing how an experimental design might try to capture confidence ratings outside of these confounds. For instance, Maniscalco & Lau (2012) proposed a measure *meta-d'* which reflected the *Type 2 Task* sensitivity and could be calculated alongside the typical *d'* found in a *Signal Detection Theory* experimental paradigm (where an observer is discriminating between two distributions, a noise and a signal).

One interesting approach to measuring an observer's confidence ratings is the *Matching Probability* methodology, which uses a lottery-styled gambling game. (Massoni et al., 2014; Dienes & Seth, 2010). In this approach, an observer first gave a *Type 1* response to some discrimination task. They were told that they would be entered into a lottery based on their accuracy. If their response was correct, they would be awarded some amount. They then were asked to respond from 0% to 100% the minimum chance p they would be required to exchange their lottery ticket (based on their response) for a new lottery ticket that would be based instead on the chance p they gave. Once they gave this second response, the lottery began. First, a number l_1 was drawn (between 40% and 100%). If $p > l_1$, the observer would keep their initial lottery ticket. In this case, the observer would be rewarded if their *Type 1* discrimination response was correct, and otherwise receive nothing. If $p < l_1$, they would be forced to exchange for the second lottery ticket. In this case, a new number l_2 was drawn between 0% and 100% and if $l_1 > l_2$, the observer would be rewarded.

Under the *Matching Probability* methodology, the percentage response p given by the observer should directly relate to their belief in their discrimination task response. There are some key details to notice about this methodology. First, if they had high confidence that their response was correct, then the optimal outcome is for the observer to stick with their first lottery ticket which rewards them based on that response. The higher the percentage p they respond, the higher the chance that they will keep this initial lottery ticket. In the case that l_1 is drawn higher than their responded p and they are forced to exchange tickets, they are still highly likely to be rewarded since l_1, which will have had to be a higher number above an already high p, will be compared against the newly drawn l_2. Thus, if the observer is confident that their discrimination task response is correct, they are rewarded the most if they respond with a high p. Second, if the observer is not confident in their discrimination task response, the optimal outcome is for the observer to exchange the second lottery ticket. While this second ticket is not directly based on the percentage p responded by the participant, it holds a critical detail where $p < l_1$. This means that the lower the observer responds p, not only does the chance of exchanging lottery tickets increase but so too does the chance of l_1 *also* being a low number. In the lottery ticket 2 scenario, the observer is rewarded when $l_1 > l_2$. This means that the best chance for winning is with as high an l_1 value as possible. This prevents a strategy of simply "bottoming out" the p response when the observer has low confidence, as this hurts their chances of winning the second lottery ticket. As it turns out, the most optimal response in this scenario is to respond with a p that matches the actual probability that their *Type 1* discrimination task response is correct (Massoni et al., 2014).

Attempts were made to implement *Matching Probability* into our experimental design for our categorization task. We ran a pilot experiment with a small number of participants through such a design. However, it proved to be too complex in the context of our experiment for our participants to follow along with. The stimuli and task design in our experiment are a bit more complex than typical discrimination task experiments. Especially considering that our human participants were tasked with learning the novel categories throughout the experiment, it was determined that implementing the *Matching Probability* was outside the scope of our experiment.

In the end, our attempt to capture confidence ratings from our participants was naive. We were aware of this naivety as we designed our experiment, and while we attempted to address this with a more clever methodology, it proved to interfere with the main focus of our experiment in teaching human observers novel categories of objects through touch.

5.10 Future Directions

Throughout this thesis, we have explored and empirically tested the fundamental framework for a categorization task within the sense of touch using physical objects and allowing for the testing and exploration of complex probabilistic models. Our Bayesian observer model which was detailed in length in Chapter 2 illustrates a starting point for a Bayesian Categorization Model testable in an experimental setting. However, while it was outside of the scope of the current thesis to explore this model further, there is a wealth of variation and modification that could be made.

