Exploring metacognitive confidence in categorisation judgements



Sarah Ashcroft-Jones

The Queen's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Experimental Psychology

Trinity 2023

"Is fearr eolais ná iontas"

SEANFHOCAIL

Acknowledgements

It is not easy to begin to name and adequately thank everyone who has helped me to finish this thesis. It truly takes a village and I have been incredibly lucky with the village I have built around me while here in Oxford. It is here that I have learnt the meaning of *"ar scath a chéile a mhaireann na daoine"* - that we are shielded from the sun by each other.

First and foremost, I have an incredible academic support network to thank. To my supervisors, Prof. Nick Yeung and Prof. Jill O'Reilly. Thank you for your guidance, support, and patience. To Nick and the ACCLab - thank you for creating such an incredible community within academia where we can learn (and make up terrible scientific album names) together. To all the members of the ACCLab past and present thank you for your support, ideas, and problem-solving skills. A special thanks to Joshua for his guidance and time in solving computational conundrums, to Matt for teaching me that programming is as much about getting things wrong as it is about getting things right, and to Sriraj for the many coffees and friendship. And most of all to Nick - thank you for your time, kindness, and wisdom. For giving me the time and space to learn, and the guidance to direct that learning effectively. It has been such a privilege to work with you and to grow as a scientist under your mentorship.

Within academia, over the course of my DPhil, I have also had the privilege of engaging with inspiring and supportive networks such as Reproducible Research Oxford (RROx), ReproducibiliTea, the Junior Researcher Programme (JRP), the Framework for Open and Reproducible Research Training (FORRT), the Oxford Statistics Bookclub, and the UKRN. I

am grateful to the individuals and collaborations that have emerged from these networks to support me on both a personal and professional level. A special mention to Dr. Kai Ruggeri for the guidance, friendship, and endless opportunities over the years.

It also goes without saying that this thesis would not have been possible without the generosity of my funders - the Medical Research Council and the Department of Experimental Psychology - or the support of The Queen's College Oxford. Thank you for the chance to begin my career in science here at Oxford and for enabling this work to happen. Thank you also to my many participants for giving their time to complete my tasks, without whom no behavioural data would have been collected, and none of this science would have been completed!

On a more personal level, an enormous gratitude goes to my family. To Mama and Papa, thank you for the decades of support, your words of quiet confidence, and your unending belief that I will succeed at whatever I put my hand to. To Papa for always indulging my "satiable curtiosity" and Mama for believing I could rule the world if I was so inclined, thank you both. To Lolo, thank you for being my constant cheerleader, for being my first-ever teacher (those long division sums have gotten me far!), and for being the best big sister (my very best skin 'n' blister). *Grá mo chroí go deo sibh.*

To my friends - *is fearr cairde ná ór.* Thank you for being yourselves and for reminding me that I am myself whenever I forget. To Laura and the boys, thank you for being there for me in every sense of the word - for lifting me up and grounding me in equal measure. I would be lost

without you. To Hannah for your unfailing support, the many cups of tea, the countless dinners, and most of all the authentic warmth of your friendship. To Ed and Jade for being your wonderful selves, for your genuine excitement for my successes, and for the many mugs of hot chocolates - thank you. To Manuela for the laughter, the hugs, and the boundless cheerleading - I am so grateful. To Ken, thank you for your friendship, for the many hours of conversation, and for your sincere support as I take this next step. To the Board Game crew, the Film Club gang, the Paperboat folks, and all those I will unwittingly fail to mention by name - thank you.

Go raibh míle maith agaibh go léir.

Contents

Abstract **i**

1 Introduction **4**

1.1 Introduction 5

1.2 Thesis Outline 10

1.3 Open Science & Materials 13

2 Literature Review **16**

2.1 Introduction 17

2.2 What is a decision? 18

2.3 Metacognition 22

2.4 Decision Confidence 28
2.4.1 Measuring Decision Confidence 30
2.4.2 Metacognitive Sensitivity & Bias 32
2.4.3 Uses of Decision Confidence 35
2.4.4 Different Types of Uncertainty 38
2.5 Modelling Decision Confidence 41
2.5.1 Single process models of confidence 42
2.5.2 Dual process models of confidence 45
2.5.3 Ideal observer models of confidence 48
2.5.4 Comparing different models of confidence 56
2.6 Categories & Categorisation 59
2.6.1 What is a category? 59
2.6.2 Models of Categorisation 60
2.6.3 Multi-dimensional Categories 65
2.6.4 Categories & Confidence 68
2.6.5 Multi-dimensional Categories & Confidence 71
2.7 Thesis Overview 77

3 Computing Confidence **80**

Abstract 81
3.1 Introduction 83
3.1.1 Reanalysis Rationale 89
3.1.2 The Original Experiment 94
3.1.3 The Original Models 99

3.2 Exploring the Original Models 110

3.3 The Model Extension 122

3.4 General Discussion 145

4 Category Information & Confidence Sensitivity 152

Abstract 153

4.1 Introduction 155

4.2 Study 2 163

4.2.1 Experimental Questions 164

4.2.2 Methods 164

4.2.3 Experimental Design 166

4.2.4 Results 177

4.2.5 Discussion 187

4.3 Study 2b: Direct Replication 188

4.3.1 Methods 188

4.3.2 Results 190

4.3.3 Discussion 194

4.4 Study 3: Conceptual Replication 196

4.4.1 Methods 196

4.4.2 Results 201

4.4.3 Discussion 204

4.5 Model-based Analysis 206

4.5.1 Model Derivation 207

4.5.2 Model Fitting & Results 211

4.5.3 Discussion 214

4.6 General Discussion 216

5 Discrete Categories **221**

Abstract 222

5.1 Introduction 224

5.2 Study 4 231

5.2.1 Introduction 231

5.2.2 Methods 237

5.2.3 Results 246

5.2.4 Discussion 275

5.3 Study 5 277

5.3.1 Introduction 277

5.3.2 Methods 280

5.3.3 Results 283

5.3.4 Discussion 300

5.4 General Discussion 301

6 Discussion & Synthesis **309**

6.1 Introduction 310

6.2 Overview of Results 314

6.3 Themes 318

6.4 Future Directions 323

6.5 Conclusions 329

References **332**

A Appendices **361**

A.1 Study 2: Confidence Simulations 362

A.1.1 Simulation Constraints 362

A.1.2 Defining the Category Boundary 362

A.1.3 Simulation Method 363

A.1.4 Simulation Results 365

A.2 Study 2: Response Analysis 367

Thesis Abstract

This thesis explores metacognitive decision confidence in category judgements. It aims to investigate how different types of information affect our confidence in the choices we make about category membership.

Over three experimental chapters, this thesis explores the interplay between the properties and structure of category spaces and subsequent confidence reports. Chapter 1 takes a computational model-based approach to contrast normative Bayesian and heuristic rule-based characterisations of confidence, with ambiguous results. The second experimental chapter, therefore, takes a more direct approach, using an experimental manipulation of category variance to directly contrast the predictions of different models of confidence. The third experimental chapter extends this approach to consider more complex category structures that involve multiple dimensions and differing logical structures to the categories.

In this way, over the course of this thesis, decision-making and decision confidence are investigated in multiple behavioural tasks with different parameters, decision boundaries, and distribution properties. Collectively this research seeks to better understand how information about categories is used or represented by decision-makers, and, in turn, where the cognitive computations of decision confidence may derive from and in what contexts.

Through this combination of behavioural analyses and computational modelling exercises, the sensitivity of confidence to the properties and structure of category distributions and bounds is evidenced. The tension between heuristic and normative Bayesian models of decision confidence is

explored and the results suggest that the relationship between confidence and heuristic distance-to-bound models is complex – with the simplicity of the original theory missing out on the rich signal about the category distribution representations that is conveyed in these data. The theoretical, practical, and methodological implications are discussed.

CHAPTER 1

Introduction

1.1 | Introduction

This thesis explores metacognitive decision confidence regarding a particular kind of choice – categorical decisions. Therefore, it aims to investigate how different kinds of uncertainty and different types of information affect our confidence in the choices we make about category membership. To explain this approach more clearly, consider the case of a clinical dermatologist who is trying to determine if an uneven blemish on the skin is benign or malignant. To come to a decision, they will necessarily rely on multiple complementary sources of information such as visual assessment, age of the patient, previously observed cases, and general occurrence of such malignancies. They might even have a criterion boundary against which to compare this particular case to decide a cut-off point for advising further treatment or not. By combining these pieces of information, all with their own intrinsic levels of certainty and uncertainty, the clinician can make an informed decision, with some level of confidence, about what the correct course of action might be. However, what this process looks like in the "black box" of human cognition remains opaque. Indeed, it is largely unclear how a metacognitive confidence judgement emerges from these intersecting sources of information. For example, one unanswered question is whether decision-makers weigh certain pieces of information differently and how they might account for the uncertainty, or variance, intrinsic to the choice categories they are deciding between.

This thesis aims to elucidate some of these issues. Through a combination

of behavioural decision-making tasks and computational modelling, this project will investigate the different levels of information processing that co-occur when making a single decision and reporting confidence in that choice. For example, the uncertainty faced by an individual at the level of the sensory systems could be separated out from uncertainty in the environment and explored using a computational modelling framework. Similarly, how the different levels of variability in different categories (for example, in the above-described categories of benign and malignant skin lesions) influence the confidence of a decision-maker is also an area of research that is, as yet, largely unexplored. Investigating how these uncertainties are utilised by the observer, and how they are expressed in confidence judgements, is a particularly novel aspect of this project. Therefore, this thesis will attempt to quantify and probe the independent levels of uncertainty that may influence individuals' sense of decision confidence and explore how confidence might map onto the functional representations of categories in cognition. By drawing together theories from cognitive science, behavioural economics, categorisation, and decision-making models, this research will attempt to demonstrate how different sources of evidence are used to compute confidence judgements. It is the hope that this will allow not only for an improved understanding of decision confidence as a metacognitive report but also of how information about categories and their mental representations are conceptualised by decision-makers.

Explorations of decision-making often have to balance the tension between predictions of rational or optimal behaviour and heuristic approaches (Adler &

Ma, 2018; Kahneman & Tversky, 1974; Tversky & Kahneman, 1974). In particular, Bayesian models lean into a definition of optimal and rational decision-making which involves the full and complete use of all available decision-relevant information both past and present (Beierholm, 2013). However, in certain environments, an imperfect but functional representation of categories and their boundaries may be more effective in terms of providing a "good enough" guide to the decision-maker without exhausting their cognitive resources. Therefore, understanding how different sources of evidence, and their variances or uncertainty, are combined will lead to a more precise understanding of how traditionally "sub-optimal" decisions may be made in terms of over- or under-confidence. For example, returning to a medical diagnosis scenario, a radiographer may have better-calibrated confidence in their diagnosis when they are dealing with a novel pathology that has not yet been reported in the UK, as compared with a situation where they are conducting a scan of a more routine pathology for which they can rely on heuristics to make a diagnosis. In the first instance, the novelty of the pathology and the high uncertainty around the nature of the sample, despite clear imaging, may lead the radiographer to be highly cognisant of the uncertainties of the treatment courses being selected. By comparison, the identification of what seems to be a more routine pathology may lead to over-confidence in the treatment selection, despite the fact that perhaps the common pathology may have a wide-ranging presentation that is highly variable and, thus, should reduce the confidence in the diagnosis. However, perhaps, in the case of the more common pathology, the regularity of the assessment of

these types of samples means that the clinician can safely underweight this source of uncertainty in favour of providing a swift diagnosis and treatment plan for a patient. These differences in approach by a decision-maker are rarely formally considered by cognitive models, despite the value they hold to explain the ease with which human decision-makers navigate in diverse, complicated, and information-rich environments. Furthermore, in the study of decision confidence, many studies rely on simple decision tasks where all information is knowable and, if known, will lead to a definite decision. However, these conditions are often not met in real-world decisions, perhaps meaning that interesting experimental questions remain unaddressed or simply unnoticed.

One area of inquiry that may lend itself to addressing some of these questions is that of categorisation and category judgements. For example, thinking back to the dermatologist conducting a routine examination, there are potential features of the categories of benign versus malignant growths that are simply not captured by common perceptual decision-making tasks. Indeed, perhaps the categories have differing variances such that malignant growths tend to have similar properties whereas benign ones are quite variable, or perhaps some of these category exemplars are actually indistinguishable according to available diagnostic tests. It is also possible that these categories are best distinguished based on the intersection of multiple dimensions such as hue, size, growth rate, symmetry, and irritation. It is further possible that these dimensions are inequitable according to their relevance to diagnosis – their relevance to category classification. For example, maybe some properties such as either strong pig-

mentation or fast growth alone are enough to indicate malignancy, whereas light pigmentation and slow growth must both be true to have confidence that something is benign. By leveraging some of the features of what is a very large and rich space of possibilities more complex questions about confidence can be addressed and a more nuanced understanding of the role of uncertainty in decision-making can be gleaned.

Therefore, to explore some of this complexity, my research investigates how decision-makers utilise information about categories and category distributions to make confidence reports about their choices. This research aims to increase our understanding of how such information may be represented and utilised. Whereas much research has looked at the uncertainty that arises from perceptual judgements (e.g., Fleming & Lau (2014)), comparatively little attention has been devoted to uncertainty stemming from more naturalistic categories that are low in perceptual uncertainty, and how category uncertainty could be reflected in metacognitive judgements. In this way, exploring different distribution types, non-overlapping stimulus sets, multi-dimensional categories, and variability in learning periods, may all yield fruitful results. Thus, this thesis aims to explore how people make decisions and determine their confidence when making judgements about an important decision type: category membership.

1.2 | Thesis Outline

To explore some of these questions and topics, this thesis consists of three interlinked threads of work which are documented in three experimental chapters. Although each chapter can stand alone in its own right, collectively they work to bridge and plot a study area that is, as yet, relatively sparse in terms of research. Taken together, the experimental chapters of this thesis work to broach a better understanding of the computation of metacognitive confidence by exploring different category spaces and decision boundaries. This allows us to not only establish a better understanding of how confidence judgements arise but also to explore what these reports can teach us about the cognitive representation and information processing that occurs in category membership judgements.

The first experimental chapter consists of a computational modelling exercise. This was conducted based on secondary data analysis from an experiment I had previously conducted to explore the computation of confidence judgements under different task difficulties. This work tries to test the idea of differential task representation by decision-makers when placed under different experimental demands. Much like the radiographer who may rely on heuristics and be overconfident when a case seems routine, a perceptual decision-maker may simplify their perceptual category representations when the task appears to be less demanding, resulting in inaccurate assessments of their confidence relative to the reality of the stimuli they are examining. By comparison, when a task is more difficult, traditionally "optimal" full representation of the stimulus

categories may be prudent and may result in more accurate corresponding confidence reports. Throughout this chapter, the assumption that "optimal" task representation requires a full and complete representation of category distributions is challenged based on the literature, and alternative models of confidence are explored.

The second experimental chapter refocuses more specifically on category judgements in a novel experimental paradigm. Once again challenging the need for full category distribution representation to produce useful confidence judgements, in this chapter the "distance-to-bound" model of confidence is explored and extended using more naturalistic category distributions with variable spread. Across a series of behavioural studies, decision-makers were tasked with identifying the category to which a stimulus belongs when perceptual uncertainty is low. Uncertainty in these tasks and decisions instead arises from the overlapping category distributions (e.g., some benign and malignant skin growths looking identical) which were created to have differences in variances (e.g., malignant growths looking more similar to one another than benign growths) such that the observer should be more certain about stimuli that arise from one category than the other. How confidence reports vary across and within these categories gives insight into the kinds of information that are used by decision-makers to make their metacognitive judgements. This work is accompanied by a modelling exercise that seeks to further clarify the way in which individuals are capturing the task space and how that relates to their confidence reports.

The third experimental chapter of this thesis steps further away from perceptual judgements and deeper into the category learning literature by exploring confidence judgements in discrete category spaces with multiple dimensions. Here the "distance-to-bound" model of confidence is again explored and tested, as the notion of distance itself must be redefined to account for the characteristics of the categories themselves in this stimulus space. A key feature of the category space considered is its asymmetry: Whereas one category is defined by a conjunction of features, the other is defined disjunctively. Therefore, this category space also allows for specific predictions to be made about the computation of confidence. In particular, this experimental space further allows for the exploration of inequities of information needed to map out the categories of certain decision spaces – some categories are easier to define than others and thus the computation of confidence in each category may arise from different information and models in each case. Indeed, this inequity results in the possibility of a logical rule-based model being leveraged in one category, with a more involved model being required to integrate the category dimensions in the other. This difference in model complexity speaks to the work of the first experimental chapter where task demands were varied to elicit different decision model representations. This idea, along with further models of decision confidence, is explored in this final experimental chapter.

1.3 | Open Science & Materials

Across the work of this thesis, in so far as was tenable, I have followed the principles of Open Science research and reporting. In this way, the data, materials, and supporting analysis code for the experimental sections of this thesis are available publicly online.

In a similar vein, due to the exploratory nature of much of the behavioural work conducted in this thesis, my research has been structured to allow for internal replications with pre-registered analysis plans and exclusion criteria where practical and informative. This means that an initial study in a line of work may be followed by a direct pre-registered replication, for example. In this way, the robustness of the results may be assessed in a setting where the researcher degrees of freedom are constrained (Simmons et al., 2011).

In this thesis, materials such as experimental code used to run the online experiments are available on public repositories on GitHub while data, analysis pipelines, and associated pre-registrations are housed in public projects on the Open Science Framework (OSF). This allows the reader to directly explore and interact with the materials and data to understand how the findings were concluded. Along the way, these resources will be clearly signposted with direct links to the relevant materials throughout the experimental chapters.

The necessity of such clarity in methods emerges from the replication and reproducibility crises that have motivated theoretical debate and changes of practice both within and beyond the psychological sciences (Ioannidis, 2005; Sim-

mons et al., 2011). Indeed, when considering the importance of balancing rigour and exploration in the sciences the notion of transparency becomes paramount. Transparency is particularly important when we consider the human elements of the research process and the fallibility that emerges from them (Strand, 2023). Transparency in science also increases the credibility and verifiability of the results and increases the possibility of reuse of both data and materials generated in the course of the study (Klein et al., 2018). These practices can, therefore, contribute to a more robust and interrogable literature while also promoting a collaborative scientific culture. As such, this thesis aims, above all, to be transparent in its reporting of the methods, analyses, and findings throughout both reporting in the experimental chapters and discussions of the results.

CHAPTER 2

Literature Review

2.1 | Introduction

This thesis aims to address the intersection of models of decision confidence and categorisation in order to discuss and acknowledge different sources and types of uncertainty that may arise in decision-making. These fields each have a rich history of research and theory which have contributed to the current understanding of cognitive processing and metacognitive reporting. Therefore, in this chapter, the background and history of each field of decision-making and categorisation will first be described and addressed in turn to establish the existing gaps in knowledge as well as understand the intersections of each as models of information processing. This will help to ground the experimental chapters that follow and situate the reader in the language and discourse of each field.

While an exhaustive account of each field is not possible, this chapter aims to give enough background to allow for an articulation of the main debates and understandings of both areas of study as relevant to this thesis work. In this way, the motivation for the experimental chapters that follow, alongside the reasoning for certain methodological choices, will be clear. Therefore, this chapter will begin with a discussion of decision-making before exploring metacognitive confidence and, finally, categorisation. This structure mirrors the evolution of the empirical work of this thesis which begins with more traditional decision-making tasks and confidence reports before evolving in later studies into more distinctive categorisation tasks.

2.2 | What is a decision?

Decisions are a process. In short, they involve choosing between a set of possible actions or options based on certain criteria or goals (Wilson & Keil, 1999). Decision-making, therefore, implies that there are multiple options that must be selected between, even if the choice is simply to take or leave the single option that is available. As such, given two or more world states, when making a decision, an individual expresses a preference to move into one world state over and above the other (Edwards, 1955). Studies of decision-making usually aim to predict these choices as the decision-maker moves through the cumulative stages of information-gathering, deliberation and then finalising their judgement (Wilson & Keil, 1999). Research in decision-making can go further than just exploring the pathway to the initial decision and can also study phenomena such as changes of mind after a decision has been reached (Resulaj et al., 2009), error monitoring across and within trials (Yeung & Summerfield, 2014), and, a variable with a long history, decision confidence (Metcalfe & Shimamura, 1994; Peirce & Jastrow, 1884).

Historically, several approaches or frameworks have been explored in decision-making. One that emerges from studies of economics is that of the rational decision-maker. This framework described the decision-maker as the "economic man" – a rational, informed, and infinitely sensitive individual (Edwards, 1955). This framework suggests that decision-makers should, or indeed do, operate in a predictable and grounded manner in direct relation to the information they

have about their environment and the options they can choose between.