Testing an extended stimulus object set

One of the biggest limitations of the empirical studies presented here has been the limited resolution of the stimulus set used in our experiments. The features used to make up the categories only had 5 levels each (6 to 10 sides, and 4mm to 8mm dot spacing). This resulted in 25 total objects. The coarse resolution of this stimulus set was set mostly as a compromise for the cost and difficulty of being able to support more physical objects. For instance, if we had wanted to double the resolution such that each feature had 10 levels, the total number of physical objects needed to be created would inflate to 100. This would take a considerable amount of time to 3D print the objects, as each object took multiple hours to produce. It also would become infeasible for a human experimenter to run such an experiment where they would be required to select a single object out of an array of 100.

However, there are alternative stimulus object designs that could be explored in the future that would be able to increase the feature resolution while keeping the number of physical objects needed to a minimum. One such design could be a shell-and-face stimulus set, which would feature configurable and deconstructable objects. Rather than needing to 3D print every feature combination, an experimenter could print the feature levels independently and have objects that need to be put together. For instance, we could have five "sides" objects and five "dot spacing texture" objects, resulting in 10 total physical prints that would need to be created but retaining the 25 object resolution.

Exploring complex feature dimensions and categories

A natural progression of the experimental paradigm presented here would be to manipulate the feature dimensions used to define the categories. Different types of features may prove to be better suited for learning the category distributions than the ones chosen in our design. For instance, the *sides* featured turned out to be problematic for a few reasons being an odd feature for participants to measure in the experiment. This was seen anecdotally through participant self-reports and in the results of Experiment 3, where it was found that the Sides measurements were typically worse than the Dots measurements. The *sides* feature is also not a physically continuous variable, which makes it a bit inappropriate to consider as a true Gaussian probability density. The experiment should be repeated with a more appropriate feature, such as weight or thickness. The categories could also be constructed using a combination of three features. Derivations of the Bayesian observer model that could handle more than two features can be found in 2.

The categories used throughout our experiments were designed to be conceptually easy for observers to learn. To maintain this simplicity, the two features *sides* and *dot spacing* were considered to be independent variables. This was true for our experiment design; the category distributions were created in such a way that there was no correlation between the two features, $\rho = 0$. However, this may not always be the case in real-world categorization settings. It would be worthwhile to test if observers can learn tactile categories when a correlation is present.

To approach such a problem, some modification would be needed from our previous derivations in Chapter 2. Namely, we can rewrite the Bivariate Gaussian Equation 2.6 to include the correlation coefficient ρ.

$$P(n_m|C)P(d_m|C) = \frac{1}{2\pi\sigma_{Cn}\sigma_{Cd}\sqrt{1-\rho^2}} e^{-(\frac{(n_m-\mu_{Cn})^2}{2\sigma_{Cn}^2}+\frac{(d_m-\mu_{Cd})^2}{2\sigma_{Cd}^2}-\frac{2\rho(n_m-\mu_{Cn})(d_m-\mu_{Cd})}{\sigma_{Cn}\sigma_{Cd}})} \tag{5.1}$$

A categorization experiment could also be run with multiple categories, the derivations of which can be found in Chapter 2.

These variations of our experiment manipulating the number of features and categories present that observers are required to learn will test the validity of our model applied to humans at higher complexity levels. The Bayesian Categorization Model can in theory resolve any categorization task of any complexity. However, this could prove to be too complex for participants to categorize with the same degree of accuracy as our model. Learning such complex categories themselves could prove to be too difficult a task at a certain point. The upper limits of when and how our model loses validity in predicting human categorization performance would be an important step in understanding the extent of our model's predictive power.

Changing priors

As previously discussed in this chapter, we reinforced experimentally and assumed equal priors between the two categories for observers in the experiment. However, this does not necessarily need to be the case, and it would be experimentally simple to introduce unequal priors in this experimental design. This would allow questioning and exploration of whether observers are capable of learning more complex prior distributions.

We might expect that an observer with degraded or less informative sensory data would be more susceptible to changes in the Prior probabilities of the categories (Tong et al., 2016). In the two experiments we have discussed thus far, the Prior was intentionally set to be equal between each of the categories. A simple future experiment could involve manipulating these priors, either during or after category training. Under such a manipulation an observer would rely more on the Priors rather than the Likelihoods.

One experimental design could explore the impact of having the priors change during the experiment and observe how this impacts category learning and performance. For instance, participants could be trained on equal priors for the first half of the experiment, and then be trained on unequal priors (for instance, 70% *Elyk* and 30% *Noek*) halfway through, where *Elyk* objects are then more likely to appear than *Noek* objects.

Bridging to other forms of categorization experiments

Earlier we outlined two distinct paradigms within categorization research whose methodologies are often incompatible with making valid direct comparisons: The *Nominal-Ordinal Feature Paradigm* and the *Continuous Feature Paradigm*. Our work throughout this thesis fits within the latter paradigm. However, it would be interesting and worthwhile for future work to attempt to bridge these two paradigms together and form a more unifying understanding of categorization overall. This could be accomplished by fitting a stimulus and experiment design analogous to typical *Feature Classification* experiments that can be then expanded to a *Feature Distribution* experiment.

As an illustrative example, let's examine the first experiment described in Brooks & Hannah (2006)'s *Instantiated features and the use of "rules"*. This experiment depicted a 2 category task featuring the Bleebs and Ramuses line drawings. These stimuli had four features that could vary by two levels. Prototype Bleebs and Ramuses were coded as 0000 and 1111 respectively. The training set of stimuli presented to participants featured "one-aways", which were stimuli that had one feature differing from the prototype (for example, 0100 would be a one-away Bleeb). The four features examined in this stimulus set were the number of legs, the head shape, the body pattern, and the body shape. Of interest in this particular study is its consideration of instantiations, which were the visual representation of a certain feature. For instance, the pattern of Bleebs was coded as stripes, and there might be many different visual instantiations of stripes, but they would all be coded as 0 for stripes. This was an attempt to incorporate an element of variation in feature representations. Repeating their example, both humans and birds have two legs, but those legs never share the same appearance across the two categories.

This is an interesting consideration that outlines a substantial shortcoming in the way features are encoded through this experimental paradigm. By reducing a legs feature to just the number of legs, an observer who can *only* see the number of legs and no other qualities about them would truly consider birds and humans to have the same measurement regarding legs. However, there is far more information available regarding just the animal's legs to be able to differentiate them. Brooks & Hannah (2006) attempts to get around this problem by considering these different instantiations of the same feature and examining the effect of introducing novel instantiations. However, this problem is resolved when you consider the multiple observational measurements an observer can make across stimuli, and that these measurements are not necessarily binary distinctions being made but continuous measurements that can account for statistical distributions underlying them. It is even possible under this view that sub-categories such as binary stripes vs dots classification can be formed through these measurements, and that these sub-categories are what are combined into the overall category percept.

We put forth the following hypothesis relating the two paradigms together: The *Nominal-Ordinal Feature Paradigm* represents a special case of the *Continuous Feature Paradigm* where a continuous feature distribution is assumed to have zero variance. Thus, we might be able to fit stimuli such as the ones featured in our experiments by applying this condition directly, setting the *Within-Category Feature Variability* $\sigma_{Cf} = 0$ for each feature and category. We would further be able to mimic effect instantiations explored by Brooks & Hannah (2006) by taking another measurement close to the category mean.

Since our stimuli set only contains two features, it is currently not suitable to replicate *Nominal-Ordinal Feature Paradigm* studies in this way. We would need to create a new set of complex stimuli where we could replicate the Bleebs and Ramuses stimuli set conceptually. This could be done in several ways and through the use of different senses. Keeping in line with the work we presented through this thesis, let's consider a fully tactile experimental design. We could follow the stimulus design by Schwarzer et al. (1999) and create a set of objects with four defining features: size, shape, surface texture, and weight. These features can all be potentially continuous and thus would be able to fit under the *Continuous Feature Paradigm*. For the surface texture feature, rather than the grit of sandpaper we could create a measurement of stripes-to-dots with in-between measurements being some ambiguous transform combination of stripes and dots to varying degrees. With these new stimuli, we could replicate the experimental design of Brooks & Hannah (2006), as well as Schwarzer et al., 1999 and many other experiments and an initial phase of the experiment would confirm this valid replication.

However, the critical part of this proposed study would be to change the categories to use feature distributions (in effect, set $\sigma_{Cf} > 0$). Such an experiment would allow us to unify these methodologies and examine our hypotheses in how they relate to one another. Furthermore, we could examine if our Bayesian Categorization Model fits the results found in *Nominal-Ordinal Feature Paradigm* studies as well.