This approach has been widely adopted in economic theories as the predictions it makes map onto the behaviour that one should expect to see play out in a functional economy and in economic decisions. However, the predictions made by this framework seem to deviate from the average decision-maker who may make biased or skewed choices that are not in line with this rational model (Kahneman & Tversky, 1974). Put simply, the average decision-maker often operates in a way that is far from the behaviour of the proposed "economic man". Myriad influences or reasons for the systematic deviations from this model can be postulated and have been demonstrated both experimentally and in real-world decision patterns. For example, these deviations from the assumptions of the rational model such as an assumed invariance of decision-making (Tversky & Kahneman, 1986) may be due to an under-emphasis on the modulating role of emotion in decision-making both cognitively and neurologically (De Martino et al., 2006).

Rational choice theory contrasts with the descriptive model of decision-making which, instead of focusing on *a priori* considerations of what a decision-maker *should* do, explores empirical observations using experimental methods to find the underlying psychological factors that relate to decision-making under uncertainty (Wilson & Keil, 1999). These ideas align with notions of bounded rationality which posits that decision-makers may make a rational choice within the limits of their cognitive capacity and world knowledge (Simon, 1990). Therefore, under these models, heuristics allow for what are often sensible decisions

in cases where people lack some of the information, or capacity to compute the infinite information bytes, necessary to make a purely rational decision in the sense of the "economic man".

From a psychological perspective, decisions have been captured in a number of ways. Classic approaches include those arising from signal detection theory (SDT) (Green & Swets, 1966). Under an SDT framework, in a two-choice task, the decision maker obtains an estimated observation of evidence that causes a response spike. This evidence estimate is a combination of signal and noise in a ratio that is unknown to the observer. This is then transformed into a decision variable and compared against some criterion value, such as the likelihood ratio of the two possible world states, to evaluate which world state the evidence emerged from (see Gold and Shadlen (2007) for an overview or Green and Swets for detailed derivations (1966)).

Sequential evidence accumulation models then act as the temporal generalisation of this framework which can accommodate multiple observed instances of evidence over time (Gold & Shadlen, 2007). Sequential accumulation models are valuable as they work to differentiate momentary evidence from an evolving decision variable. These models can be subdivided depending on whether the decision rule is relative and based on the evidence ratio between the two choices in a single accumulation process or an absolute value which is reached in one of two simultaneous accumulators (Smith & Ratcliff, 2004). Race models are those which accumulate evidence for the two possible responses in separate totals – the first total to reach the criterion determines the response (an absolute stop-

ping rule) (Vickers, 1970). Diffusion models, or random walks if the time points for sampling are discrete, by comparison, have a single decision variable total which moves between an upper and a lower bound as evidence for each choice is sampled. A response in a diffusion model is reached when the decision variable crosses one of these criterion bounds (a relative stopping rule) (Ratcliff & Smith, 2004).

Ultimately, the study of decision-making seeks to explore these different frameworks, to understand the decision process itself and to understand variables that might influence or bias that process. Much work has been done to bridge between these frameworks and ground the theory in neural mechanisms (Summerfield & Tsetsos, 2012). Another way to explore and differentiate the different models of decision-making is to investigate metacognitive indicators, such as decision confidence, that tell us about how information might be used in the decision-making process.

2.3 | Metacognition

Metacognition is the set of processes used to help monitor one's own cognition with the aim to guide present and future behaviours (Rhodes, 2019). The history of metacognitive inquiry, thinking about our thinking, spans back to the memory training of Greek philosophers who explored mnemonic tools and the use of memory loci to train an artificial memory (Yates, 2011). Within the domain of psychology, this concept was first expanded across the last century using the method of introspection, whereby participants under variable external conditions were asked to observe their consciousness and make verbal reports on their inner cognitive experience (Dunlosky & Metcalfe, 2008). This approach brought to light the complexity of metacognition as a field of scientific inquiry. However, this complexity was a double-edged sword. Soon the paradox of how an organ can be both the observer and the object being observed (Rhodes, 2019), often referred to as Comte's Paradox, (Nelson, 1996), became the stumbling block for the validity of introspection. In essence, the question of how the mind can validly reflect on itself seemed to undermine meaningful progress in the field of metacognitive inquiry, as the validity of the introspective reports was questioned (Dunlosky & Metcalfe, 2008).

The fruitful debate surrounding this paradox, (see Dunlosky & Metcalfe (2008) for a review), culminated in recent decades in extensive empirical work following the seminal review by Flavell in (1979). In this review, the contents and structure of metacognition and metacognitive processes were delineated

(Flavell, 1979). Importantly, in this review, Flavell separated out the notion of metacognitive knowledge from the concept of metacognitive experiences.

In this theory, metacognitive knowledge is defined as the sum total of a person's knowledge and beliefs about what factors interact to affect the outcome of cognition (Flavell, 1979). These factors may include knowledge about persons, tasks, or strategies. For example, person-based metacognitive knowledge could relate to an understanding of the need to attend to someone in order to comprehend them, it could acknowledge individual differences in the different social agents, and, ultimately, encompass everything that you know about yourself and other social agents as cognitive processors. Task-based metacognitive knowledge includes information about how successful you might be in achieving a certain goal or how demanding a task might be. Finally, the subcategory of metacognitive strategy knowledge is helpful in understanding how best to achieve particular goals. For example, memorising a list of chemical elements for an exam as compared with choosing the healthiest option for dinner, may require different cognitive approaches to succeed. This knowledge is not thought to vary in content or structure from any other form of knowledge or cognitive information and so requires no special mechanism for storage or retrieval (Flavell, 1979).

On the other hand, metacognitive experiences can vary in length and may occur before during or after a cognitive enterprise (Flavell, 1979). For example, you might suddenly feel a maths test is going very well, or become puzzled about how something was phrased by your colleague after your conversation is over.

Metacognitive experiences could be relevant pieces of metacognitive knowledge that have been brought into conscious thought, or they could be thoughts or feelings about one's own thinking such as feelings of uncertainty, puzzlement, or confidence about a given cognitive task. Metacognitive experiences can shape or update metacognitive knowledge and can guide strategies or update goals, ultimately affecting behaviour.

This distinction between metacognitive knowledge and metacognitive experiences, in turn, led to the development of a more sophisticated model of metacognition (Rhodes, 2019). This model assumes a multilevel structure to cognition with object and meta-levels of cognition being connected by directed flows of information in the form of monitoring and control (Nelson & Narens, 1990), see Fig.2.1.

The object level concerns itself with objects external to cognition, while the meta-level concerns itself with cognition about the cognition of those external objects (Nelson, 1996). Straightforwardly, object-level processes are the basic mechanisms of cognition such as perception, memory, and decision-making. The meta-level then monitors the processes and/or outputs of these object-level cognitions and exerts appropriate control. For example, the object level of cognition could handle the processing of a maths equation to be solved in real-time, while the meta-level might hold information about your confidence in maths as a subject or your feelings of uncertainty about algebra or long division calculations.

Monitoring is then the process by which the meta-level is informed by in-

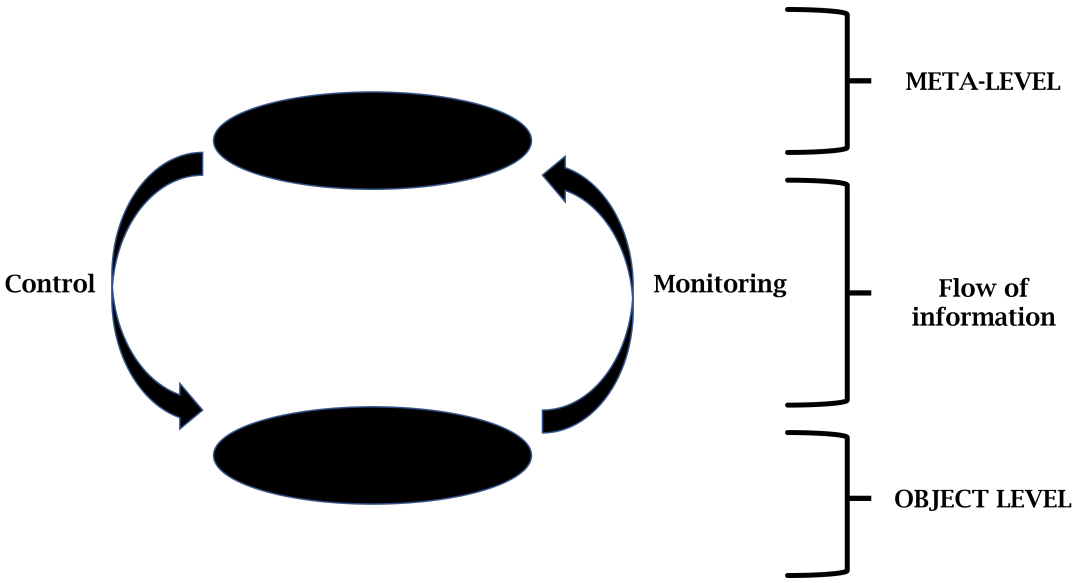


Figure 2.1: Figure adapted from Nelson & Narens (1990). The model outlines a mechanism that comprises two structures and the relationship between them. The object level is monitored by the meta-level and in turn, updates the meta-level.

formation at the object level. For example, the completion of a maths sum could bring certain pieces of meta-knowledge into conscious thought such as different strategies to simplify the task at hand or past experience with completing this type of problem successfully. Meanwhile, control is the directed process by which the meta-level modifies the object level. Your confidence or anxiety might influence how quickly or effectively you tackle the equation in front of you, for example. In this way, the paradox described by Comte of the mind observing itself, (Nelson, 1996), is then resolved if we consider that this process occurs simultaneously across a hierarchy of cognitive levels, rather than within a single unit or level.

Investigations of metacognition often focus on the accuracy of metacognit-

ive monitoring and, in turn, on the influence of metacognitive control on current and future behaviour. For example, studies of monitoring have implicated working memory capacity and other executive functions in the accuracy of these processes, with higher working memory capacity relating to higher monitoring accuracy (Komori, 2016). In terms of metacognitive control, this has been described in relation to stopping rules in cognition – when an agent should give up on a problem or challenging task (Payne & Duggan, 2011) – based on judgements of solveability (Ackerman & Thompson, 2017). However, one vital, if relatively understudied, question is when and how often human decision-makers engage in any kind of metacognitive processing. This is an important question as, if metacognitive experiences are relatively rare, their possible contribution to a wider understanding of human cognition remains limited. By looking at a combination of work from mind-wandering, thinking-aloud protocols, and spontaneous learning strategies, Jordano & Touron (2018) suggest that the frequency of metacognitive experiences varies across individuals, the life-span, and in different settings. For example, while trying to validate a questionnaire against the think-aloud method, it was found that among high-schoolers completing a think-aloud protocol, 30.6% of the recorded activities were metacognitive in nature, (Schellings et al., 2013). By comparison, in fifth-graders conducting a reading comprehension task as part of a think-aloud task, between 0-27% of the verbalisations that were made were metacognitive (Meyers et al., 1990). Though of course, the quality of these instances of metacognitive monitoring may be just as, if not more, important than the number of recorded instances (Jordano &

Touron, 2018). It seems, based on these studies at least, that when cognition is oriented toward a task metacognitive thoughts appear with some regularity in the cognitive landscapes of participants, an important piece of evidence when deciding if metacognitive processes are a meaningful addition to the study of broader cognitive processes like decision-making.

2.4 | Decision Confidence

When zooming into metacognition in more detail, metacognitive confidence is then a particular type of metacognitive experience. Sometimes described as a feeling of knowing, (Navajas et al., 2017), confidence explicitly reports on the reliability of internal cognitive operations (Aitchison et al., 2015). Having an estimate of this reliability is important to help guide behaviour and future behaviour (Persaud et al., 2007). Understanding the mechanism used to report this reliability may help understand failures of decision-making in sectors such as healthcare (Jackson & Kleitman, 2014; McKenzie, 1998) and finance (Atir et al., 2015), as well as in everyday decisions (Brenner et al., 1996).

More precisely, decision confidence can be described as a subjective estimate of choice accuracy (Adler & Ma, 2018), which arises from metacognitive monitoring of the decision process (Sanders et al., 2016). As described by SDT-based diffusion models (Gold & Shadlen, 2007; Smith & Ratcliff, 2004), effective decision-making requires the weighing-up of evidence for certain choices; including integrating ideas about the sensory reliability of the measurements (Komura et al., 2013), and, when exploring category spaces, a representation of the category space from which stimuli are drawn. These decisions are often made on imperfect evidence (Tversky & Kahneman, 1974) and the quality of the evidence must be quantified by the decision-maker to assess the confidence with which they should proceed (Kepecs et al., 2008).

A key determinant of confidence is the strength of accumulated evidence in

the decision process, as reflected in the balance of evidence modelled in evidence accumulation models or distance-to-criterion approaches or in SDT-style models (Kepecs & Mainen, 2012). This means that, in a binary choice task, confidence reflects the difference in the strength of evidence accumulated in favour of each of the two options with corresponding signatures observed in, for example, speed-accuracy trade-offs. Indeed, when accuracy is emphasised, a longer time is taken to accumulate the evidence difference for the decision, and confidence tends to be correspondingly larger (Vickers & Packer, 1982).

However, although most current theories agree that confidence depends critically on the strength of accumulated evidence during the decision (e.g., Pleskac & Busemeyer (2010), Fleming and Daw (2017)), many other factors also seem to affect confidence, including decision-relevant ones like expectations or quality of evidence (Bang & Fleming, 2018; Navajas et al., 2017; Olawole-Scott & Yon, 2023), but also less obviously relevant ones like mood (Massoni, 2014), personality (Pallier et al., 2002), and psychopathology (Rouault et al., 2018).

Thus, while observers reliably report their subjective feelings of confidence, such that their confidence reports correlate with their objective accuracy (Nelson & Narens, 1990; Peirce & Jastrow, 1884), the underlying computational models of these judgements remain somewhat debated, in particular when discussed relative to rational or Bayesian models of cognition (Adler & Ma, 2018; Xue et al., 2023).

2.4.1 | Measuring Decision Confidence

Metacognitive judgements, including retrospective confidence judgements, are a form of explicit monitoring of the understanding held of the environment or task being experienced. Decision confidence reports are most commonly elicited by self-report measures. In a way, this mirrors the original introspective techniques which asked individuals to observe their own thinking. These reports are explicit statements of the judgements that are communicated intentionally (Overgaard & Sandberg, 2012). However, to quantify these observations or reports, particularly to compare across individuals who may have a different way of communicating their observations, confidence is required to be reported in a standardised manner.

Measuring confidence, therefore, often includes the use of Likert or continuous scales that attempt to quantify the report in a way that is comparable across individuals. These scales might consist of verbal or numerical bins ranging from low to high confidence. This might include a range of confidence statements that act as verbal labels for the confidence reporting bins like "very unsure" or "very high confidence". The most straightforward labels for confidence reports are simple to ask participants to label trials as a "guess" or as "known" in a binary manner (Dienes, 2007). Indeed, a range of scale characteristics have been explored to try to understand if categorical or continuous, more or fewer bins, and verbal versus numeric scale creation were measuring meaningfully different metacognitive reports. For example, a simple binary scale of low or high

confidence has been found to be more sensitive for reports of low-level awareness as compared with a numerical scale (Tunney, 2005; Tunney & Shanks, 2003). However, the reverse of this pattern was found by Dienes and colleagues (2007). This more extensive work compared multiple scale formations (six in total) and found no meaningful differences in the way that the confidence expressed on these scales related to the accuracy of participants' choices (Dienes, 2007). Within their own judgements, across different formations of the confidence scale, participants were reasonably consistent in their scale usage relative to their accuracy.

Another way approach to measuring confidence includes asking participants to gamble or wager on the accuracy of their decision, whereby a higher wager is thought to relate to higher confidence. This was originally developed for situations where the participants may not be able to understand a verbal or numerical scale such as when working with non-human primates (Kiani & Shadlen, 2009). These post-decision wagers are favoured by some (Persaud et al., 2007) but the functional differences between these wagers and classic confidence reports are disputed (Dienes & Seth, 2010).

In this thesis, confidence reports are elicited using continuous scales that are book-ended by verbal labels of confidence levels. For example, in Chapter 3, confidence was prompted on a continuous semi-circular sliding scale ranging from "Definitely correct" through "Don't know" to "Definitely wrong". In the other experimental chapters, a continuous horizontal sliding scale that ranged from "totally UNSURE" to "sure CORRECT" was used. Therefore, while the

exact wording of these labels and the visualisation of the scale itself varies across individual experimental paradigms, the broader method is held constant throughout all studies. Within each experimental chapter, the exact wording and scale will be delineated for clarity.

2.4.2 | Metacognitive Sensitivity & Bias

The experiments in this thesis primarily study the impact of experimental manipulations on participants' confidence reports, both aggregated and on individual trials. In the wider literature, other metrics of confidence are often studied, in particular measures of metacognitive sensitivity and bias that quantify the relationship between subjective confidence and objective accuracy. The former is concerned with the individual calibration of confidence, while the latter relates to the reliability or accuracy of metacognitive reports (Fleming & Lau, 2014).

Metacognitive sensitivity refers to the level of calibration that exists within individuals between their objective accuracy and their feeling of certainty. In this way, metacognitive sensitivity is what makes confidence reports meaningful and, to an extent, worth measuring (Fleming & Lau, 2014). This calibration exists independent of the type of measurement used to capture confidence reports, discussed above. Because, regardless of which measurement is used, while confidence tends to correlate positively with the accuracy of a decision, this process is variable and prone to error (Nelson & Narens, 1990). Early investigations suggested that although confidence correlates strongly with decision time, such

that faster decisions are associated with higher confident reports, confidence was an unreliable indicator of accuracy (Henmon, 1911). The metacognitive sensitivity of confidence reports links strongly to the type of measurement used in a decision-making task.

Metacognitive bias can then be construed as the likelihood of reporting high or low confidence irrespective of decision performance, which may relate to individual differences such as psychiatric alterations to cognitive processing(Rouault et al., 2018). In general, overconfidence is the most common form of metacognitive bias – individuals are usually more confident than they should be given the accuracy of their performance (Atir et al., 2015; Brenner et al., 1996). While this pattern is almost ubiquitous across individuals, it does not hold true for all trial difficulties and is often eliminated in easy tasks – a phenomenon known as the hard-easy effect (Lichtenstein & Fischhoff, 1977).

In the absence of these biases, or by accounting for them explicitly, metacognitive sensitivity can be measured using criteria based on the relationship between the accuracy of the choices made and the confidence reported in those decisions. Therefore, different ways of calculating how well confidence reports are calibrated within individuals have emerged in the literature such as meta-d' or the use of receiver operating characteristics (ROC) (Fleming & Lau, 2014).

Meta-d' utilises the framework from signal detection theory to estimate the hit and false alarm rates of individuals, but instead of assessing the first-order decision patterns it focuses on the second-order patterns in terms of their confidence ratings conditioned on the accuracy of observer's responses. It accounts for

the number of high and low confidence ratings assigned to correct and incorrect judgements. In the case of high confidence, hits would then be high confidence reports correct trials, and false alarms would be high confidence errors. In this way, meta-d' quantifies the degree to which participants' confidence is predictive of accuracy, with above-chance metacognitive sensitivity found when correct trials elicit high confidence to a greater degree than incorrect trials (Fleming & Lau, 2014). However, the assumptions about the underlying distributions of these rates have proved incoherent for the second-order judgements of confidence (Galvin et al., 2003).

By comparison, the second-order area under receiver operating curves, aka AUROC2, can be used as a non-parametric bias-free measure of sensitivity (Fleming & Lau, 2014). These make use of the fact that with multiple confidence ratings, it is possible to treat each confidence level as a criterion bound that separates high from low confidence. Then, for each criterion, the second-order hit rate and false alarm rate can be calculated and plotted to obtain a second-order ROC curve. The AUROC2 can then be used as a measure of metacognitive sensitivity with higher values indicating higher sensitivity. Importantly, recent work suggests that, across 17 variant measures of metacognitive ability, all the measures assessed were valid and achieved reasonable and similar levels of precision (Rahnev, 2023).

In this thesis, metacognitive sensitivity and bias are not a focus of study in themselves, but considerations around these measures inform the experimental designs. For example, attempts to reduce metacognitive bias are established as

part of the methodology of confidence reporting (with the scale toggles starting at a random position on each trial to prevent systematic anchoring effects from emerging, for example). Meanwhile, across analyses, confidence reports are also pooled within each individual and in some analyses the values of confidence are binned within participants to allow for more meaningful comparisons across participants, noting previous reports of high inter-individual variability in metacognitive sensitivity and bias that potentially add noise in statistical comparisons (Ais et al., 2016; Pallier et al., 2002; Rouault et al., 2018).

2.4.3 | Uses of Decision Confidence

To the degree that they reliably track objective accuracy, decision confidence and the associated feelings of certainty or uncertainty can provide useful information to the decision-maker. They may help to indicate what actions should be taken next and whether corrective actions need to be taken after the fact. In terms of the multi-level model of metacognition, this relates to the element of "control" exerted by the meta-level over the object level processes (Nelson & Narens, 1990).

For example, confidence judgements have been shown to relate to information seeking, with reduced confidence, or increased uncertainty, relating strongly to increased information seeking in behavioural tasks (Desender et al., 2018). This link has also been established in terms of a predictive post-decision neural correlate of uncertainty which links to information seeking (Desender et al., 2019). The relationship between information seeking and confidence is thought

to be modulated by metacognitive sensitivity, with those with higher metacognitive sensitivity seeking out a second opinion on a medical decision-making task, see for example Clayton *et al* (2022). Confidence is also useful when trying to understand or predict advice-seeking (Carlebach & Yeung, 2020). This idea is also relevant to the role of confidence in information sharing and advice-taking when looking at the variable weighting of evidence in group decisions (Bahrami et al., 2010; Bang et al., 2014).

Other proposed uses of confidence include choosing tasks (Carlebach & Yeung, 2023) and deciding whether to stick to a choice when a decision is repeated (Folke et al., 2016). Furthermore, in multi-stage decisions, confidence is carried over and used to adjust the decision threshold of subsequent stages (van den Berg, Zylberberg et al., 2016). Indeed, it has been suggested that confidence acts as a "common currency" to allow different types of decisions to be compared and prioritized, for example in tasks in different modalities (de Gardelle et al., 2016; de Gardelle & Mamassian, 2014).

Feelings of certainty can also influence learning and a learner's decision to remain engaged in the active learning process. For example, when judgements of learning (JOLs) were manipulated experimentally, lower JOLs predicted the learner to be more likely to want to revise a word-pair again during a recall task (Metcalfe & Finn, 2008). These JOLs are particularly effective after a short delay in predicting future success in recalling relevant items (Dunlosky & Nelson, 1992).

We also know that confidence can be helpful in problem-solving more gen-

erally as part of a broader meta-reasoning framework (Ackerman & Thompson, 2017). A problem, described in its broadest sense, encompasses a set of information, a particular goal, and one or more obstacles that stand in the way of achieving that goal. Metacognition is intrinsically related to the processes of problem-solving as described by Davidson *et al.*, (1994). The steps of identifying the problem, representing the problem space, planning behaviour, and evaluating your performance (Davidson et al., 1994), all have clear metacognitive elements. Confidence can be directly linked to performance evaluation, but can also be considered important in terms of mentally representing the problem space and updating that mental representation by seeking new information or attempting to combine the information in novel ways, for example.

While it is clear that the question of how metacognitive control is exerted based on confidence reports is an important one, it is not the focus of this thesis. Rather than exploring how current or past confidence reports signal the observer that their decision-making process or task approach might need to be changed, this thesis explores what information current confidence reports reflect in terms of category distribution representation and how that informs both category learning and confidence judgements themselves. However, that is not to suggest that these ideas about metacognitive control are irrelevant to the study of confidence in categorisation judgements – they merely do not stand as the primary focus of this work.

2.4.4 | Different Types of Uncertainty

When discussing confidence, and its uses, it is important to acknowledge that different kinds of uncertainty exist that may be captured by a decision confidence report. The information we use when making a decision is rarely perfect or complete. Therefore, decisions are rarely made with absolute certainty. Some scholars have defined this uncertainty as confidence itself (Iyer et al., 2022), others focus on the process of decision-making under uncertainty whereby the observer lacks knowledge regarding which outcome is connected to which choice (Platt & Huettel, 2008).

One way to explore uncertainty would be to differentiate between "informational conflict" and "informational ambiguity" (Iyer et al., 2022). Informational conflict is the uncertainty that arises when the evidence is conflicting between choices, while informational ambiguity arises from incomplete or unclear information regarding choice outcomes. Informational conflict can be further disentangled from cognitive or decision conflict which arises when multiple similar outcomes must be selected between, (Pushkarskaya et al., 2015). These differences have meaningful effects on judgements and behaviours with insurers setting higher premiums for situations with information conflict than informational ambiguity (Cabantous et al., 2011), for example. However, by contrast, individuals appear to prefer to place wagers based on conflicting rather than ambiguous information (Pushkarskaya et al., 2015).

Furthermore, beyond the information or precision of knowledge held about

the outcomes, decisions may also vary in terms of their *certainty* of outcomes. For example, the availability of the chosen option is fairly certain in a restaurant setting when choosing between different pizza toppings – the outcome of ordering a pepperoni pizza is fairly certain. However, when choosing numbers for the lottery, or selecting a "winning" horse at the races there is an element of risk involved as the outcome is far from certain, and by definition cannot be predicted with absolute certainty. Here it is perhaps valuable to also define the concept of "risk" as separate from ambiguity in decision-making. The former refers to the outcome being probabilistic, such as in a coin toss or lottery draw where a precise probability can be calculated as the outcome space is discrete and known. The latter relates to the probabilities of the outcome space being unknown or unclear such as in a horse race where the weather, noise of the stands, the jockey's breakfast choices, or even the colour of an opponent's silks may have some immeasurably small effect on the outcome. While individuals appear to prefer risk to ambiguity on average (Platt & Huettel, 2008), patterns of risk aversion are domain-specific (Weber et al., 2002) and highly culturally variable (Bontempo et al., 1997) meaning that this preference is not universal.

These ideas speak to the complex and multifaceted nature of uncertainty. This complexity might bear out in confidence reports on categorisation judgements which in themselves have a scope for rich and complex structures. Therefore, exploring these spaces through the lens of confidence, as an expression of certainty, is a useful venture. Investigating experimental equivalents to the dermatologist grappling with situations where benign and malignant growths

could look identical, where they might differ in their distributional properties like variance, or might differ in terms of which dimensions are most relevant, leverages that richness of category judgements to explore the complexities of uncertainty and confidence.

2.5 | Modelling Decision Confidence

Just as decision-making has been modelled in a number of ways, the computation of decision confidence has also been represented in a number of different models. Some of these models are outlined and discussed below. As with this literature review more broadly, this section does not aim to be exhaustive in listing all possible variations of confidence models. Instead, different families of relevant models will be introduced and, in some cases, compared, in order to illustrate the landscape of thought that surrounds confidence modelling. Importantly, these discussions are prefaced by the acknowledgement that "all models are wrong, but some are useful" (Box & Draper, 1987, pg. 424). In particular, when looking to explore specific aspects of metacognitive confidence, certain models may lend themselves more readily to such discussions than others.

Furthermore, in some situations, it can be the lack of fit of a model that is most informative as it can highlight a particular point at which behaviour deviates from the theorised computations. Here the question of whether a model is accurate becomes less valuable, and instead, one might ask "How wrong do they have to be to not be useful?" (Box & Draper, 1987, pg.74). The perspective of Box is again pertinent when comparing models with behavioural data and exploring the emergent deviations - in those moments it is "inappropriate to be concerned about mice when there are tigers abroad" (Box, 1976, pg.792). If a model fundamentally fails to capture large-scale patterns, then smaller effects should not be the focus of model adjustment. With these thoughts in mind,

single and dual process models, and ideal observer models of decision confidence will be introduced in the following sections.

2.5.1 | Single process models of confidence

Single process models of decision confidence suggest that the decision process and the judgement of confidence are intertwined, in the sense that the same evidence informs both the decision itself and ratings of confidence in that decision. Models of confidence that utilise a signal detection framework imply a single process model with a "decision locus" for confidence (Baranski & Petrusic, 2001).

As reviewed by Balakrishnan and Ratcliff, (1996), classic models of signal detection assume that a stimulus, when presented, will create an information state in the perceptual system known as a strength value. This strength value is a uniquely identifiable state that corresponds to the signal strength of that stimulus. It represents the (noisy) perceptual measure of the stimulus on the decision-relevant dimension such as line length or luminance. The observer must then determine which stimuli are likely to cause which strength value, estimating the likelihood of each option. Over time, it is assumed that the observer creates a mapping from strength values to responses, though how that decision rule is actually chosen remains unexplained by these models (Balakrishnan & Ratcliff, 1996). Confidence, then, is mapped from the decision strength value and read off. In other words, it is a mapping or scaling of the distance between the stimulus observation and the decision criterion as defined by the

strength value (Baranski & Petrusic, 2001). Therefore, signal detection models have become synonymous with distance to the decision bound models of decision confidence which articulate that confidence is a "read off" of the distance of the stimulus measurement from the decision criterion (Kepecs et al., 2008; Komura et al., 2013). These distance-to-bound models predict that confidence increases with distance from the decision boundary, with the lowest confidence reports lying closest to the decision boundary where uncertainty is at its highest (Kepecs et al., 2008; Komura et al., 2013), see Fig. 2.2 for a visualisation of this model. Single process models make specific predictions – that the decision process should be unaffected by the requirement to make confidence judgements and that the reporting time for confidence should be a relatively fast constant which is unaffected by the trial difficulty or confidence read out (Baranski & Petrusic, 2001).

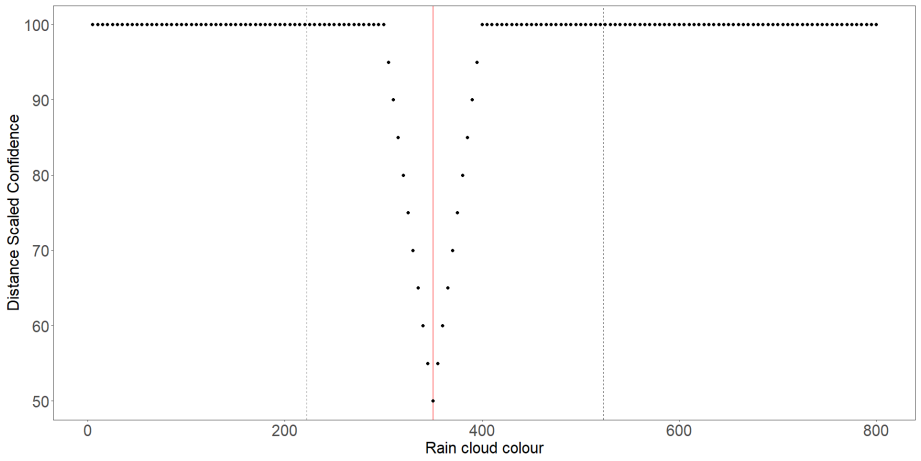


Figure 2.2: Simulated distance-to-bound model of confidence. Confidence is plotted on the y-axis and stimulus value – here, colour of rain clouds – is on the x-axis. The confidence values rise, here in a linear manner, from the decision boundary marked by the red vertical line. In this space, the decision bound stands at the point of equi-probability between the two category means, denoted by the vertical dashed grey lines.

The theory of probabilistic mental models (PMMs) is another example of a single process model of decision confidence (Baranski & Petrusic, 2001). PMMs were developed to address two clear confidence effects that were previously mentioned, (Gigerenzer et al., 1992); the overconfidence effect whereby confidence reports are higher than the frequency of correct answers for a given question (Brenner et al., 1996), and the hard-easy effect where the degree of over-confidence increases with the difficulty of the question (Lichtenstein & Fischhoff, 1977). These models lean on the Brunswikian notion of the natural environment and the reference classes of that environment that are known to the observer. They posit that the decision process is based on the activation of contextual relevant cues (Gigerenzer et al., 1992). For example, if you are asked whether coffee is

more expensive in one chain compared with the other you might recall the level of luxury in the standard decor of each in order to gauge their price points. Confidence then relates to the assessed validity of this cue – the hard-easy and over-confidence effects then emerge from an overestimation of the validity of these cues relative to the decision process. Overconfidence is thought to emerge when the possibility of category samples for which the cues are less valid is not accounted for – e.g., cafes with less luxurious interiors but locations in expensive neighbourhoods, or worn-out cafes that hold a monopoly in an amusement park, or purposefully shabby outlets with spartan interior design choices to keep the focus on the coffee tasting notes (Brenner et al., 1996). The hard-easy effect falls naturally from this, with those examples with higher difficulty having the greatest mismatch between the estimated validity of the cue and the accuracy of the classification (Lichtenstein & Fischhoff, 1977). This is deemed to be a single process model of confidence because the confidence and decision are still both judged from the same activated contextual cue (Baranski & Petrusic, 2001).

2.5.2 | Dual process models of confidence

Dual process models, in contrast to the single process framework, suggest that decisions and confidence judgements arise from at least partially separable processes, and may even utilise different sources of information. A large family of such models, sequential sampling models (Baranski & Petrusic, 1998), suggest that decisions are formed as part of an information accumulation mechanism and confidence is judged based on a *post hoc* evaluation of that accumulated

information. For example, confidence judgements might be drawn from the calculation of the proportion of evidence in favour of the decision (Juslin & Olsson, 1997). These are akin to the diffusion or accumulator-based decision-making models. Due to the post-decisional nature of these judgements, the requirement for confidence reports should not affect the decision process but, in contrast with single process models, confidence reporting may take time and the time taken may vary across trials (Baranski & Petrusic, 2001).

Based on these ideas, and in order to verify the separability of these processes, the locus of confidence judgements was probed by Baranski & Petrusic (1998). They found that the different predictions of single and dual process models played out depending on whether the task was speed or accuracy-pressured. In other words, how confidence was calculated or, more accurately, where confidence reports were drawn from, varied depending on the task demands. When accuracy was stressed, the computation of confidence reports began during the decision-processing phase; when the task was speed stressed, all decision-processing resources were devoted to the decision itself and the confidence computation appeared to occur entirely post-decision. These different task demands had effects on error monitoring with a post-decision locus for confidence resulting in improved error detection (Baranski & Petrusic, 1998). Furthermore, it has been shown that the task demand of requiring a confidence report in itself may affect the confidence processing mechanism (Baranski & Petrusic, 2001). This was seen in an increase in decision time when a confidence report was required, suggesting that some processing for confidence occurred

pre-decision and was completed post-decisionally (Baranski & Petrusic, 2001).

One classic dual process model of confidence is the Two-Stage Dynamic Signal Detection (2DSD) model from Pleskac & Busemeyer (2010). This model leverages a drift-diffusion process to account for the response and decision time, the evidence accumulation post-decision is then thought to relate to confidence: people are more confident to the degree that continued processing after the decision leads to a strengthening of that choice, and tend to experience low confidence (or even changes of mind) if evidence accumulated in post-decision processing goes against the initial choice. The post-decision accumulation is interrupted when a confidence report is collected and the evidence is categorised into a suitable confidence judgement. Evidence for these types of models emerges in a number of ways.

For example, there is evidence that post-decisional evidence accumulation influences confidence. Indeed, an asymmetry has been found between post-decision confidence on correct and error trials. There is a deceleration of accumulated evidence consistent with the current choice and a contrasting acceleration of accumulated evidence toward the opposite direction relative to the current choice in error trials, with consequences for confidence reports (Yu et al., 2015). This links to ideas of error monitoring which necessarily relies on post-decision processing (Yeung & Summerfield, 2012, 2014). Indeed, it has also been shown the confidence reports and the detection of error both rely on the balance of stimulus evidence accumulated in the periods both before and after the decision point (Charles & Yeung, 2019). Similarly, by probing confidence

and decision using a single mechanism, it has been shown that changes of mind about an initial response and confidence in that choice are both informed by the same evidence (van den Berg, Anandalingam et al., 2016). While some of these findings suggest that the calibration of post-decision confidence reports is improved by the extended evidence accumulation, biases in post-decision evidence accumulation may also result in biased confidence reports. For example, an increased rate of accumulation for confirmatory post-decisional evidence could lead to overconfidence (Navajas et al., 2016).

2.5.3 | Ideal observer models of confidence

Much like the economic and rational man framework in decision-making, the ideal observer model of decision confidence suggests that the observer makes optimal use of the information available to them when given a statistical description of a task; integrating all available evidence to make an accurate confidence judgement, (Adler & Pointon, 1993). The ideal observer model can, therefore, be described as the computational theory for a task (Marr, 1982) and so is abstracted away from the implementations themselves. The Bayesian framework is a popular conceptualisation of this model (Beierholm, 2013; O'Reilly et al., 2012). It describes the brain as a dynamic evidence-processing machine, combining knowledge about the statistical structure of the world with the current input, to compute a posterior probability distribution. This distribution describes the probability of different world-states by integrating new evidence with the prior knowledge held by the observer. This enables optimal behavi-

oural responses by the observer (Knill & Richards, 1996; Pouget et al., 2016; Sanders et al., 2016).

Both neural firing patterns, (Lak et al., 2017), and sensorimotor integration seem to adhere to these optimal observer principles, (Ernst & Banks, 2002; Körding & Wolpert, 2004). For example, when participants are asked to make judgements about the size or shape of a stimulus while exploring it both visually and using touch (haptic cues), they seem to combine these sources of information in a near-optimal way (Ernst & Banks, 2002): To minimise variance in the final estimate, in accordance with the maximum likelihood estimate, the information from the different senses of touch and vision is weighted according to their variance. This means that less reliable sources of evidence, with high variance or noise, are down-weighted such that they do not affect the decision as strongly. Therefore, if the visual input is perturbed, evidence from the visual input is down-weighted, despite it being the more dominant sense (Körding & Wolpert, 2004). To solve such integration problems cognition may also lean on prior beliefs to help with object recognition and processing (Kersten et al., 2004).

Given these explorations of Bayesian aspects of cognition, it is not surprising that ideal observer models of metacognitive confidence are also popular (Kepecs & Mainen, 2012; Pouget et al., 2016) – with the confidence reports being assumed to be related to $P(Correct)$ in a given choice. The fact that the idea of "confidence" itself is meaningful to decision-makers (Peirce & Jastrow, 1884), however, suggests that these assumptions of Bayesian inference are in fact just a testable framework (Adler & Ma, 2018). While some recent evidence suggests

that the prior expectations held about the precision of evidence that will be encountered influence metacognitive reporting – lending credence to this "rational" or Bayesian framing of confidence (Olawole-Scott & Yon, 2023) – some of the proposed signatures of Bayesian confidence (Hangya et al., 2016) have not been supported quantitatively (Adler & Ma, 2018).

Bayesian inference emerges from a discussion of conditional probability. When thinking about conditional probability we often describe scenarios where we ask about the probability of obtaining a certain set of data given a particular model of the world. For example, we might think about the probability of a coin landing on heads given what we know about coin flips. We can annotate this as $P(data|model)$, conventionally denoted as $P(y|\theta)$, or, indeed, $P(heads|coinToss)$. However, when we think about understanding cognition and metacognition, we are often more interested in understanding the probability of a certain model given the data, $P(\theta|y)$, reversing the classic direction of conditionality. Reversing this directionality is not as straightforward as subtracting the original probability from 1, as we might initially expect given that probability occupies a $0-1$ space. Thinking about it in real terms, we can take the simple example that the probability of my having my Christmas decorations up, given that it is Christmas Day, is not the same as the probability of it being Christmas Day given that I have decorated my Christmas tree. The theorem required to achieve this reversal, which therefore allows for the calculation of $P(\theta|y)$, was devised by Reverend Thomas Bayes (1702-1761) and is outlined below.

Taking first one of the laws of probability which states that the probability of two outcomes occurring, $P(a, b)$, when the outcomes are not independent is:

$$P(a, b) = P(a|b) * P(b) \tag{2.1}$$

The same is true for the opposite direction of conditionality:

$$P(b, a) = P(b|a) * P(a) \tag{2.2}$$

Given that $P(a, b)$ and $P(b, a)$ are equitable we can establish the following:

$$P(b|a) * P(a) = P(a|b) * P(b) \tag{2.3}$$

Dividing both sides by $P(a)$ leaves:

$$P(b|a) = \frac{P(a|b) * P(b)}{P(a)} \tag{2.4}$$

This is Bayes's theorem which, in more conventional terms, is ordinarily expressed as:

$$P(\theta|y) = \frac{P(y|\theta) * P(\theta)}{P(y)} \tag{2.5}$$

This equation establishes the assumed mechanics of the computations behind the ideal observer models described above. What follows is a deep dive into the terms of the theorem derived above (for a more instructional and hands-on

exploration of these principles see Farrell & Lewandowsky (2018)).