What may prove to be especially powerful with this approach will be the ability to include *Measurement Noise* in the *Nominal-Ordinal Feature Paradigm*. Currently, variation in how observers may measure the features under this paradigm is not considered. However, as we demonstrated in Chapter 4 this factor can potentially have a major impact on participant categorization accuracy. If *Measurement Noise* is a present factor then even under the special case of the *Nominal-Ordinal Feature Paradigm* where $\sigma_{Cf} = 0$, the resulting distributions including this sensory noise variance are $\sigma_{Cf+f_m} > 0$.

Further extensions of our work

We have focused our considerations of categories on representations that are generally rooted in the natural world. These are mainly comprised of features that can be measured directly through the senses resulting in a holistic, sensed percept. However, the concept of categories can extend to be far more abstract than this; things such as concepts, ideas, emotions, or descriptions of things that do not pertain to a physical, tangible representation can be understood through categories. Furthermore, if we can understand the factors that influence these kinds of abstract categories, we could theoretically apply approaches like our Bayesian Categorization Model to understand how these categories are perceived.

Perception of identity is one such abstract concept that could be understood through the lens of categorization. Of particular interest is the perception of gender identity: how does one person perceive another person's gender? This question is especially relevant to transgender populations whose experiences usually involve an incongruence with their gender identity and how other people identify them. What are the different factors that someone is using, either intentionally or not, in their perception of another's gender? What are the main observations being made that contribute to gender perception, and how are these observations and perceptions changing as a trans person takes steps to transition? Considering these questions through our Bayesian Categorization Model reveals that to understand how categorical gender perception is made, we must understand what is driving the underlying observations and categories people hold on gender. Human physical characteristics vary wildly across people. Even if an overall average measurement differs for male-identifying individuals and female-identifying individuals, this variance means these measurements overlap greatly, and so too should the categories. With so many potential features to make observations over, gender perception is likely a complex categorical problem.

Exposure to these gender categories is an important driving force for these underlying categorical distributions, and poor representation of people of differing identities and appearances throughout media might lead to very biased and inaccurate categorical gender distributions throughout the population. The treatment of gender as a binary category might be driven by a biased sampling of the population (through movies, television, etc) that doesn't properly represent an even sampling of the human population. By just considering the observations involved in gender perception as distributions with variation, and that these distributions are combined to form an overall gender percept, we can see how gender perception becomes reduced to a bimodal distribution. We might be able to calculate an overall posterior for either gender category using our model if we had access to all the observations involved and the distributions of those observations within the categories. Of course, gender is not truly a binary. Gender-nonconforming people and non-binary people might not fit nor want to fit within a binary gender categorization at all. Finding ways to shift or expand the gender categories that are learned throughout the population would lessen the amount of incongruent gender perception overall. An interesting study along this line of thinking might involve seeing if these underlying category distributions could be changed simply through further exposure. In other words, does more diverse representation change people's perception of other genders to better reflect the variation found in people?

Of course, studies of this nature would need to be done with extreme care and be informed by representatives of transgender and non-binary communities. Scientific study on marginalized groups, even with good intentions and in pursuit of fundamental research to further understanding, can very easily be used in turn to further marginalize, stigmatize and pathologize those groups.

This consideration of identity perception as a problem approachable through categorization could also hold fascinating exploration when considering plural people. Plurality is a phenomenon of experiencing some form of multiple selves within one body, sometimes referred to as headmates or alters. This is often related to and known as Dissociative Identity Disorder (DID), although plurality and DID are not exactly the same thing; while DID is a specific and narrow diagnosis, plurality represents a wider umbrella of diverse experiences as varied as the night sky. Plural experiences are often associated with trauma, however, this is not always the case and some plural people might have no history of trauma at all (Yarbrough, 2018). A plural person may have two or more people or unique identities within their system—a term that plural people sometimes use to refer to themselves as a collective.

Spending time getting to know a plural system and the members of their system can eventually lead to recognizing and identifying those persons based on even subtle behavioural cues. We could again attempt to understand this identity perception by an outside observer on individual identities through our Bayesian Categorization Model.