The *prior* is arguably the most intuitive of the terms to be discussed. It is denoted in Equation 2.5 by the term $P(\theta)$. The prior is formally described as the knowledge held about the world before an experience or event. In modelling terms, it is the information or beliefs held about the model and its parameters prior to the experiment or the data collection (Farrell & Lewandowsky, 2018). It is a probability distribution across a set of possible values. This term is intuitive when we consider that the Bayesian framework emerges from the formalisation of Helmholtz's ideas of perception as an unconscious inferential process (Kersten et al., 2004). Helmholtz articulated the ambiguity of retinal images and emphasised the importance of prior knowledge to disambiguate our percepts (Helmholtz, 1867). Indeed, identical retinal images could arise from different objects viewed from different angles (Kersten et al., 2004). For example, shape perception can depend on viewpoint and the angle of illumination (Mamassian & Goutcher, 2001). To accurately perceive and recognise objects, the retinal image is combined with priors regarding the statistical likelihood of a retinal image equating to a particular object's reality. A classic "popular science" example of this phenomenon is that of scientists initially misidentifying the moon's craters as mountains, based on noisy imaging. Depth perception is aided by a prior for illumination coming from above which makes surfaces look concave or convex under different conditions (Mamassian & Goutcher, 2001). In this way, the moon's craters might look like mountains under the right (noisy) optical conditions.

Whether such priors in cognition or perception are innate or learned has been much discussed. Early experiments showed that chickens raised in an environment lit from below still behaved as if the light was coming from above suggesting there is an "innate" perceptual parameter to cause the interpretation of retinal images in this way (Hershberger, 1971). It appears, however, that this prior develops as a result of the experience of illumination relative to the position of the head (Howard et al., 1990). More specifically there is evidence that visual perception precedes based on the assumption that light is coming from the above and left of an object (Mamassian & Goutcher, 2001). Further evidence of the development of perceptual priors comes from the study of audition, (Bao, 2015), whereby infants develop priors for prototypical speech sounds from their native language environment (Kuhl, 1991) from between 6-12 months (Kuhl et al., 2006). These priors, which emerge through our developing cognition based on experience and repeated interaction with the world (Jazayeri & Shadlen, 2010), lend support for the Bayesian framework, as well as provide clear examples of cognitive priors that might influence cognition more broadly.

In terms of metacognition, there is also clear evidence that confidence is affected by prior experience, for example, as shown by the formation of "global" self-performance estimates from the aggregation of local decision confidence reports (Rouault et al., 2019). Here a distinction can be made in terms of the interpretation of prior in terms of the first-order decision and second-order confidence. In the case of the former, the prior can be described as the likelihood of a particular decision being correct or a particular stimulus being present on a

given trial. In the second order, there is the probability of responses in general being correct.

The *likelihood* is the second of the terms to be discussed from Equation 2.5. It is denoted by the term $P(y|\theta)$ and describes the probability of obtaining a certain dataset, or for example recognising a particular object, given what we know about the world or model (given the prior: $P(\theta)$). Importantly, the likelihood considers the values of $P(\theta)$ to be fixed and held constant and the data or outcome to be variable. This differentiates the likelihood term in Bayes theorem from the likelihood expressed in maximum likelihood estimates ($L(\theta|y)$) which is not an expression of probability (and is not normalised between $0-1$) and which assumes the data is fixed but the parameters of a model to be variable. This distinction will be explored more in subsequent sections on the modelling process. In terms of confidence, the likelihood might be described as the way in which prior knowledge exerts influence over a confidence report, for example in terms of expertise linking to overconfidence and reduced tendency to seek further information (Atir et al., 2015; Friedman et al., 2005).

The *evidence* is the denominator of Equation 2.5 and it acts as a normalising factor which scales the posterior between $0-1$. It is the probability of the data irrespective of the prior. It is also known as the marginal likelihood, referencing the marginalisation of probabilities whereby the overall probability of an outcome is summed across all possible contexts. For example, if we look at the probability of students identifying as a certain gender across different undergraduate courses, the marginal probability of being a female student is the

summation of the self-identifying females across all courses divided by the total number of students. In modelling terms, this might equate to the probability of the data across all possible values of the model parameters. In the context of confidence, this might relate to the different scaling of confidence relative to the number of possible choices, for example. In light of this, we could expand out the term $P(y)$ as follows:

$$\sum_{\theta} P(y|\theta) * P(\theta) \tag{2.6}$$

This allows Bayes theorem to be re-described in a particularly accessible way; as the fraction formed by the probability of a *particular* outcome observed in this instance, over the sum total of all *possible* outcomes that might have emerged based on our prior knowledge of the world (Farrell & Lewandowsky, 2018).

Finally, the *posterior* is the last term to be described, $P(\theta|y)$. This is the probability of a certain model or parameter values given the data. It describes the outcome of the process of Bayes theorem whereby the prior knowledge we hold is combined with the current experience or data to update our understanding of the world.

The relevance of this explanatory introduction will become evident in the following chapters where these terms and the overarching framing are leveraged in computational models of confidence.

2.5.4 | Comparing different models of confidence

Given the various models that could be used to describe confidence, it is necessary to think about how one can differentiate between each model and understand what observers actually do when making metacognitive judgements. This requires model comparison studies to be conducted in parallel with advancing behavioural studies. In essence, this means gathering a set of data and seeing which model best approximates what people were doing when making choices in decision-making paradigms. In such an exercise we once again hold in mind that "all models are wrong, but some are useful" (Box, 1976, pg.424), as even the best-fitting model to the data might be a poor approximation of the cognitive mechanisms underpinning decision-making and confidence judgements.

For example, the Bayesian framework operationalises decision confidence as a measure of probability correct, $P(Correct)$, (Aitchison et al., 2015). However, decision confidence reports appear to relate only imperfectly to this description. Indeed, Bayesian models of decision confidence have been shown to systematically under-predict confidence when reliability is high and/or stimulus values are extreme (Adler & Ma, 2018). Deviations from Bayesian optimal decision confidence are frequently cited. For example, the sub-optimal over-confidence of financial professionals in their own knowledge when judging the validity of nonsense financial terms appears inconsistent with Bayesian confidence (Atir et al., 2015). Indeed, even the broader patterns of over-confidence, (Brenner et al., 1996), undermine the notion that metacognitive confidence is an "optimal"

or Bayesian enterprise. Understanding how these sub-optimal confidence values are computed may shed light on the cognitive processes leading to the poor inferences that these judgements reflect (Adler & Ma, 2018).

Therefore, the computations that underpin confidence judgements have been explored in several model comparison studies, (Adler & Ma, 2017, 2018; Aitchison et al., 2015; Barthelmé & Mamassian, 2010; Xue et al., 2023). These studies have examined how well predictions of more sophisticated computations, e.g., a Bayesian inference of $P(Correct)$, describe decision confidence, as compared with simpler heuristic approximations. Heuristic models reflect "rule-of-the-thumb" computations that use a reduced amount of information to generate confidence more efficiently, though arguably less accurately, than more involved models. For example, observers could base confidence on a linear function of the stimulus value without accounting for the wider statistical structure of the task. In the random-dot motion task that has been commonly used in studies of decision-making (e.g., Shadlen & Newsome(1994)) and decision confidence (e.g., Kiani & Shadlen (2009)), this would mean increasing confidence judgements with an increased percentage of coherent motion. While some heuristic models have been ruled out completely (Barthelmé & Mamassian, 2010), heuristic models that approximate Bayesian ideals have demonstrated a good fit to decision confidence (Adler & Ma, 2017, 2018; Xue et al., 2023). However, whether this suggests that decision confidence is best described as a heuristic approximation of a Bayesian ideal is less clear. Building on this work, this thesis seeks to explore and compare some of these descriptions of confidence

while remaining focused on a specific type of decision that lends itself both to rich model exploration and comparison: category membership.

2.6 | Categories & Categorisation

The latter sections of this chapter seek to address the fundamentals of categorisation to allow for a discussion of confidence reports pertaining to category membership judgements. These judgements are central to the core aims of this thesis. By understanding the kinds of information used to learn, form, define, and adjust categories – how observers decide what category an item belongs to becomes clearer. In turn, this allows for a better understanding of the confidence judgements being made in relation to these types of decisions. This is important as a key aim of this thesis is to understand whether and how confidence reports can map out categories and category representations in a meaningful way. In essence, we ask what can decision confidence teach us about categorisation, and vice versa?

2.6.1 | What is a category?

A category pertains to a particular class of objects or experiences that are represented mentally by a concept. In terms of their importance, concepts have been described as "the glue that holds our mental world together." (Murphy, 2002, p. 1). This statement suggests that the cognitive processes that underpin the construction and maintenance of concepts are what hold together perception and meaning and, ultimately, help us to make sense of the world. Indeed, concepts include the relevant information pertaining to the properties of certain classes of objects: categories. Concepts also connect present experiences with

past knowledge to allow us to understand new objects as novel exemplars of known categories (Murphy, 2002). Concepts function to allow the determination of whether something classifies as being part of a category or not (Smith & Medin, 1981). The theories and models of categorisation have been heavily debated and have remained a readily evolving area of psychological and cognitive theory for decades. The following section will outline a brief history of some of these models, examining their strengths, weaknesses, and relative successes in trying to capture the often elusive features of a "category".

2.6.2 | Models of Categorisation

Early psychological studies (Hull, 1920), and philosophical discussions such as Aristotle's "Categories" (Barnes, 1991), centred around the definitional approach to categorisation. This approach is often described as the "Classical View" (Smith & Medin, 1981) and suggests that a category of objects could be defined verbally as distinct from another and was learned via experience. For example, the category of a house cat might be defined as follows "A cat is a small furry animal, with four legs, two pointy ears, and a tail. It meows a lot and is partial to cream, fish, and chasing mice." This approach includes the idea of necessity – certain definitional properties must be present in order to be eligible for category membership – and sufficiency – if all defining properties are present it must be a member of the defined category (Barnes, 1991; Hull, 1920; Murphy, 2002). Importantly, the categories suggested by this approach are clear and distinct with well-defined boundaries for membership. This approach was

pervasive from Aristotelian times through to the late 1900s but has since been usurped (Murphy, 2002; Smith & Medin, 1981).

The downfall of the classic definitional approach was two-fold. Theoretically, it is fundamentally difficult to define categories as distinctly as required by the definitional approach to categorisation (Wittgenstein, 1953). The existence of a three-legged cat or mute kitten, for example, would mean that such an animal would not fit the above-defined category of the house cat and would in essence lack a category or exist in a category of "a-not-quite-house-cat-creature" by itself. Empirically, the existence of a graded continuum for category membership (Hampton, 1979), "fuzzy" category members (McCloskey & Glucksberg, 1978), and variable typicality ratings (Barsalou, 1985; Rosch, 1973), all undermined the rigidity and clear cut nature of the classical viewpoint.

One way to adapt away from the classic definition approach to categorisation is to separate out information pertaining to the category member into defining and characteristic features (Murphy, 2002). This approach borrows from the semantic memory literature to describe certain features as essential to a concept while others – such as a house cat being able to meow effectively – are merely characteristic of a particular concept (Smith et al., 1974). It is then possible to construct a category based on a set of semantic features that vary in the extent to which they define the category (Smith et al., 1974). However, this "Characteristic Feature Hypothesis" has since been tested empirically by asking individuals to list and sort features of certain objects and concepts. Contrary to the original theory, this work has shown that both defining and characteristic features are

important to constructing category membership criterion (Hampton, 1979).

The typicality phenomenon of categories and concept structures, whereby some category members or features are more typical than others, gave rise to more recent and competitive theories about categorisation (Murphy, 2002). For example, the prototype view of category construction arose from the vacuum left by the fall of the classical view. Proponents of this view, (Rosch & Mervis, 1975), suggested the importance of a unified summary representation for each category (Murphy, 2002) and suggested that the likelihood of categorisation depended on the similarity of a category member to the category prototype. Items with greater similarity to the prototype would be more likely to be categorised as belonging to that category.

The notion of typicality remained central to prototype theory with certain members of a category being atypical to their concept group, such as penguins being atypical of the category of birds as they are flightless. Thus, just as with the classical view, defining exemplar typicality again became a central problem for the prototype theory of categorisation. A forerunner for quantifying typicality – natural frequency – was quickly overturned as the simple frequency of an item was not predictive of its typicality when tested across eight common categories (Mervis et al., 1976). To overcome such definitional hurdles about typicality, the idea of a prototype itself was redefined. Rather than the prototype being a singular item or object in a category it could be viewed as the summation of all category-relevant features, weighted based on family resemblance (Rosch & Mervis, 1975). In this way, based on this approach, typical items will have many

highly weighted features as compared with atypical category members. More specifically, typical category members will have many highly weighted features associated with their category, and few overlapping features in common with other categories (Murphy, 2002).

The typicality-based work of Rosch and Mervis (1975) was expanded on to further include the idea of category *ideals*. These ideals represent the degree to which an item fits the primary goal of a category (Barsalou, 1985). Perhaps unsurprisingly, these ideals were more relevant for man-made or functionally defined items such as tools than for natural kinds (Barsalou, 1985). This differentiation is inherent to the idea of the goal of a category, as tools are defined based on their functions which are goal-oriented, for example. Despite these extensions, a unified prototype theory of categorisation has remained elusive and decisive empirical support stalled (Murphy, 2002).

Over time, the exemplar theory of categorisation emerged in contrast to prior theories. Instead of defining concepts as either the sum total of all category members or a definitional (or otherwise) list of features, a concept under exemplar theory is the totality of all instances remembered by the individual (Medin & Schaffer, 1978). Also known as the "context theory" of classification, this theory overcame many of the challenges of the classical views such as being able to explain more clearly the differences in reaction times for typical and atypical category members as well as the intransitivity of categories (Murphy, 2002). Under this theory, classifying a new category member is based on the distance between exemplars from the category in question and the item being categor-

ised across the dimensions of the category, in a multiplicative fashion (Medin & Schaffer, 1978). However, the degradation of memories about category members, but not abstracted category prototypes, over time suggests that support for the exemplar framework over the original central tendency-based prototype models is not entirely clear cut (Posner & Keele, 1970). Indeed, the prototype and exemplar theories are not entirely separable in and of themselves, as the prototype model of categorisation does not deny the fact that people remember category exemplars with experience (Murphy, 2002). Work on connectionist models of semantic concepts exemplify this blurring with examples of networks that are trained on specific instances, but through iterated learning gradually exhibit categories that reflect the typicality structure of the categories experienced (Rogers & McClelland, 2004). The bottom-up approach of these models, which begin with neural mechanisms and then work to understand what macro-level functions emerge, is contrary to probabilistic models of cognition (Griffiths et al., 2010) which leverage the Bayesian framework with the aim to explain the exceptional generalisation ability of human cognition and it's ability to learn about new instances of a category with minimal training or input (Tenenbaum et al., 2006).

The above-described context model of categorisation, under the exemplar framework, was later extended and adapted into the Generalised Context Model (GCM). In the GCM, the multidimensional scaling no longer relies on the multiplicative rule of similarity distances proposed by Medin & Schaffer (1978) but instead comprises a more general distance metric that weights the different di-

mensions and square-roots the sum of the combined distances (Nosofsky, 1984, 1992). This weighting is data-driven, rather than model-specified, resulting in more naturalistic category structures emerging under this framework (Murphy, 2002).

These different models of categorisation emphasise different and complex categorisation criteria and the plethora of convergent, and divergent, ideas held by the field over time. This discussion is non-exhaustive and cannot begin to capture all the detailed debates and differentiated models over the decades. Instead, it stands to capture a flavour of some of the historically important discussions and families of models. However, a common thread throughout this discussion is that just as in decision-making where individual instances can be found nearer or further from a decision boundary, category membership itself can be uncertain. Indeed, category membership is often fuzzy and categorisations can be made with more or less certainty, just as in perceptual discrimination tasks.

2.6.3 | Multi-dimensional Categories

Beyond the structure of concepts, how stimuli from categories that are defined in multidimensional ways are perceived and classified is, in itself, a topic with a long history. Two opposing arguments have developed: the Combination Theory and the Differentiation Theory. Both of these frameworks try to explain how cognition proceeds, when time and cognitive resources allow, in successfully classifying stimuli with multiple relevant features.

Combination Theory can be traced to work around ideas of Feature Integra-

tion Theory from the attention literature (Treisman & Gelade, 1980) and relates directly to the GCM, (Nosofsky, 1984) (and the extended GCM - EGCM which includes a time course for sampling and processing different features (Lamberts, 1995, 1998)). This theoretical framework suggests that multi-dimensional category instances would be first perceived in terms of each individual dimension with these features later "combined" together, in a bottom-up manner. The Combination Theory of classification tracks well with ideas of how information processing proceeds at a neural level, with feature-level processing emerging earlier in the neural pathways before processing at the object level emerges (Hubel & Wiesel, 1959). By contrast, Differentiation Theory instead suggests that multi-dimensional concepts are processed in a top-down manner, starting with the object level before the individual features or dimensions are differentiated out. This suggests that stimuli may instead be processed holistically as a feature "blob", the components of which may be decomposed later, given time or cognitive resources (Lockhead, 1972).

The experimental research conducted based on the predictions of these two frameworks is extensive. In line with Combination Theory, when time pressure or cognitive load was increased, the similarity scores on a match to standard task dropped (Milton et al., 2008). However, the opposite effect is seen when time pressure is applied in another categorisation task – the triad task, in which participants have to state which two of three stimuli go together best, or which is the odd one out, when the stimuli are manipulated to be more or less similar on a single dimension of two – seemingly supporting Differentiation Theory (Ward,

1983). Recent work, replicating original studies of these frameworks, suggests that data are often compatible with both theories. However, it appears that Combination Theory alone is sufficient to explain the trends in the data (Wills et al., 2015). By replicating classic categorisation studies across seven experiments, results were shown to be consistent with Combination Theory in all studies, while some results were inconsistent with Differentiation Theory. For example, by analysing data from the triad task looking at the overall response-set (rather than as isolated trials), classification on the basis of a single attribute or dimension under time pressure, in line with Combination Theory, was detected. Differentiation Theory would suggest that processing on a single dimension takes time because the holistic "blob" must first be decomposed, therefore, this should not occur when time pressure is applied (Wills et al., 2015).

These discussions about single versus multiple dimension classification, with suggestions of integrated versus unintegrated dimension processing (Shepard, 1964), highlight the idea that, in some cases, the integration of dimensions is not necessary to differentiate categories. For example, when trying to decide if something is a cup or a bowl – as ordinarily defined by some combination of diameter and wall height – there comes a point when the walls of the object are too low for it to be reasonably described as a cup, regardless of the diameter. In this instance, a single dimension can be used to define the multi-dimensional space, leading to further questions about the serial or parallel processing of such dimensions under different constraints (Nosofsky, 1992) These kinds of instances introduce the notion of logical rules in differentiating stimuli in the

category space (Little et al., 2011), a framework that will be explored in more detail in subsequent sections.

As with the above section discussing models of categorisation, this section serves to demonstrate the complexity of the structure of categories. The potential complexity of category structures is highlighted by work describing the individualised and context-specific nature of semantic concepts which demonstrates why it is that concepts are often so inherently "fuzzy" and hard to define (Yee & Thompson-Schill, 2016). Here, in particular, the discussions served to highlight the nuances surrounding the weighting and processing of different dimensions. These kinds of differences are particularly clear in categories defined on multiple dimensions where there are both continuous and distinct dimensions, such as classifying mugs from bowl, or where one category can be defined based on a single dimension more readily than another (e.g., see the stimulus space of Little *et al* (2011)). These kinds of spaces will be explored further both in this review and in subsequent experimental chapters with reference to how confidence meaningfully differentiates these features.

2.6.4 | Categories & Confidence

As evidenced by the theoretical debates regarding how classification may occur, understanding how decisions about category membership are made is a complex topic. Even in everyday situations, classifying objects and events into appropriate categories is an involved process. How this may relate to reporting decision confidence is even less apparent. To illustrate these complexities it is

helpful to ground the discussion in a concrete example, here we will do so by taking the day-to-day example of the weather. When looking at the sky, light-coloured clouds often indicate dry weather while darker clouds predict rainfall. Based on this information it should be possible to classify days into dry or rainy day categories. In terms of behaviour, to decide if an umbrella will be needed, the observer could create a representation of rain-producing clouds and of non-rain-producing clouds. They could then compare the current sky to these distributions. However, this relationship is not absolute. Even if an observer knew the exact colour of the clouds, with no perceptual uncertainty, predicting the weather with any kind of certainty may still not be possible.

In this way, the weather observer must try to account for differences in the category means, variances, and perhaps even distribution functions to accurately represent a category distribution, make an appropriate decision, and later report their confidence. Adding to the complexity, naturalistic categories, such as rain clouds, may cluster together in stimulus-feature space resulting in overlapping category distributions for similar categories, adding further uncertainty. This contrasts with standard perceptual tasks in the cognitive sciences, such as luminance-gradient judgements, where the category boundary is discrete and uncertainty arises only from the noisy estimation of where the stimulus lies along that dimension. Regardless of which model of categorisation we appeal to, computationally speaking, categorisation involves the application of decision boundaries in this stimulus-feature space (Qamar et al., 2013). However, how this process relates to the computation of decision confidence is far from clear.

As previously discussed, one prominent proposal suggests that decision confidence is approximated on the basis of the distance from the stimulus measurement to the decision boundary, e.g., Kepecs *et al* (2008), Komura et al., (2013). However, applying this proposal to a categorisation task may under-describe the reliability of the stimuli. For example, consider decisions relating to categories with different levels of variance. Returning to the example of predicting weather from rain clouds, perhaps rain-producing nimbostratus clouds are more varied in colour, meaning their distribution has a wider spread across cloud-colour space. This would mean that a medium-grey cloud is more likely to produce rain than not, see Fig.2.3. Indeed, perhaps some nimbostratus clouds might be lighter than some non-rain-producing stratus clouds but will still produce rain. In Fig.2.3, consider the ratios of the curves to the immediate left and right of the red mid-line (which for the purposes of this graphic is placed at the point of equi-probability). An observer should be less certain that it will not rain when the stimulus falls just left of the mid-line than they are confident that it will rain just to the right, a violation of the strict "distance to the bound" model of confidence. While this idea has been explored with nested categories with shared mean values (Adler & Ma, 2018), this has not been explored with categories whose means lie on either side of the category boundary.

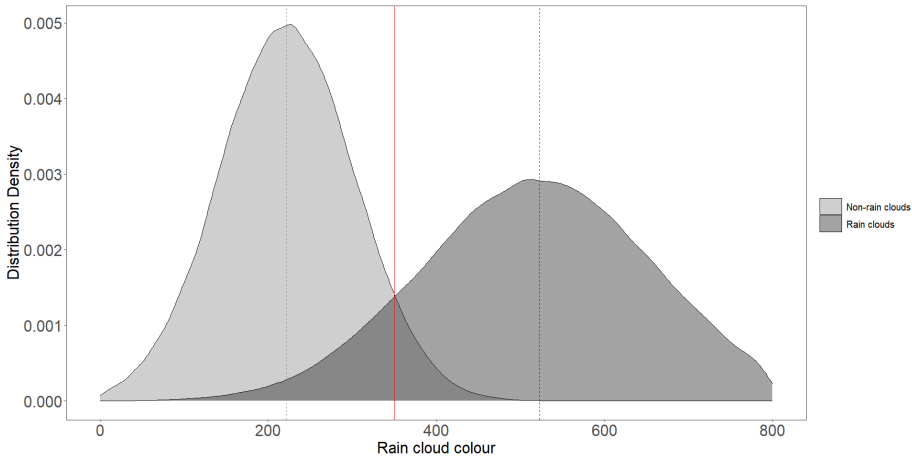


Figure 2.3: Simulated rain-cloud distributions. The density of category stimulus occurrence on the y-axis and cloud colour on the x-axis. Two discrete but overlapping categories of rain-producing nimbostratus (right) and non-rain-producing stratus (left) clouds are shown.

By extending work into these types of overlapping category distributions, we allow for the comparison of Bayesian predictions of decision confidence with predictions of simpler heuristic models of confidence such as the distance-to-bound model. Indeed, while Bayesian models would utilise the differences in variance to inform confidence reports, it is unclear how the "distance to the bound" model would hold given circumstances such as Fig.2.3.

2.6.5 | Multi-dimensional Categories & Confidence

Natural categories, as opposed to the artificial perceptual categories imposed by experimenters, provide further scope for these kinds of comparisons. Importantly, they are often multi-dimensional, allowing for more complex probabilistic judgements to arise. For example, apples and pears differ in both shape and

colour, and species of flowers might vary in terms of size, colour, and petal shape and number.

To discuss this in more detail, let's return to our dermatologist who is trying to decide if a skin lesion requires treatment or not based on size. Consider that, beyond the size of the growth (which may be seen as a proxy for the likelihood that will require treatment), the colour of the lesions may also vary systematically (e.g., from light to dark). For example, perhaps malignant growths, due to their growth rate, are more likely to also have a difference in colour because darker hues can also indicate a change in cell division. Therefore, the percept value of the dimension of colour hue changes the probability of category occurrence and should influence confidence reporting. Arguably, just as category-relevant information – like category variance – should change the probability of a stimulus falling in a category, thus affecting confidence, probabilistically related dimensions such as colour hue should also influence confidence. Again, while Bayesian models, by creating a statistical model of the world, encompass these dimensions, computations of confidence based on distance-to-bound models would not immediately account for these complexities.

Relative to a simple one-dimensional perceptual task, an n-dimensional category space the category boundary is shifted such that a classic distance-to-bound model may no longer be an accurate predictor for confidence reports, see Fig. 2.4 for a simple example. Perhaps, here a trainee dermatologist is asked to classify treatment plans based on the size of the lesion along the x-axis. Colour on the y-axis is probabilistically related to this size dimension. Malignant le-

sions occupy the narrow space in the top right of the heat plot where the size is large and the colour is darker. The benign lesions occupy a wider space where size and colour are generally lower, in the lower left corner of the space. Here although, by definition, the accuracy of the response, and the classification rule itself, is based on size, the use of the colour dimension may change confidence reports. For example, confidence may be higher in lesions that are both large and dark than those which are just large. In the distance-to-bound model of confidence, this variation is neither predicted nor explained.

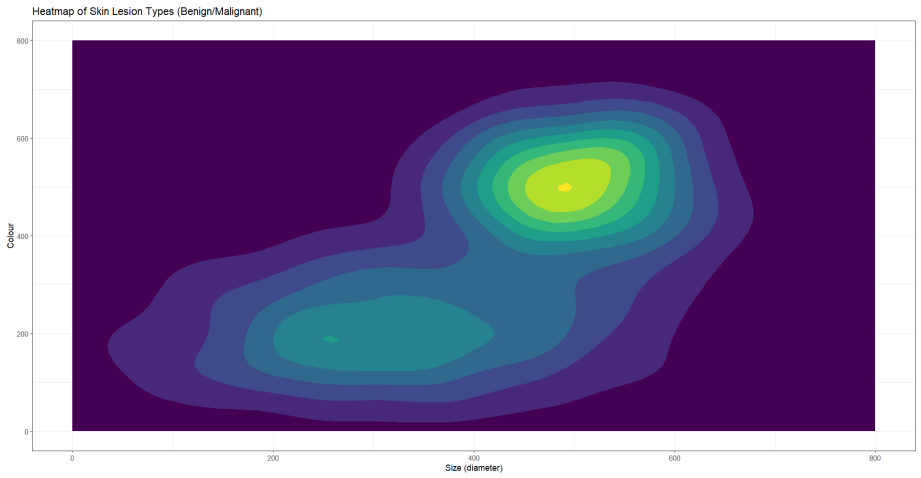


Figure 2.4: Simulated heat-map of two categories of skin lesion that vary across 2 dimensions, size and colour. Variation in colour on the plot relates to differences in the density of category stimuli, with lighter colours equating to more frequent occurrences. Note that the two categories have different spreads, relating to the different variances that they have. Note also that a decision boundary relating to the location on the x-axis only may not be as informative to confidence in this 2-dimensional category space whereas a boundary along the diagonal, or distance from a category prototype may become more informative to decision confidence.

Indeed, how individuals utilise and weigh multiple stimulus dimensions in categorisation decisions and, critically, how this affects how they judge their con-

fidence in those decisions remains largely unexplored. It is clear that decision-makers have the ability to form and utilise complex category distribution representations to inform their decisions and confidence, as this ability is necessary to navigate even the most everyday of tasks. However, the mental representations of these task spaces in relation to their accompanying computations of decision confidence are relatively under-articulated.

Therefore, over the course of its work, this project will investigate if decision confidence reports reflect a change in decision boundary position, for example, when category dimensionality shifts, and in turn if confidence approximations are better predicted by distance to that bound or if an alternative computation better describes the data such as a normative Bayesian model or an alternative heuristic-based model such as distance to the category prototype.

Similarly, there are instances where a discrete category boundary can alter the notion of "distance" in a multi-dimensional category space. In a discrete category space, there is no overlap in any dimension between the categories. As such, distance to the boundary could be a linear read off of the confidence judgements in the space. However, when the dimensionality of the categories increases, the measurement of "distance" that is being judged becomes largely unspecified: does it depend on the distance to the category boundary in one dimension (and, if so, which) or some combination of the different dimensions that define the stimuli?

To try to explain some of the potential operations of categorisation in a multi-dimensional space, a family of models known as logical-rule models of

categorisation have emerged (Fific et al., 2010). These models vary in their approach for categorising category item members based on either the serial or parallel application of logical rules in the category space (Little et al., 2011). They focus on the variable reaction times of response for different items in neighbouring categories. For example, consider again a category stimulus space that defines the categories of bowls versus cups. Arguably, for an item to be considered a cup it must have both relatively high walls and a narrow diameter. By contrast, a bowl can have high walls (such as in the case of a cereal bowl) or lower walls (as per a pasta or soup bowl) and a varied diameter (a sugar bowl and a fruit bowl are both still considered bowls). These type of categories, and their delineation, are interesting because they lend themselves to exploring the idea that some categories are easier to define than others. For example, in the cups and bowls comparison, once it has been established that the diameter of the object is wide enough then it becomes unlikely that it is a cup and so by virtue of this the height of the walls does not need to be attended to. By contrast, knowing that something has a smaller diameter does not eliminate the possibility that it could be a bowl, so the height of the walls must also be attended to in order to classify a cup as such.

This inequity of information required may have consequences for confidence reports, and these ideas become particularly important when we consider the role of confidence in category judgements. If the computation of confidence makes use of available information, such as category variance and location of decision boundaries, it may map out the feature space in a way that accur-

ately reflects the mental representation of the stimuli learned by the observer. Indeed, confidence may drawn on the most salient or relevant dimension or feature of a category, following a single-dimension classification approach in a multi-dimensional space, for example. In turn, the asymmetrical approaches that are used by the observer to categorise in that feature space may play out in confidence just as in reaction times. Importantly, if differences in confidence suggest that different forms of classification are being used for different categories, it would mean that not only are observers representing the category space in a flexible fashion but the underlying confidence computations might differ across and between category spaces. This may reflect an approach that optimises not just cognitive but also meta-cognitive efficacy, in terms of using single-dimension classification where possible, over absolute accuracy, for example.

In conclusion, these discussions indicate that categories have a rich structure that usefully extends models of confidence beyond simple (perceptual) decision tasks that have mostly been used. It is clear that, methodologically, the richness of categorical structures will allow for interesting explorations into the responsiveness of confidence to these variations. Furthermore, the variability of confidence in these spaces might further have theoretical implications for the conceptualisation of the computations of metacognitive confine more broadly. In turn, although not a subject of study here, these ideas also suggest that confidence might be a useful measure in future categorisation research.

2.7 | Thesis Overview

This thesis aims to map out the use of confidence in categorisation in order to understand what information is being attended to in what types of categorisation tasks, and how that translates to confidence computations. By allowing the intersection of these theories of categorisation and decision confidence to become central in the discussions, the usefulness of confidence as a tool to understand the flexibility of cognitive representations becomes paramount.

In Chapter 3, I re-analyse data from a previously collected dataset from a task modelled on standard perceptual decisions but begin to implement key features of categorisation tasks such as there being a probabilistic (rather than deterministic) relationship between perceptual value and the correct response, and the fact that the response has to be learnt from feedback rather than being specified in the design. The goal of the modelling exercise is to try to understand how confidence reports might reflect differences in the representation of the underlying distributions used. This work was grounded in the idea that, by leveraging differences in task difficulty across experimental sessions, the task representations used by participants could be manipulated. A computational modelling exercise was completed to explore these ideas in depth. This served as the starting point for thinking about different models of confidence, and how categorisation tasks might help to tease them apart.

Chapter 4 then explicitly manipulated the distributional properties of stimulus categories in three behavioural studies, examining a specific feature of the

category distributions – namely their variances – to explore specific predictions about confidence. Models of confidence such as the distance-to-bond and Bayesian normative models are discussed as relevant to these spaces.

Finally, in Chapter 5, I step into the richness of multi-dimensional category structures looking at the specific predictions about confidence both across and within the bounds of a discrete category space. Leveraging a two-dimensional space with a discrete bound, the predictions of distance models of confidence and discussions of rule-based models of categorisation are explored as relevant to the confidence reports elicited in this space.

Therefore, over the course of the subsequent chapters of this thesis, this series of empirical and computational works will be described with the aim of exploring the intersection of confidence and categorisation. This work aims to further the understanding of the value of traditional models of confidence as relevant to the field of categorisation and our understanding of cognition more broadly.

CHAPTER 3

Computing Confidence

Abstract

This first experimental chapter of this thesis comprises a secondary data analysis and computational modelling exercise exploring confidence in a perceptual decision-making task. By leveraging differences in task difficulty across experimental sessions to try to manipulate the task representations used by participants, as expressed in their metacognitive decision confidence, this experiment sought to understand when and why observers might engage with a Bayesian normative or a heuristic rule-based model of the task space.

In this chapter, after exploring the initial ambiguous model fit to this data, by conducting model-based simulations to understand where the models were failing to capture the patterns of the data, a model extension and the subsequent model fitting procedure were carried out.

The results, while still ambiguous in terms of what the "winning" model might be, allowed for discussion about individual differences in task representations, notions of "optimality" under contextual constraints, and the persistence of task representations over time.

3.1 | Introduction

This thesis has the overarching aim to understand the interplay between confidence and categories from both a methodological and theoretical standpoint. To explore these questions, different kinds of categorisation tasks will be deployed over the course of this thesis. However, the first empirical chapter of my DPhil draws from work on a more traditional perceptual discrimination task, in which participants are asked to perform a simple, binary categorisation of a sensory attribute – specifically, judging whether the orientation of a stimulus fell above or below the particular angle of plane.

This task was inspired by a standard perceptual task but began to move away from this framework in two key ways. Foremost, participants were not told the classification rule but had to learn it via trial-by-trial feedback over the course of the experiment and training. Secondly, the stimuli were generated from two overlapping category distributions, so while accuracy was absolute relative to the experimenter-imposed response-bound, the generating distributions were more complex than the underlying uniform distributions often leveraged in a perceptual discrimination task.

The aim was to investigate if task demands would alter participants' representations of the distribution of stimulus orientations and whether this would be reflected in their confidence reports. The challenges and limitations of this work inspired the later shifts in design to tasks defined by their categorical rather than their perceptual uncertainty – which will be observed in subsequent chapters.

However, the questions explored here are, in themselves, important. When and why do individuals engage in differential task representation, how dynamic are these approaches, and why might it be "optimal" to engage different category representations under different contexts?

Optimality is a complex notion. Within the field of cognition, to claim that the mind is, at some level, operating optimally aligns with a theory of cognition known as the Bayesian Brain Hypothesis. This theory suggests that humans operate as ideal Bayesian observers in their interactions with the world (Beierholm, 2020). As outlined in the previous chapter, the Bayesian perspective suggests that the world's statistical structure is combined with current sensory input to calculate the likelihood of specific outcomes or states of the world (Adler & Pointon, 1993) – such as the likelihood that a berry is edible versus inedible – and that this input is used to guide behaviour accordingly. Importantly, these claims do not suggest that human cognition is, in itself, optimal. Indeed, for example, imperfect sensory processing (Arnold et al., 2019) and sub-optimal memory processes are widely acknowledged across scientific subdisciplines (Loftus, 1996). However, despite these imperfections, the Bayesian Brain Hypothesis suggests that the ideal Bayesian observer follows a process that combines our beliefs, the noisy evidence we have, and our objectives in an optimised manner. That is, despite the sub-optimality of the design of human cognition, our decision-making process can lead us to make optimal decisions given the information we have via Bayesian inferences (Bowers & Davis, 2012). This kind of "optimality" has been seen in some multi-sensory cue integration

tasks (Battaglia et al., 2003; Butler et al., 2010; Ernst & Banks, 2002), and has been used to explain common phenomena such as the perceptual "Ventriloquist Effect" (Alais & Burr, 2004) and cognitively the rapid acquisition of priors (Jazayeri & Shadlen, 2010).

In the mind of a Bayesian observer of a two-choice perceptual decision-making task, the individual trial-by-trial stimuli during a training phase would be used to build an understanding of the world of the task – how the stimuli relate to each category and how the categories relate to each other. Under a common task structure, with category stimuli sampled from a standard Guassian distribution centred on some mean, the Bayesian observer would integrate that information about the category distributions, their means, and therefore their position relative to each other, into both their choice and their subsequent confidence judgement. This would mean that the observer "represents" on some level the normal distributions of the categories in order to produce optimised responses. By doing so, the observer also represents, for example, the overlap of those category distributions – areas of the stimulus space which should evoke lower confidence reports because the likelihoods of the two categories would be very similar. This compares with a rule-based observer who would learn about the task structure – perhaps through rote memorisation or indeed through some kind of sensory integration that aligns with Bayesian principles – and then insert a hard and discrete boundary in the space around which to make their binary choices. In this way, a rule-based observer would represent the space very differently than a Bayesian observer. Their representation of the category

distributions might be better described as uniform distributions on either side of the bound (if their rule is absolute) or, if the category is centred strongly on a prototypical stimulus, perhaps constructing symmetrical normal distributions on either side of the bound.

Although Bayesian conceptualisations of cognition have become increasingly popularised (Beierholm, 2013), the theoretical basis of the Bayesian perspective remains primarily computationally described rather than algorithmically determined (Marr, 1982). That is to say that they are described in terms of the logic and goals of the information-processing task but do not concern themselves with how this computation can actually be implemented and by what mechanisms the input or output will be represented (McClamrock, 1991). This can lead to a questioning of the relevance of Bayesian models to theories of cognition (Bowers & Davis, 2012), where understanding the mechanisms behind cognition is often a primary goal.

However, Bayesian theorising itself can be thought of as falling into one of several levels (Bowers & Davis, 2012). In principle, extreme theorists would claim that understanding how the mind functions mechanistically is irrelevant when compared to the importance of exploring how individuals navigate uncertainty to solve problems in their environment, for example. However, this is a viewpoint that few would commit to within the realms of the cognitive and psychological sciences as it essentially dismisses many aspects of cognition research (Bowers & Davis, 2012). However, some have extolled its utility for forwarding understanding of behaviour - such as in Movellan and Nelson's commentary on

the "probabilistic functionalism" framework of generalisation (2001).

A more popular view does not dismiss the idea that implementation is important but posits that numerous mechanistic implementations of cognitive processes could achieve parallel performance in cognition (Geisler & Ringach, 2009). This methodological viewpoint sees Bayesian models as tools to create benchmarks for optimal performance in particular tasks and, therefore, makes no claim about the specific cognitive mechanisms that might lead to these outputs.

By comparison, the theoretical Bayesian suggests that the computations themselves adhere to Bayesian principles, such that near-optimal outputs in behavioural tasks are seen as evidence of Bayesian operations at the algorithmic level. Indeed, some reviews of the evidence have suggested that the presence of Bayes-optimal behaviour in different task domains makes it likely that the mechanisms underpinning that behaviour could be computationally described as Bayesian (Körding & Wolpert, 2006). However, this chapter – and this thesis more broadly – adopts the above-described methodological Bayesian viewpoint and makes no direct claims about the mechanisms that may underpin these behavioural data at the algorithmic level.

This theoretical stance is partially due to the fact that it is entirely possible that "optimal" behaviour could arise from different types of computations (Geisler & Ringach, 2009). However, another important argument in favour of the methodological standpoint lies in the fact that it is also possible that what is optimal, in terms of balancing cognitive load or managing a complex social environment, is better achieved through non-Bayesian computations either at

the computational or algorithmic level. Indeed, although contrary to Bayesian models, which suggest observers fully represent category distributions, human observers often rely instead on heuristic-like (Tversky & Kahneman, 1974) rule-based category divisions (Wills & Mackintosh, 1998) to guide their behaviour. It has been found that observers' use of these rule-based strategies depends upon the ease of rule creation as a function of stimulus complexity (Livesey & McLaren, 2009). For example, in an easy task, observers can abstract a simple rule in place of full distribution computation, whereas in complex tasks a simple rule cannot be derived and the observer must either compute the distribution fully or learn it through experience via a trial-by-trial stimulus mapping (Livesey & McLaren, 2009; Wills & Mackintosh, 1998). Counter-intuitively, this suggests that human observers appear to compute less accurate category representations when stimuli are simple, and conversely are more accurate in their distribution representations when stimuli are complex. Based on these ideas, it could be predicted that for a simple perceptual decision, participants would develop a rule-based heuristic model in place of full Bayesian computations, whereas in more complex decisions they would exhibit more Bayesian-like performance. This does not refute the notion that individual sensory neurons may be firing in line with Bayesian predictions (Lak et al., 2017) but suggests that, at the level of decision processing, another computation may be dynamically employed by human observers to navigate in particular circumstances. These ideas form the basis of this chapter which aims to perform a comparative model fitting on an existing dataset to better understand how decision-makers computing their de-

cision confidence, whether a Bayesian model of confidence best describes the data, and to explore more broadly how individuals are representing information about perceptually uncertain categories during the decision-making process.