Each headmate within a plural system could be understood as a category with different behavioural observations being the features an observer is measuring, such as patterns in speech, vocal pitch, body language, and interests. These behavioural cues might naturally have some amount of variance within a headmate "category" of what an observer might expect or can accurately measure themselves. These categories can also overlap to varying degrees, making identification a complex categorization problem. However, given enough time and information an observer may learn to accurately identify headmates in a system. Understanding this learning process could further inform how identity is perceived overall.

5.11 Conclusion

Throughout the study of categorization, researchers have employed various methodologies, stimuli, and models to understand how the brain undergoes this process. Each of these approaches has its own strengths and weaknesses for studying different aspects of categorization. However, we found two gap existed in this research. There is a lack of direct Bayesian approaches that consider categories and stimuli based on the statistical information underlying them. There has also been limited exploration on the tactile system in categorization. This thesis aimed to bridge both of these gaps by exploring categorization through our Bayesian model and investigate human performance on a categorization task when limited to their sense of touch.

The empirical and theoretical work presented here puts forth an experimental framework for training and testing novel categories in human participants using only the sense of touch. Furthermore, we demonstrate that a Bayesian Categorization Model which includes sensory *Measurement Noise* can predict overall human performance in a tactile-based categorization task. This work supports the hypothesis that a Bayesian-based computational perceptual model may represent the underlying processing that a human observer undergoes during categorization tasks.

Bibliography

Ahn, W.-K. & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science* 16(1), 81–121.

Alais, D. & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology* 14(3), 257–262.

Allen, S. W. & Brooks, L. R. (1991). Specializing the Operation of an Explicit Rule. *Journal of Experimental Psychology: General* 230(2), 3–19.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review* 98(3), 409–429.

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science* 12(2), 157–162.

Ashby, F. G. & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(1), 33–53.

Bankieris, K. R., Bejjanki, V. R. & Aslin, R. N. (2017). Sensory cue-combination in the context of newly learned categories. *Scientific Reports* 7(1), 10890.

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical Transactions of the Royal Society of London* 53, 370–418.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E. & Pouget, A. (2008). Probabilistic Population Codes for Bayesian Decision Making. *Neuron* 60(6), 1142–1152.

Bejjanki, V. R., Clayards, M., Knill, D. C. & Aslin, R. N. (2011). Cue Integration in Categorical Tasks: Insights from Audio-Visual Speech Perception. *PLoS ONE* 6(5). Ed. by D. G. Pelli, e19812.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances.

Brooks, L. R., Squire-Graydon, R. & Wood, T. J. (2007). Diversion of attention in everyday concept learning: identification in the service of use. *Memory & cognition* 35(1), 1–14.

Brooks, L. R. & Hannah, S. D. (2006). Instantiated features and the use of "rules". *Journal of Experimental Psychology: General* 135(2), 133–151.

Churchland, A. K., Kiani, R. & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience* 11(6), 693–702.

Dienes, Z. (2007). Subjective measures of unconscious knowledge. In: *Progress in Brain Research*. Vol. 168. Elsevier, 49–269.

Dienes, Z. & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition* 19(2), 674–681.

Ernst, M. O. & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415(6870), 429–433.

Ernst, M. O. & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences* 8(4), 162–169.

Feldman, J. (2012). Bayesian models of perception: a tutorial introduction. *Handbook of Perceptual Organization*.

Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. (2003). A Comparison of Primate Prefrontal and Inferior Temporal Cortices during Visual Categorization. *The Journal of Neuroscience* 23(12), 5235–5246.

Galvin, S. J., Podd, J. V., Drga, V. & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review* 10(4), 843–876.

Gepshtein, S. & Banks, M. S. (2003). Viewing Geometry Determines How Vision and Haptics Combine in Size Perception. *Current Biology* 13(6), 483–488.

Hannah, S. D. & Brooks, L. R. (2009). Featuring familiarity: How a familiar feature instantiation influences categorization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 63(4), 263–275.

Healy, A. F. & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory* 7(5), 344.

Heekeren, H. R., Marrett, S., Bandettini, P. A. & Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature* 431(7010), 859–862.

Hillis, J. M., Ernst, M. O., Banks, M. S. & Landy, M. S. (2002). Combining Sensory Information: Mandatory Fusion Within, but Not Between, Senses. *Science* 298(5598), 1627–1630.