3.1.1 | Reanalysis Rationale

What information people use to help assess and report their confidence, and when, is central to this doctoral project, as it helps to clarify our understanding of the information used by decision-makers during the decision-making process. Therefore, the first step in this DPhil was to reanalyse an existing dataset, drawn from my own unpublished MSc project (2018-2019). This thesis explored whether Bayesian or simple rule-based models better predict participants' perceptual decisions and confidence and if the model fits varied across easy and difficult task versions. The initial model-fitting work, conducted during my MSc, failed to find a best-fitting model and there was little clarity on where the models were failing to describe the data. The current reanalysis in this thesis aimed to clarify my previous work by extending the models under comparison to include new variables and to restrict the models by forcing them to predict both responses and confidence reports simultaneously. This project, therefore, tried to disentangle the models from each other and better understand the behaviour of each model in relation to the actual data. This work, therefore, attempts to shed light on the representations of categories that people hold, under different contexts, as reflected in their metacognitive decision confidence reports.

The original "Computing Confidence" project tested the prediction that

people would form different category representations when performing simple versus complex decision tasks, with distinct impacts on confidence in each case. Central to this prediction is the idea that intrinsic to the decision-making process is the ability to generalise from learned stimuli to new category exemplars. This generalisation is achieved via the creation and access to category distribution representations. A key marker of the ability of the decision-maker to represent the distributional properties of categories, and to represent them is a phenomenon known as the peak shift effect (Andrew et al., 2014).

While the peak shift effect itself is not the main focus of study in this chapter, it does serve to illustrate two key points – that human decision-makers can and do represent category distributions and their overlap (Wills & Mackintosh, 1998), and second that human observers do not always do so and may instead invoke simplified distribution representations in some contexts (Livesey & McLaren, 2009). Therefore, this effect will be explored in some detail here.

The peak shift effect describes a shift, or transposition, of the expected stimulus-response patterns relative to the objective "peak" of the likelihood distributions of each category when exemplars are drawn from two categories with overlapping probability distributions (Cheng et al., 1997), see Fig. 3.1 to visualise this effect. The stimulus-response patterns to Category A have been "shifted" away from Category B and are plotted as Category A*.

The mechanism behind this effect unfolds when two stimulus categories along a dimension, such as colour hue, are drawn from overlapping distributions on that dimension. This setup is a classic approach to stimulus sampling in the ex-

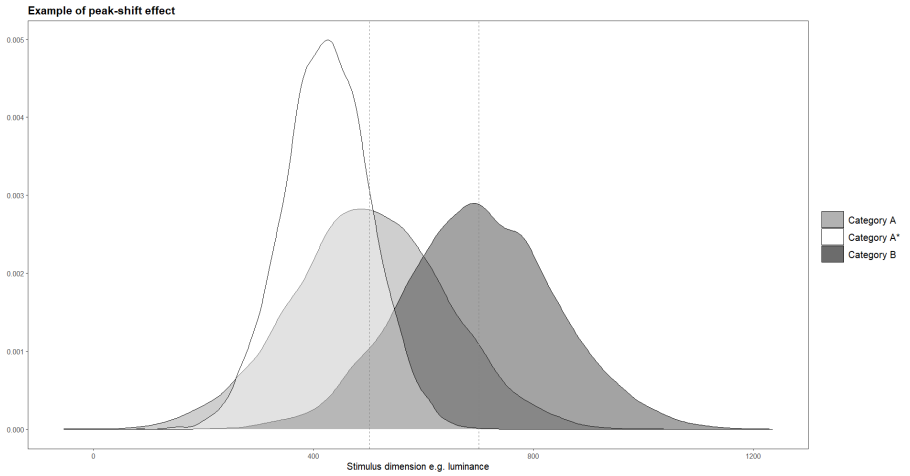


Figure 3.1: A simple illustration of the type of patterns seen in classic peak-shift studies is shown. Two category distributions A and B are plotted with some overlap, their means are demarcated by the dashed vertical lines. Because of the inhibitory effect of the overlap between the two distributions the representation of Category A, and its "peak" is shifted away from Category B. Note that the peak of Category A has shifted away from the mean and that the gradient of the "shifted" representation of Category A* is much steeper.

perimental design of perceptual decision-making tasks, for example. Narrowing in on the area of overlap of the distribution, at this point in the category distribution space, when trying to identify a Category A stimulus, the excitatory response from the Category A distribution will be subtracted from the inhibitory response of the B stimuli in that distribution region. This means that the response is reduced. If we compare that muted response from the trained Category A stimuli to another stimulus that sits on the opposite end of the Category A distribution away from the overlap we see the "peak" of the response shift. The response of the untrained stimulus is higher than the corresponding response of the trained stimulus in the distribution overlap - and the "peak" of responsiveness is now offset away from the maximal likelihood of the category

distribution. This difference of response, this shift in responsiveness, effectively indicates that via the computation of category distributions, and their overlap, the category representations themselves are influenced by a discounting away of the exemplar values. This has been replicated extensively and can be seen across many animals from tasks in spatial learning in pigeons (Cheng et al., 1997) to olfactory learning in honey bees (Andrew et al., 2014). The peak shift effect provides evidence of the computation of complete category representations during categorisation tasks that are parallel to assumptions made by Bayesian models of decision-making and, indeed, decision confidence.

However, despite the popularity of Bayesian models of human cognition, surprisingly, human decision-makers rarely display this peak shift effect (Wills & Mackintosh, 1998). Instead, it appears that human decision-makers rely more heavily on rule-based approaches that may be best described as approximate heuristic or rule-based computations (Tversky & Kahneman, 1974). Indeed, in standard perceptual tasks, "unless stimuli are presented very quickly, concurrently with another task, or are very complex and difficult to verbalize, human subjects will almost invariably try to implement cognitive strategies, the use of verbal labels and relational concepts, the formulation and evaluation of hypotheses, for example to solve discrimination tasks" (Wills & Mackintosh, 1998, pg.28). Indeed, more recent studies suggest that the use of rule-based over associative learning necessitates the ability to verbalise the difference in stimulus categories (Livesey & McLaren, 2009). However, importantly, when this simplified rule-based approach learning is prevented due to the complexity of

stimuli, the peak-shift effect can be observed (Wills & Mackintosh, 1998). For example, when challenged by an artificial category dimension that comprised a grid of black and white icons with the presence, absence, and prevalence of some icons indicating the category membership of each trial, human decision-makers demonstrated the peak shift effect in their responses to novel and untrained stimuli (Wills & Mackintosh, 1998). It seems that the more complex a stimulus, the higher the dimensionality of the categorisation process, and, consequently, the more difficult it is to articulate a simple rule to delineate the stimulus space. When a stimulus becomes complex, creating a simplified heuristic or rule-based logic for the decision process becomes untenable. By preventing the ability to learn using rules, observers are forced to return to a reliance on associative learning and, in turn, learn or compute category distributions more comprehensively. The ability to create rules to categorise stimuli appears to align with higher-order cognitive functions as the observer must abstract away from trial-by-trial associations, a cognitive leap which requires higher level manipulation of stimulus knowledge than is achieved by most species (Andrew et al., 2014; Cheng et al., 1997; Wills & Mackintosh, 1998). In categories defined by higher dimensionality, a simple verbal rule becomes harder to abstract out of the task and accompanying perceptual experience (Livesey & McLaren, 2009; Wills & Mackintosh, 1998).

In this way, given that human observers appear to seldom show the peak shift effect (Wills & Mackintosh, 1998), it seems that they seldom fully represent category distributions when making decisions. Instead, it appears that

we often rely instead on the heuristic-like (Kahneman & Tversky, 1974) rule-based division of categories that do not represent full category distributions or their overlaps. As the point at which human observers move from this rule-based strategy to full category distribution representation is dependent on ease of rule-creation as a function of stimulus complexity (Livesey & McLaren, 2009; Wills & Mackintosh, 1998), counter-intuitively, human observers compute less accurate category representations when stimuli are simple and conversely are more accurate in their representations when stimuli are complex. The behavioural experiment and model comparison conducted in the original "Computing Confidence" project aimed to elicit and delineate these different modes of category distribution representation using metacognitive decision confidence as the modelled outcome. Therefore, this chapter will first explore the original work, and its limitations, before detailing the model extension and exploring those results.

3.1.2 | The Original Experiment

In the behavioural study, participants were asked to classify Gabor patches into one of two categories based on their orientation and then report their confidence in that choice. Thus, they made a decision and confidence judgement based on perceptual information as it related to the categories defined in the task. Importantly, to test the idea of rule-based versus associative learning under different task demands, the experiment was designed to have easy and hard versions of the trials. These versions aimed to specifically elicit either rule-based

or full category distribution representations, respectively. In the simple version of the task, the target Gabor was presented alone. The orientation of that Gabor defined the category to which each trial stimulus belonged and therefore the accuracy of the participants' responses. By comparison, on complex trials, the target Gabor was presented with a distractor Gabor. The target and distractor Gabor had different orientations and were presented simultaneously. The location of the target and distractor remained constant on-screen within each participant. Therefore, in the complex version of the task the participant had to identify the target Gabor and then make a choice based on the orientation of that Gabor alone. This increased the dimensionality of the task and made a simple rule verbalisation such as "respond left when the Gabor orients to the right" more difficult to achieve. On each trial, either hard or easy, after inputting their decision about which category they felt the trial was drawn from, participants reported their decision confidence on a continuous sliding scale ranging from "Definitely correct" through "Don't know" to "Definitely wrong" (Fig.3.2). The participants were given trial-by-trial feedback on their response accuracy as well as block-wise feedback.

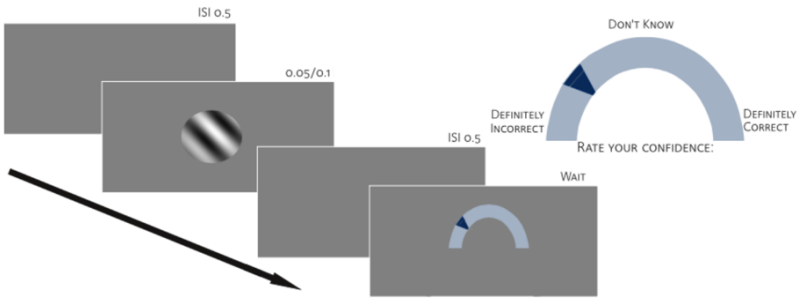


Figure 3.2: On-screen experimental trial-by-trial set-up. Figure reproduced from Ashcroft-Jones, 2019 - unpublished MSc thesis.

In order to produce a well-powered dataset, the target sample size was established to match the 35,640 test trials used for model comparison in the study by Adler and Ma (2018). Thus, the target sample size was 14 participants with a total number of 2520 test trials collected per participant. To account for the number of trials and therefore testing time required, the experiment consisted of four testing sessions across multiple days. Each testing session was designated to have either an easy or a hard stimulus set, with a counterbalancing across and within participants of the session order. The study recruited 14 participants, of whom 13 completed all 4 testing sessions. The participant who did not complete all sessions is excluded from all analyses, in line with the study preregistration. The behavioural results and subsequent modelling are therefore mapped to the data of 13 participants and their combined 32,760 trials unless otherwise indicated.

The orientation of the Gabor patches varied across trials. The orientations were drawn from one of two category distributions which overlapped in percep-

tual space and straddled the horizontal point. This was the deciding factor as to the participants' accuracy of response – whether the response aligned with the category from which the target stimulus was drawn. In complex task trials, the category from which the orientation of the distractor stimulus is drawn was also randomly allocated. However, the categories were designed to overlap such that perceptually the orientation may appear to be drawn from one category when in fact it was drawn from the other. This manipulated the categorical certainty of each trial. Therefore, trials were more or less categorically uncertain dependent on from where in the category distributions the stimulus orientation was drawn. Trials drawn from the overlap of the categories would be highly uncertain as to their generating category and so a low confidence participant response would be expected if full category distribution representation was being utilised by the participants. In tandem, the contrast of all Gabor patches was also manipulated across trials. On complex trials, the same contrast level was applied to both patches. The sinusoid used in the calculation of the Gabors was multiplied by a constant drawn randomly from an array of five values, [0.1, 0.2, 0.3, 0.4, 0.8]. This variable manipulated the perceptual certainty of the trials.

Thus, across the experiment, individual trials varied in their difficulty in three ways. The first, which varied between sessions, was the presence or absence of a distractor Gabor to increase the dimensionality of the task which was hypothesised to evoke differential category representation – either rule-based or full distribution representation. The second, which varied trial-by-trial within and between blocks (and sessions), was the categorical uncertainty which de-

pended on where in the stimulus space the stimuli were drawn such that a participant might know what the orientation of the Gabor was – but not know what it means in terms of category identification. Finally, all of the trials varied in terms of how perceptually certain they were based on the stimulus contrast applied to the Gabor patch on screen, changing the perceptual ease with which the individual stimuli and their orientations could be discriminated.

The core experimental question was whether decision-makers computed a full category distribution or if they "threw away" information by simplifying the category distribution representations. In this study, I predicted that participants' confidence would be better explained by a discrete rule-based model (Fig. 3.3b) when the task was simple, and by a full category distribution model (Fig. 3.3a) when the task was complex. In Fig. 3.3, the left panel (a) shows the true category distributions which include the represented overlap of the distributions. This overlap would mean that even perceptually distinct stimuli would be uncertain as to their category membership, and this would in turn be reflected in the confidence of the participants. By comparison, in the rule-based strategy, shown in the right panel (b) of Fig. 3.3, a key idea is that participants will always report a particular category choice in line with a binary perceptual discrimination – therefore only perceptual noise will corrupt this relationship and the corresponding confidence reports. These predictions were in line with the previous findings around decision response patterns (Livesey & McLaren, 2009; Wills & Mackintosh, 1998) but shift the focus to metacognitive confidence as the key dependent variable.

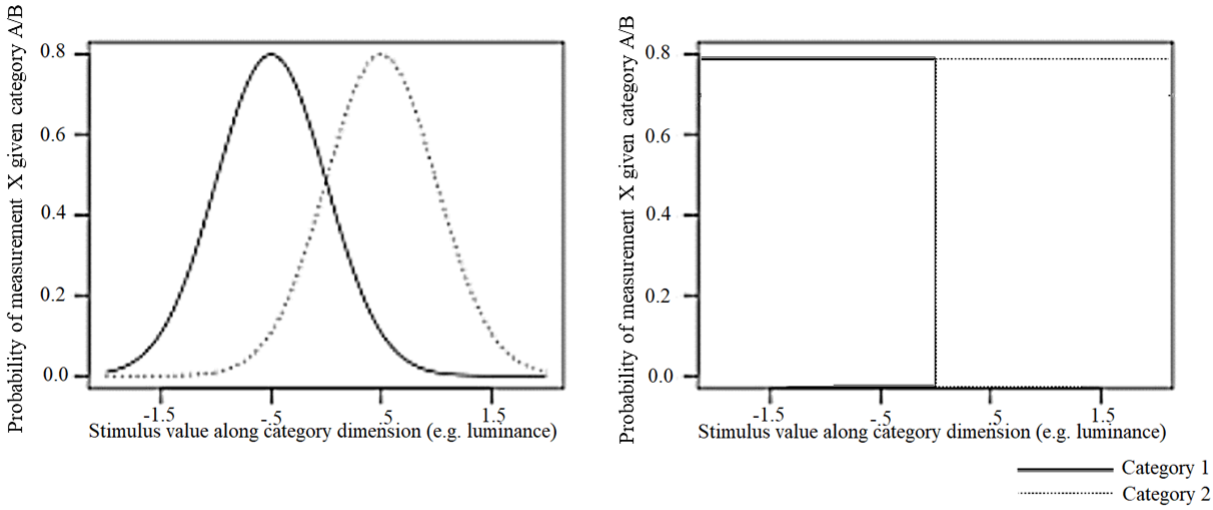


Figure 3.3: The hypothesised cognitive representations of the true category distributions computed when stimuli categories are complex (left) as compared with rule-based category distributions computed when stimuli categories are simple (right). Figure reproduced from Ashcroft-Jones, 2019 - unpublished MSc thesis.

For more details of the original analysis plan visit the OSF preregistration. To view the experimental code or the model-fitting visit the relevant GitHub repository.

3.1.3 | The Original Models

The original model fitting included a model representation of a Bayesian observer with fully constructed and represented category distributions and an alternative, simple rule-based model. The key predictions of this study were

about which model the participants used under which experimental conditions.

Whether decision-makers used rule-based strategies or computed full distribution representations was tested using a comparative model fitting procedure for a normative Bayesian model for complex trials and a rule-based model for simple tasks (Model 1 - Proposed Model), and the converse (Model 3 - Rev-proposed Model), as well as a normative Bayesian model (Model 2 - Full-Bayes Model) and a rule-based model (Model 4 - Full-Rule Model) across all trials. These four models all modelled confidence as binned into 10 bins within participants from the continuous input scale. This binning was conducted within participants such that the continuously measured confidence rating is replaced with a number from 1-10 according to the decile bin it falls in after pooling across all sessions, allowing for more meaningful comparisons across participants.

In deriving the original models (described in more detail below), it was assumed that a Bayesian observer with a correct generative model would use the environmental statistics, and a noisy stimulus measurement, to infer the probability that a stimulus belonged to a category, see Fig. 3.4. For each decision, the observer would be expected to select the most probable category, choosing the category C with the highest posterior probability $p(C \mid x)$, where x is the stimulus measurement. This equates to reporting and responding that a trial belongs to category 1 (C=1) when $\log \frac{p(C=1|x)}{p(C=2|x)}$ is positive and to category 2 (C=2) otherwise.

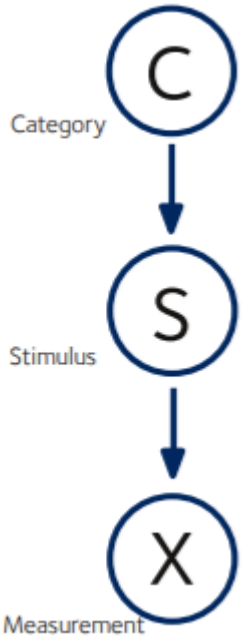


Figure 3.4: Observer generative model: Schematic of the Bayesian observer generative model. Figure reproduced from Ashcroft-Jones, 2019 - unpublished MSc thesis.

To understand Fig. 3.4 in detail, we must imagine that if an observer has a perfect perceptual system that provides direct and unperturbed access to the stimulus, they must then decide whether the stimulus, s, is an example of Category 1 or Category 2 stimuli, both of which follow a normal distribution centred on their respective means. That is,

$$p(s|C=1) = N(s; \mu_s, \sigma_s^2) \tag{3.1}$$

$$p(s|C=2) = N(s; -\mu_s, \sigma_s^2) \tag{3.2}$$

$$p(C=1) = p(C=2) = \frac{1}{2} \tag{3.3}$$

However, our Bayesian observer may face both unreliable evidence and unreliable perception and must instead rely on a noisy measurement, x, of the stimulus, s.

$$p(x|s) = N(x; s, \sigma^2) \tag{3.4}$$

Given these assumptions and our understanding of the decision being made by an observer based on the log posterior ratio, we can calculate how the observer will translate their perceptual measurement x into a decision under a normative Bayesian model as follows;

$$\frac{p(C=1|x)}{p(C=2|x)} \quad = \quad e^{-\frac{2x\mu}{\sigma_s^2+\sigma^2}} \tag{3.5}$$

The alternative rule-based model was founded on the ideas of rule-based decision-making. It assumes that the observer uses their knowledge of the midpoint between the two category distributions, but incorrectly represents the

category distributions as uniform, and divides the stimulus space into two parts with orientation >X falling in Category A, and orientations <X belonging to Category B;

$$p(s|C=1) = \begin{cases} k, & \text{if } s < 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

$$p(s|C=2) = \begin{cases} k, & \text{if } s \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.7}$$

The log posterior ratio under the alternative generative model was therefore:

$$\frac{p(C=1|x)}{p(C=2|x)} = \frac{\Phi\left(\frac{-x}{\sigma}\right)}{1-\Phi\left(\frac{-x}{\sigma}\right)} \tag{3.8}$$

where Φ indicates the standard normal cumulative distribution function.

Given the assumption that stimulus orientations are corrupted by normally distributed noise, a probability distribution over the observer's measurements is inferred and, using the observer's decision rule, a probability distribution over confidence reports can then be computed. On a given trial in which an observer picks $C = 1$, under either model, an optimal confidence report would reflect the probability of a stimulus being drawn from $C = 1$ given the measurement x that the observer has made, $p(C = 1|x)$. Therefore, the relationship between confidence and the perceptual measurement x of stimulus S can be derived as

follows. Given that;

$$p(C = 1|x) + p(C = 2|x) = 1 \tag{3.9}$$

confidence reports should then reflect the following;

$$p(C = 1|x) = \frac{1}{1 + \left(\frac{p(C=1|x)}{p(C=2|x)}\right)^{-1}} \tag{3.10}$$

Whether or not the predicted confidence values under these two models differed across the stimulus space was investigated using model-based simulations. The results of these simulations are plotted in Fig.3.5. Note the difference in gradient and variability of confidence reports moving away from the simulated overlap of the categories which lies at the centre of the stimulus space (around the 0 point on the x-axis). The Bayesian Normative Generative model (left panel) has lower overall confidence reflecting the representation of the overlap of the categories which encompasses a level of uncertainty across much of the stimulus space. By comparison, the rule-based Alternative Generative model (right-hand panel) has a much wider spread of confidence reports including much higher confidence values being reported moving away from the rule-based boundary inserted in the overlap of the distributions (at 0). This model does not account for the categorical uncertainty that exists in the overlap, as it does not represent that overlap. This difference creates a distinct pattern of confid-

ence under each model in the proposed stimulus space. This means that, based on these simulations, there are clear qualitative differences in the predictions of the two model families being used. In the analysis of my MSc thesis, a model recovery exercise also indicated that the model predictions were distinct.

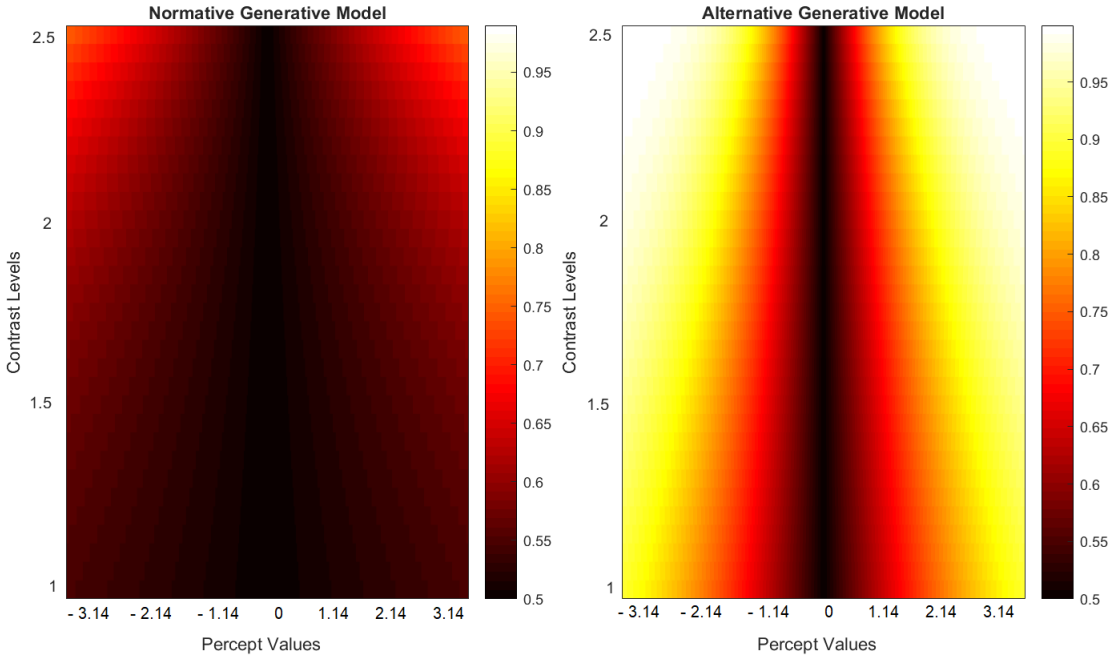


Figure 3.5: Model simulated confidence values as a function of contrast and percept value. The Bayesian Normative Generative Model is depicted on the left panel, the rule-based Alternative Generative Model is illustrated in the right panel. The perceptual values are mapped between -*pi* to *pi*. Higher contrast levels relate to greater perceptual certainty. Lighter colours indicate a higher confidence rating. Figure reproduced from Ashcroft-Jones, 2019 - unpublished MSc thesis.

The behavioural data were then examined to see if the experimental design had successfully elicited differences in the confidence reports of participants in line with the predicted differences in category representation across blocks. These data indicated that there were differences elicited by the two trial types,

hard and easy, in line with the predictions of the models Fig.3.6. In the easy simple trials, which were predicted to elicit a rule-based cognitive model of the task space, the confidence values showed a good variance of reports and reached high levels of confidence in the easiest trials (high contrast and far from the boundary point) - see the upper panel of Fig. 3.6. This, on average, is in line with the model simulations illustrated in the right-hand panel of Fig. 3.5. By comparison, the harder complex trials, which were predicted to elicit a Bayesian normative model of confidence, mapped onto the simulated confidence values with low and largely invariant confidence across the space. Again these data map well to the patterns simulated under the Bayesian Normative Model in Fig. 3.5. This suggests that the different blockwise trial types elicited confidence in line with the relevant predictions of the two model simulations and that the mapping was in the direction that was expected.

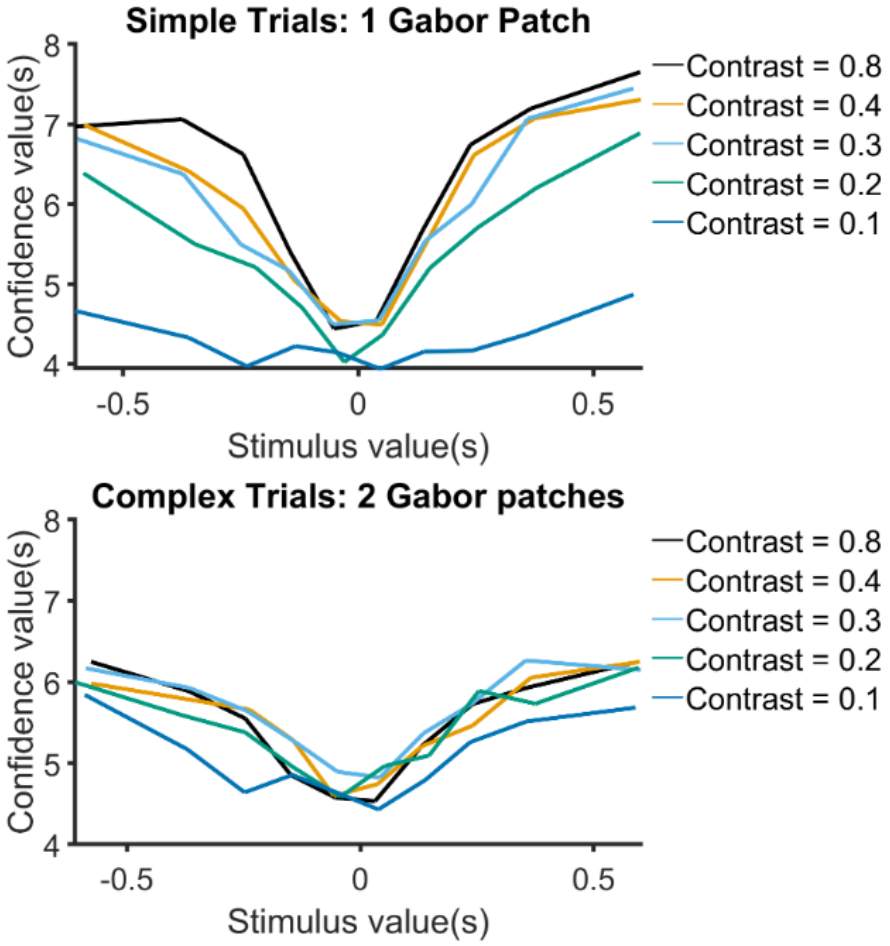


Figure 3.6: Confidence values (binned - the continuously measured confidence rating is replaced with a number from 1-10 according to the decile bin it falls in after pooling across all sessions) as a function of stimulus orientation at different levels of visual contrast on trials with 1 Gabor patch (simple - above) or 2 Gabor patches (complex - below) Darker colour lines indicate higher level of contrast(corresponding to lower perceptual uncertainty) Figure reproduced from Ashcroft-Jones, 2019 - unpublished MSc thesis.

The four models were then fit to the behavioural data using a custom likelihood function and the *fmincon* algorithm to generate a probability distribution of confidence reports which was divided into deciles within participants to match

the true participant data. The models had free parameters for a decision lapse, the thresholds of the 10 confidence bins, and a value for the perceptual noise internal to the observer. The likelihood of the actual confidence reported by participants trial by trial was then evaluated. However, despite both the simulation work, Fig. 3.5, the model recovery, and the behavioural data, Fig. 3.6, showing these distinguishable confidence patterns for the easy and hard versions of the task, the fit of these models were somewhat inconclusive, see the left panel of Fig. 3.7.

As part of the model fitting, within each participant, for all blocks simultaneously, the BIC and AIC values were calculated for all four models. As all models under analysis have the same number of parameters, differing levels of model flexibility are not present and the AIC and BIC values are equitable, Here, only the BIC results are reported for brevity. The aggregate BIC indicated the “winning” model - model 2 (Model 2 - Full Bayesian). This allowed the Δ BIC, the BIC of each model minus the BIC of the winning model, to be calculated within each participant. This process removed the within-individual variability of model fit. A boot-strapped 95% confidence interval was then calculated on the Δ BIC values. These confidence intervals must not straddle the zero point for the fit of the "winning" model to be reliably deemed superior to all other models tested. As seen in Fig. 3.7, this condition was not met: Although the Full-Bayes model (model 2) provided the best overall fit to the data; it was indistinguishable from the Full-Rule model (model 4). The error bar of the ΔBIC on the Full-Rule (model 4) just crossed 0, indicating that it

was not significantly different from the fit of the Full-Bayes model (model 2). Furthermore, the Full-Bayes model (model 2) was not the best fitting model for many participants, just 5 of 13 participants as compared with 4 of 13 who favoured the Full-Rule model (model 4) – see the right panel of Fig. 3.7.

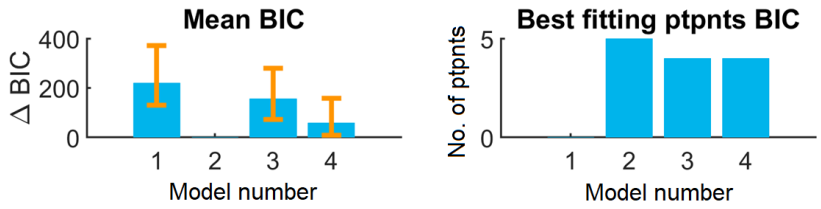


Figure 3.7: Fit of the four observer models to the behavioural data; the best fitting model was the Full-Bayes model overall (model 2 - a normative Bayesian model applied to all trial types). The left panel indicates the mean Δ BIC values of the model fits. Note the confidence interval of the Full-Rule model (model 4 - a rule-based model applied to all trials) which rests across the zero line indicating that it cannot be meaningfully distinguished from the Full-Bayes model (model 2). The right panel indicates the number of participants for which each model was the best fit. Note the Proposed model (model 1) did not fit any participants best, however, the spread of the other models across participants was extensive. Figure reproduced from Ashcroft-Jones, 2019 - unpublished MSc thesis.

Although inconclusive in terms of which model can be declared the "winning" model, these results do indicate two key points. Foremost, they are conclusive in terms of suggesting that Model 1, the hypothesised model for these data, was not a good fit to the data for any participants. Indeed, the two competing models, 2 and 4, are those models which stand in contrast to the hypothesised idea that people would use different strategies across the task conditions. Instead, these data suggest that people were internally consistent – being best fit by either the Full Bayesian or Full Rule models – but with different participants seeming to favour different strategies.

3.2 | Exploring the Original Models

Given these ambiguous fits, despite the clear sensitivity of confidence to the changes in contrast and difficulty and the qualitative differences predicted by the models, I was interested to explore these data and investigate the model fitting further. This work comprises two stages - an initial exploration of the original modelling approach and then a separate extension of the models which will be described in turn here.

This exploration is the focus of the remainder of this thesis chapter and the subsequent model refitting was a natural evolution of the original data analysis. The model refitting acts to address some of the limitations of the original approach, creating a more refined approach, which might allow me to distinguish the models more conclusively. This work is original to this DPhil thesis and acts as the initial steps of the subsequent secondary analysis of the original dataset.

Likelihood Simulations

To begin with, to see where the models differ in terms of their goodness of fit, I first conducted likelihood-based simulations of decision confidence under the four models, see GitHub for details. These were simulations that explicitly checked that the output of the likelihood function, when given fixed and plausible values of the parameters, is, qualitatively, in line with the model predictions. These simulations served two distinct purposes. Foremost, they acted as a valid-

ity check of the likelihood function used in the original model fitting, ensuring that there were no bugs in that code that could have led to erroneous model fits. This function was a custom-built function that computed the trial-by-trial likelihoods – it therefore was the crux of the model-fitting pipeline and its validity is paramount to the robustness of the model-fitting results. Secondly, these simulations enabled a local minima investigation which checked if the different models could reliably converge on a stable set of parameters as an indication of whether each model type can provide a reasonable fit to the data.

In order to conduct these likelihood simulations, the values from the best-fitting model (Full-Bayes model - model 2) were first used to help select some plausible parameter values against which to test the likelihood function. From the original model fitting, the fitted lapse rate and confidence threshold values were extracted for each participant. The fitted values for the variance assumed by the model under both the 1 and 2 Gabor patch block conditions were also extracted within participants. These values were then averaged across participants and were input into model-based simulations to create a simulated dataset that was computed based on plausible fitted parameter values. Using these trial values, the confidence predicted on each trial was simulated using the likelihood function. In this function, for each of the 10 confidence bins, the probability of a confidence report in that bin was calculated trial-by-trial. This entailed converting the likelihood thresholds into the stimulus space and then normalising by the response probabilities. These data then allowed for further local minima investigations and can be compared with the standard model predictions.

In terms of the validity of the likelihood function used in the model fitting, the confidence predictions of the likelihood function when given the true values of the parameters used to simulate the data show clear sensitivity to the number of Gabors and the contrast level of each trial, see Fig. 3.8. This suggests that the initial model fitting was free of major bugs or errors which could have led to erroneous results.

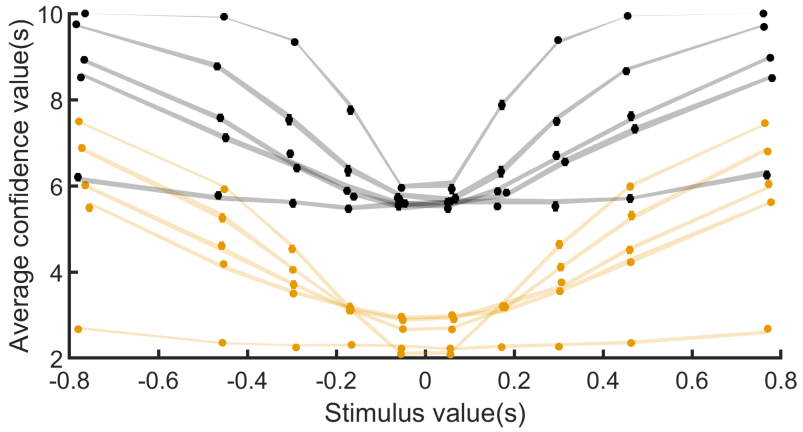


Figure 3.8: Confidence predictions of the likelihood function (filled data points) when supplied the true parameters used to simulate the data (shaded lines). The plot shows the estimated confidence at different contrast levels, and the different numbers of Gabors separately. The yellow lines indicate the two Gabor simulations, the grey indicates the one Gabor trials. These values are simulated under the Full-Bayes model (model 2) - the best-fitting model.

Based on the validity of the likelihood function, the stability of the model fits, based on their ability to reliably, and quickly, converge on a stable set of parameters, were explored in a local minima investigation. To begin with, across 10 runs of each model, within each participant, the number of starts required for the model to eventually converge successfully (within one log-likelihood of the best fit - model 2) was estimated via a bootstrapping procedure. For this

calculation, a bootstrapped sample of 10000 samples was used to calculate the probability of a "regret" smaller than 1. Where regret is defined as the difference between the actual best value of the log-likelihood found and the bootstrapped optimum (following the optimisation procedures of Acerbi *et al* (2018)).

If the number of restarts estimated was greater than the number of runs simulated (here N= 10) the number of restarts required was marked as Inf. Therefore, if the estimated number of starts was marked as infinity the model run required >10 restarts to successfully converge for that participant on that run – i.e., the behavioural data of that participant could not be reconciled with the model-predicted values under those parameters using the likelihood function within 1 log-likelihood of the best fit within 10 restarts.

See Fig. 3.9 for details of model starts. Note the lower average estimated start number for the Full-Rule model (model 4) compared with the other models. This could indicate that the Full-Rule model (model 4) did not struggle to fit the data and required less scoping of the parameter space to achieve a fit. From this, we could also suggest that the Full-Rule model (model 4) appeared to not get caught in local minima as much as other models.

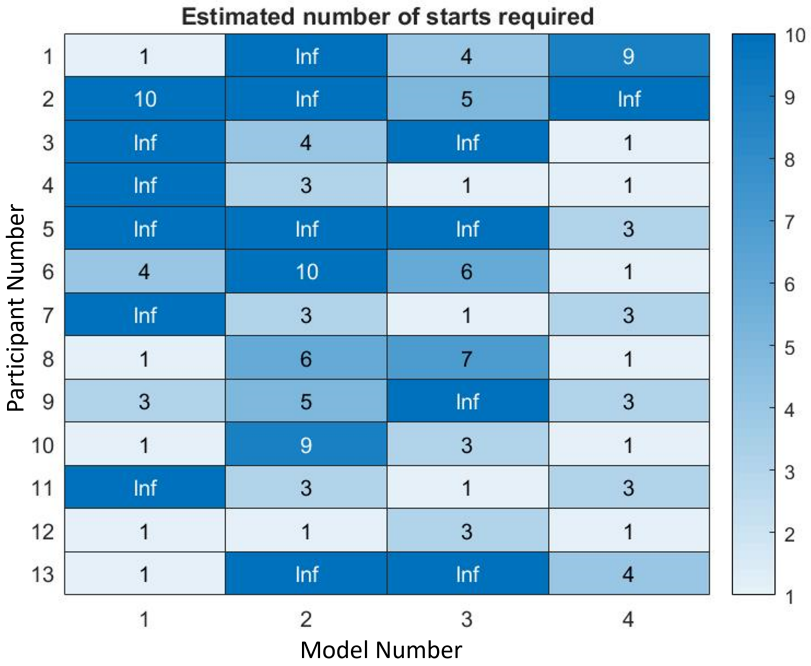


Figure 3.9: Across 10 runs of each of the models (1-4), within each participant (1-13), the estimated number of restarts required to achieve a successful model convergence within one log-likelihood of the best fit identified was calculated. If the number of restarts estimated was greater than the number of runs simulated (here N = 10) the number of restarts required was marked as Inf.

The success of these fits was then determined. A fit was considered a success if it lay within one log-likelihood of the maximum found. Across the 10 runs of each model, within each participant, the proportion of runs that resulted in success was calculated and recorded in Fig. 3.10. These data indicate that, when given plausible parameter values, the fit of the Full-Rule model (model 4) to the maximum was tighter across many participants compared with the Full-Bayes model (model 2).