Iordan, M. C., Greene, M. R., Beck, D. M. & Fei-Fei, L. (2015). Basic Level Category Structure Emerges Gradually across Human Ventral Visual Cortex. *Journal of Cognitive Neuroscience* 27(7), 1427–1446.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences* 6(8), 345–350.

Kalckert, A. & Ehrsson, H. H. (2014). The moving rubber hand illusion revisited: Comparing movements and visuotactile stimulation to induce illusory ownership. *Consciousness and Cognition* 26, 117–132.

Kemler Nelson, D. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior* 23(6), 734–759.

Klatzky, R. L., Lederman, S. J. & Matula, D. E. (1993). Haptic exploration in the presence of vision. *Journal of Experimental Psychology: Human Perception and Performance* 19(4), 726–743.

Klatzky, R. L., Lederman, S. J. & Metzger, V. A. (1985). Identifying objects by touch: An "expert system". *Perception & Psychophysics* 37(4), 4.

Knill, D. C. (2007). Robust cue integration : A Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *Journal of Vision* 7(5), 1–24.

Knill, D. C. & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences* 27(12), 712–719.

Knill, D. C. & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research* 43(24), 2539–2558.

Koenig, P., Smith, E. E., Glosser, G., DeVita, C., Moore, P., McMillan, C., Gee, J. & Grossman, M. (2005). The neural basis for novel semantic categorization. *NeuroImage* 24(2), 369–383.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B. & Shams, L. (2007). Causal Inference in Multisensory Perception. *PLoS ONE* 2(9). Ed. by O. Sporns, e943.

Kubovy, M., Rapoport, A. & Tversky, A. (1971). Deterministic vs Probabilistic strategies in detection. *Attention, perception & psychophysics* 9(5), 427–429.

Lakshminarasimhan, K. J., Pouget, A., DeAngelis, G. C., Angelaki, D. E. & Pitkow, X. (2018). Inferring decoding strategies for multiple correlated neural populations. *PLoS computational biology* 14(9), e1006371.

Lederman, S. J. & Klatzky, R. L. (1993). Extracting object properties through haptic exploration. *Acta Psychologica* 84(1), 29–40.

Lee, W. & Janke, M. (1964). Categorizing externally distributed stimulus samples for three continua. *Journal of Experimental Psychology* 68(4), 376–382.

Li, L., Chan, A., Iqbal, S. M. & Goldreich, D. (2017). An Adaptation-Induced Repulsion Illusion in Tactile Spatial Perception. *Frontiers in Human Neuroscience* 11, 331.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic bulletin & review* 9(4), 829–835.

Macmillan, N. A. & Creelman, C. D. (2005). *Detection theory: a user's guide*. 2nd ed. Mahwah, N.J: Lawrence Erlbaum Associates. 492 pp.

Maniscalco, B. & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition* 21(1), 422–430.

Markman, A. B. & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin* 129(4), 592–613.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman and Co.

Massoni, S., Gajdos, T. & Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology* 5(1455).

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines* 1(2), 185–196.

Medin, D. L. & Schaffer, M. M. (1978). Context Theory of Classification Learning. *Psychological Review* 85(3), 207–238.

Minda, J. P. & Ross, B. H. (2004). Learning categories by making predictions: An investigation of indirect category learning. *Memory & Cognition* 32(8), 1355–1368.

Moreno-Bote, R., Knill, D. C. & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences* 108(30), 12491–12496.

Murray, R. F., Patel, K. & Yee, A. (2015). Posterior Probability Matching and Human Perceptual Decision Making. *PLOS Computational Biology* 11(6). Ed. by J. X. O'Reilly, e1004342.

Newell, F. N., Ernst, M. O., Tjan, B. S. & Bülthoff, H. H. (2001). Viewpoint Dependence in Visual and Haptic Object Recognition. *Psychological Science* 12(1), 37–42.

Norman, J. F., Kappers, A. M. L., Beers, A. M., Scott, a. K., Norman, H. F. & Koenderink, J. J. (2011). Aging and the haptic perception of 3D surface shape. *Attention, perception & psychophysics* 73, 908–918.

Norman, J. F., Cheeseman, J. R., Adkins, O. C., Cox, A. G., Rogers, C. E., Dowell, C. J., Baxter, M. W., Norman, H. F. & Reyes, C. M. (2015). Aging and solid shape recognition: Vision and haptics. *Vision Research* 115, 113–118.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General* 115(1), 39.

O'Connell, R. G., Shadlen, M. N., Wong-Lin, K. & Kelly, S. P. (2018). Bridging Neural and Computational Viewpoints on Perceptual Decision-Making. *Trends in Neurosciences* 41(11), 838–852.

Peters, R. M., Staibano, P. & Goldreich, D. (2015). Tactile orientation perception: an ideal observer analysis of human psychophysical performance in relation to macaque area 3b receptive fields. *Journal of neurophysiology* 114(6), 3076–96.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research* 41(24), 3145–3161.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology* 77(3), 353–363.

Pouget, A., Dayan, P. & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience* 1(2), 125–132.

Pouget, A., Dayan, P. & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience* 26(1), 381–410.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology* 3(3), 382–407.

Rock, I. & Victor, J. (1964). Vision and touch: An experimentally created conflict between the two senses. *Science* 143(3606), 594–596.

Rorie, A. E., Gao, J., McClelland, J. L. & Newsome, W. T. (2010). Integration of Sensory and Reward Information during Perceptual Decision-Making in Lateral Intraparietal Cortex (LIP) of the Macaque Monkey. *PLoS ONE* 5(2). Ed. by V. Brezina, e9308.

Rosas, P., Wagemans, J., Ernst, M. O. & Wichmann, F. A. (2005). Texture and haptic cues in slant discrimination: reliability-based cue weighting without statistically optimal cue combination. *Journal of the Optical Society of America A* 22(5), 801.

Ross, B. H. (2000). The effects of category use on learned categories. *Memory & Cognition* 28(1), 51–63.

Saizman, C. D., Britten, K. H. & Newsome, W. T. (1990). Cortical microstimulation influences perceptual judgments of motion direction. *Nature* 346(6280), 174–177.

Saizman, C. D. & Newsome, W. T. (1994). Neural mechanisms for forming a perceptual decision. *Science* 264(8), 231–237.

Samad, M., Chung, A. J. & Shams, L. (2015). Perception of body ownership is driven by Bayesian sensory inference. *PLoS ONE* 10(2), 1–23.

Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review* 117(4), 1144–1167.

Schwarzer, G., Küfer, I. & Wilkening, F. (1999). Learning categories by touch: On the development of holistic and analytic processing. *Memory & Cognition* 27(5), 868–877.

Shadlen, M. N. & Newsome, W. T. (2001). Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey. *Journal of Neurophysiology* 86(4), 1916–1936.

Sigala, N. & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415(6869), 318–320.

Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J. & Ashby, F. G. (2010). Implicit and explicit category learning by macaques (Macaca mulatta) and humans (Homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes* 36(1), 54–65.

Smith, J. D., Berg, M. E., Cook, R. G., Murphy, M. S., Crossley, M. J., Boomer, J., Spiering, B., Beran, M. J., Church, B. A., Ashby, F. G. & Grace, R. C. (2012). Implicit and explicit categorization: A tale of four species. *Neuroscience & Biobehavioral Reviews* 36(10), 2355–2369.

Sripati, A. P. (2006). Spatiotemporal Receptive Fields of Peripheral Afferents and Cortical Area 3b and 1 Neurons in the Primate Somatosensory System. *Journal of Neuroscience* 26(7), 2101–2114.

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.

Tong, J., Ngo, V. & Daniel; G. (2016). Tactile length contraction as Bayesian inference. *Journal of Neurophysiology* 116(2), 369–379.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 1: behavioural study: Macaque visual categorization. *European Journal of Neuroscience* 11(4), 1223–1238.

Weiss, Y., Simoncelli, E. P. & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience* 5(6), 598–604.

Yamauchi, T. & Markman, A. B. (2000). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory & Cognition* 28(1), 64–78.

Yarbrough, E. (2018). *Transgender mental health.* In collab. with A. P. A. Publishing. First edition. Arlington, VA: American Psychiatric Association Publishing. 323 pp.

Yuille, A. & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In: *Perception as Bayesian Inference*. Cambridge, UK: Cambridge University Press, 123–161.