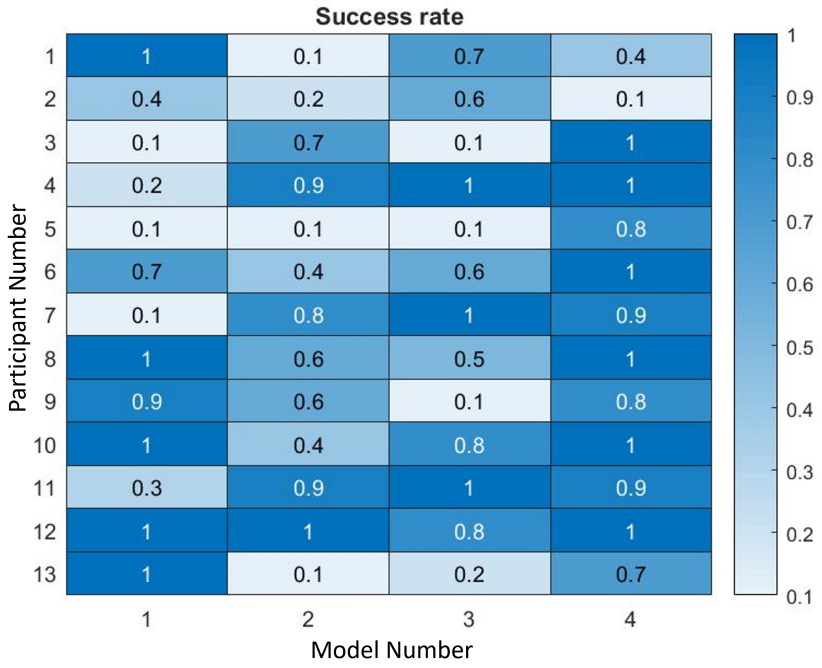


Figure 3.10: Across 10 runs of each of the models, within each participant, the proportion of model fits resulting in a fit within one log-likelihood of the maximum found was recorded. This proportion runs from 0-1 with a value of 0.7 indicating that 7/10 runs of that model within that participant were deemed a success.

Taken together these analyses indicated that the original model fitting results were not unreliable. They also suggest that the Full-Rule model (model 4) was a strong contender model despite the Full-Bayes model (model 2) being identified as the best-fitting model overall. There is a tension between these two analyses, the Δ BIC values and the local minima investigations, which could suggest that while Model 2 gives close fits when it does succeed, it may struggle to converge. By comparison, Model 4 may represent a reasonable approximation of the data which does not match the close fit of Model 2 but more often succeeds

in converging on a reasonable fit to the data. Another key pattern in these data is that some models seem clearly better than others for some participants (e.g., Model 1 for Participant 1, Model 4 for Participant 3), but for others, the fits are more ambiguous. This suggests that there are individual differences both in terms of the strategies being used by the participants in this task and also in the extent to which one model as compared with another better describes the approach of a participant.

Model-Based Simulations

Following on from this, given that the model fitting function appeared to be valid, and that the Full-Rule model (model 4) appeared to be a candidate of interest beyond the overlapping confidence interval it produced, I then conducted model-based simulations (see this GitHub Repository for details) and compared the model-simulated confidence reports to the behavioural data. This exercise aimed to simulate data under each model to understand how, in relation to the actual participant data, the different models are succeeding or failing to capture the patterns of confidence. This exercise investigated the behaviour of the model across the full scope of the stimulus space to explore deviations from the behavioural data which may explain the poor fitting of certain models, for example. Looking at these results, in Fig. 3.11, the Full-Rule model (model 4) is shown to systematically under-predict confidence for difficult stimuli near the category boundary (which is plotted at 0 on the x-axis). That is to say that the Full-Rule model predicts lower confidence at the point of overlap between

the two stimulus distributions than participants actually reported – see lower panel of Fig. 3.11. By contrast, the Full-Bayes model (upper panel of Fig. 3.11) does not under-predict confidence. However, the fit of this model is quite noisy – as indicated by the wide error shading for the model simulated values across the stimulus space. This means that the predictions from the Full-Bayes model are, perhaps, not specific enough to be discriminated from other models and are overly flexible in values computed.

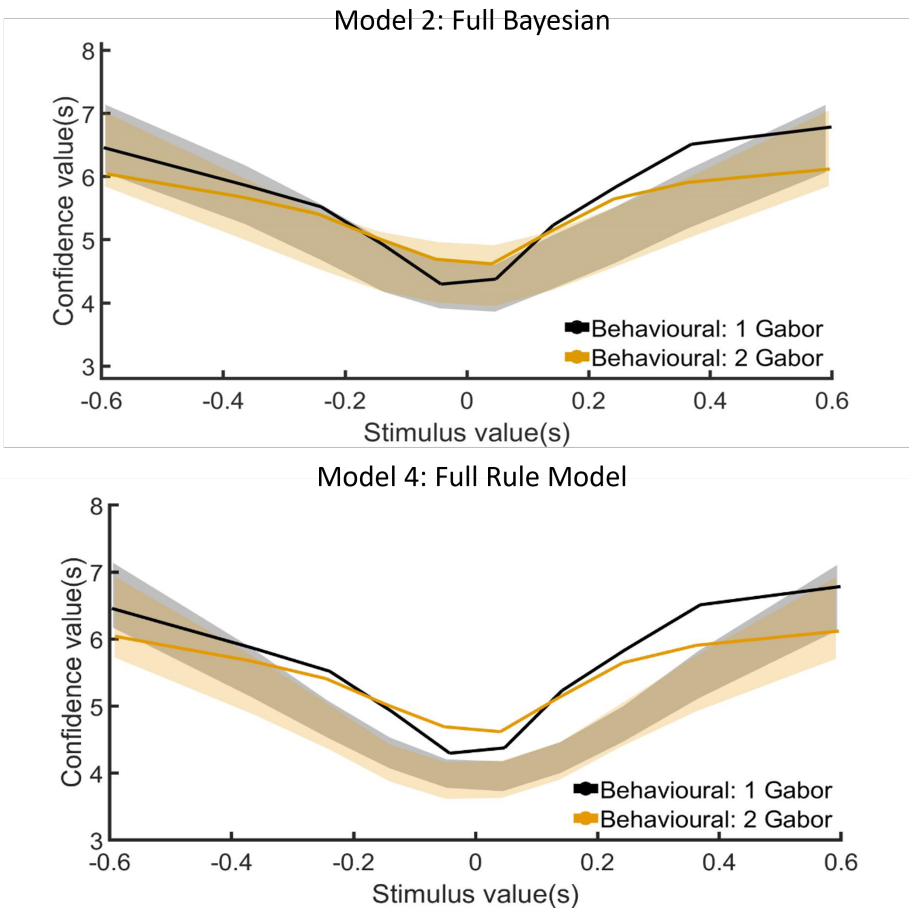


Figure 3.11: Fit of the Full-Bayes model, model 2 (above), and the Full-Rule model, model 4 (below), to the data, for confidence reports across stimulus values. Data are plotted separately for 1 Gabor (easy) and 2 Gabor (complex) trials. The model-based simulations are represented by the error-shading while the actual data are plotted as lines. Note the wider error on the model 2 simulation lines and the under-fitting of model 4 for confidence reports closer to the category bound at zero.

Taking this exploratory investigation one step further we can then look to differentiate between confidence predictions at different contrast levels across both hard and easy trials, see Fig. 3.12. As shown in the left-hand panels the one Gabor trials show greater differentiation of confidence across the different

contrast levels. Both models (upper panels - model 2, lower panels - model 4) under-predict confidence at the lowest contrast levels – that is to say, they under-predict confidence for the hardest trials in the one Gabor version of the task. In complex two Gabor trials (right-hand panels), the confidence reports are much less undifferentiated. As seen before, model 4 (lower panels) under-predicts confidence relative to model 2 (upper panels), with this distinction clearest in the two Gabor version of the task. As described above, confidence is less differentiated in these more complex trials, as seen by a distinct lack of separation of confidence across the different contrast levels. Of interest, however, is also the over-prediction of confidence in both models of confidence at the extremes of the stimulus space. While both models predict that the gradient of confidence continues to rise in fact the behavioural data deviates, and plateaus after a certain point. This shift in gradient is not apparent in either model but is particularly distinctive in the case of the Full-Bayes model (model 2).

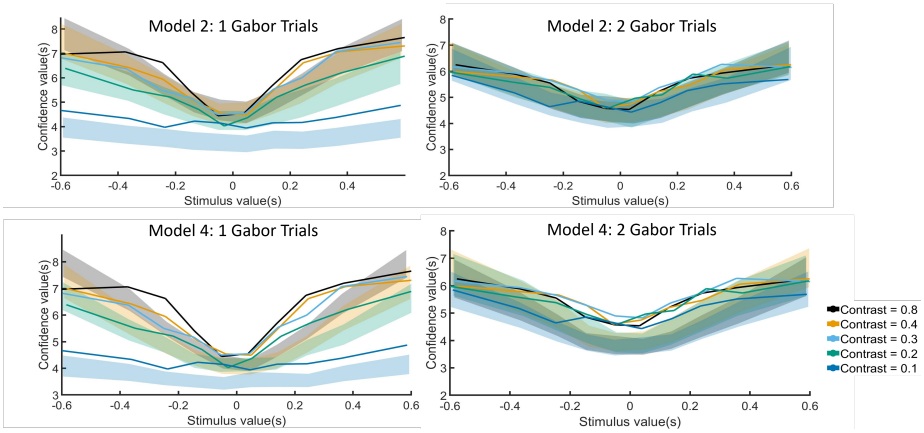


Figure 3.12: Fit of model 2 (above) and model 4 (below) to the data for simple (left) and complex (right) trials, for confidence reports across stimulus values. Data are plotted separately for each contrast level of the stimuli varying from 0.1-0.8. The model-based simulations are represented by error-shaded lines while the actual data are plotted as simple lines. Note the wider error on the model 2 simulation lines in the 2 Gabor case (upper right). Both models show the distinction between contrast in the one Gabor case (left), and both models under-predict confidence for the lowest contrast level for those trials. Conversely, both models match the clustering of confidence reports, across contrasts, in the two Gabor cases. As above, note the under-fitting of model 4 for confidence reports closer to the mid-line at zero.

Discussion

Collectively, these investigations suggested that the two best-fitting models were qualitatively distinguishable in their fits to the behavioural data. The likelihood simulations indicated that the model fitting was valid, and the model simulations indicated that the models made distinctive predictions about confidence at different points in the stimulus space relative to the behavioural data. For example, the Full-Bayes model appears to over-predict confidence for extreme values, while the Full-Rule model seems to under-predict confidence near the category bound. However, it appears that the original model fitting was not

stringent enough for this differentiation to be quantitatively delineated.

Therefore, stricter models were required to constrain the modelling space in order to distinguish these models analytically. Of course, it is also possible that different participants adopt different strategies – a finding that is hard to reliably evidence but could explain the spread of model fits and the variability in the local minima investigation success rates across participants.

3.3 | The Model Extension

Following on from the analyses above, a model extension was derived to try to distinguish the fits of the four models in the data more definitively. This extension aimed to restrict the model fits by requiring the models to fit both the response and the confidence simultaneously, and included extra parameters to better describe both the noise and the signal in the data. The original models accounted for the probability of a confidence report, given the stimulus and response of the observer. That is:

$$p(C|S,R) \tag{3.11}$$

To restrict the models further, the decision response can also be accounted for. Therefore, the extended version of the models must account for the probability of both the confidence report and the response, given the stimulus.

$$p(C,R|S) \tag{3.12}$$

As this constrains the models based on the response, response lapses were included to allow for the possibility of error in choice either due to inattention or motor response error, for example. Furthermore, accounting for the imperfect transfer of evidence to the mechanisms for confidence reporting, metacognitive noise was also included. Metacognitive noise can be described as random noise that reduces metacognitive efficiency but does not affect the perceptual

decision, it can affect either the signal of the information being used to compute confidence or the computation itself by, for example, jittering the criteria for confidence ratings (Shekhar & Rahnev, 2021). The relationship between these variables is shown in more detail in Fig. 3.13.

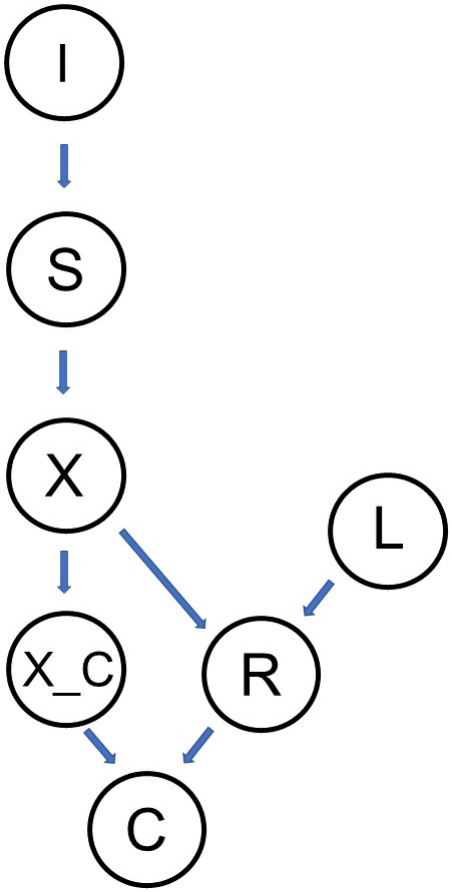


Figure 3.13: Extended Bayesian generative model of observer decision and confidence judgements. I indicates the category identity, from which a stimulus S is drawn. The observer takes a noisy measurement of this stimulus value to produce a measurement X. X informs the response, R, which may be corrupted by a lapse, L. X is then affected by metacognitive noise to produce X_c. Both X_c and R inform on the final confidence judgement, C.

This extended model was established in this way for two reasons. First,

by forcing the models to account for both the response and confidence report, they are restricted more stringently by the behavioural trends themselves. This aimed to allow the models to be distinguished more clearly if one model better accounts for the behavioural data. Given the above-described patterns of distinct deviations of the models from the behavioural data, the need for the models to account for more data points (both the decision response and the confidence report for each trial) should allow the models to be better distinguished – particularly in the case of the normative Bayesian model predictions which were overly noisy in the original model.

Secondly, by explicitly including lapse and metacognitive noise terms, the models should be able to account better for responses, and confidence, at the extremes in the data both very near to and far from the category bound where they previously struggled. Importantly, a second and independent lapse term was established on confidence reports to allow for the possibility that confidence reporting errors could occur. Indeed, the inclusion of these parameters allows the models to have less exacting predictions of confidence for more extreme stimulus values. Thus, increasing the chance that the parameter estimates and confidence predictions will be more plausible. The inclusion of metacognitive noise also aligns these computations with the current understanding of confidence as a noisy readout of the observer's cognitive state (Guggenmos, 2022; Shekhar & Rahnev, 2021).

Simulating Model Likelihoods with IBS

Due to the complexity of the new model fitting – with the requirement for the models to predict both response and confidence reports and the corresponding increased number of parameters – full mathematical derivations became impractical. Importantly, the confidence function and decision rules were matched to the original models, with the addition of the relevant noise and lapse parameters, and the assumed relationship between confidence and the log-likelihood, was maintained. This means that the broad structure of the models, and their contrasting predictions, remain intact relative to the derivations described in the original models. However, calculating a full likelihood function for $p(C, R|S)$ was untenable.

Therefore, an alternative method was adopted - complete "slow" simulations of each individual trial, (see the relevant GitHub repository here). These simulations were conducted using Inverse Binomial Sampling (IBS; (van Opheusden et al., 2020)) in MATLAB (MATLAB R2020b) to simulate the log-likelihood on each trial. IBS estimates the log-likelihood of a model in an efficient and unbiased manner (van Opheusden et al., 2020). Thus, IBS is adept at providing likelihood-based inferences when analytic evaluation is not feasible.

In this analysis, the IBS functions were adapted to take the actual participant data for both response and confidence from the dataset. They were then used to simulate the same pattern of responses using both the experimental design and trial-level stimulus information as described by custom-built,

model-specific simulation functions. These functions map out the path of the data through the expected pathway from perception to reporting, following the structure described in Fig. 3.13.

To enable IBS to converge more effectively, the confidence report bins were reduced from 10 to 4 in both the model and the original behavioural data. This was necessary because IBS works by re-simulating the response and confidence reports until they match the data trial-by-trial. Thus, IBS outputs the log-likelihood of these response patterns under a specified model. However, when the model attempts to match the behavioural data across 10 confidence bins the specificity required in the output of the simulations slowed the modelling procedure exponentially. To understand this better, imagine that an individual made an error in their choice but was highly confident about it, and used the highest bin for their report. This is a difficult case for IBS to solve, and it is almost impossible without lapse parameters included in the model on both the decision and confidence. In this case, the IBS functions would likely have to re-sample and re-simulate the trial numerous times to achieve that matched output. The longer it takes for the IBS sampler to converge on the behavioural data for a particular trial, the lower the likelihood of the generating model on that trial. If IBS fails to match trials after several thousand iterations, which can occur in these outlier trials, it will give a convergence error and the sampler will stall. However, by reducing the number of bins that IBS has to match to, although it may still struggle with mismatched errors and high confidence, for example, it will manage to match the confidence reports with greater success

than if there were 10 bins to match across while still illustrating a variability of confidence level in the dataset.

The addition of the lapse rates was, therefore, not only a theoretical addition to the models in this extension but also became a pragmatic addition based on the methodology. Without these parameters being included in the model, giving the possibility that for example a seemingly random confidence bin could be reported for a given trial, the IBS sampler would struggle on specific trials which may not hold any meaningful signal in their distinguishing features due to their random and erroneous nature.

Model Fitting with BADS

I optimised the model fitting procedure using the complementary Bayesian Adaptive Direct Search (BADS; (Acerbi & Ma, 2017)). BADS optimises the search of the parameter space and the log-likelihood between a series of fast, local Bayesian optimisation steps and a systematic, slower exploration of a mesh grid. BADS utilises a mesh-adapted directed search that moves between polling steps and search phases in an alternate manner. When polling, the mesh points are evaluated by taking steps in one direction at a time, and doubling the step size when an improvement is successfully noted. Unsuccessful steps halve the distance on the next iteration. In the search stage, based on the points that have been evaluated, a Gaussian process is fit. This allows BADS to select the next points for evaluation trading off between exploration of uncertain regions and exploiting known and promising regions. In this way, as a semi-local op-

timiser, the BADS search strategy is efficient and can often effectively escape local minima.

BADS takes input, in the form of a function handle calling to IBS, a vector of parameter start-points as well as plausible (PLB, PUB) and hard upper and lower boundaries (UB, LB) for each parameter. BADS then iterates through the IBS simulations at each step while moving efficiently through the plausible parameter space to find the maximum likelihood. A wide-ranging parameter space was chosen to prevent over-fitting of the model to the data by not constraining the search space to an area occupied only by the behavioural patterns in the data. The parameters to be fitted and the values of the bounds set on the parameter spaces are shown below, see Table 3.1. Details of each parameter are outlined in the next section.

Free Parameter Initialisation & Bounds			
Parameter	Start Point	Upper | Lower Bounds	Plausible Bounds
$Sigma_X$	Random [PLB:PUB]	pi - $pi/1000$	pi - $pi/100$
$Thresholds$	Random [PLB:PUB]	$0.99 - 0.01$	$0.9 - 0.1$
$LapseRate$	Random [PLB:PUB]	$1 - 0.01$	$0.45 - 0.01$
$MetaNoise$	Random [PLB:PUB]	$4 - 0.0001$	$2 - 0.0087$
$ConfLapse$	Random [PLB:PUB]	$1 - 0.0001$	$0.1 - 0.01$

Table 3.1: Parameter space established for BADS to search for the maximum likelihood.

The Parameter Space & Fitting

The models were comprised of both fixed and free parameters. This section will first outline the fixed parameters before describing the free parameters and the search space that BADS was allowed to poll from and grid search in relation to these values.

The fixed parameters included the means of the distributions, the trial-by-trial experimenter imposed contrast levels, a neutral prior on the decision rule, and a fixed noise level for the perceptual process. These will all be described in more detail here. The means of the two distributions, μ, corresponding to the average deviation from the horizontal for each stimulus category, were set symmetrically on either side of π in the circle geometry of possible stimulus orientations. Trial-by-trial contrasts on the Gabor patches could be assigned a value of 0.1, 0.2, 0.3, 0.4, or 0.8 – with lower values indicating more difficult-to-discriminate stimuli. These contrasts represented external, experimenter-controlled noise in the perceptual process. A neutral prior was set on the decision process such that there was no bias towards one response over the other, as was assumed by the experimental design and the behavioural data. Finally, a fixed level of noise was assumed to apply within the visual system which was denoted by a fixed parameter of σ of a value $\sqrt{1/7}$, where 7 was the value of κ the concentration parameter of the vonMises distribution used to produce the orientations for the experimental trials. This fixed parameter of noise was separate from a free parameter for σ which was allowed to vary across trials depending on the stimulus properties.

Next will follow a brief description of the free parameters including the σ values that were not fixed and will deep dive into the values displayed in Table 3.1. The σ_x parameter represents perceptual noise in the observer's cognitive system. Thus, it represents uncertainty at the level of the stimulus percept. The value of this parameter was multiplied by a sample from a random normal

distribution with a mean of one and is then added to the objective stimulus orientation. Therefore, this parameter acts to jitter the perceived orientation away from the actual value to a greater or lesser extent depending on the parameter value itself. As the original experimental data was generated under five different levels of perceptual contrast which could directly influence the perceptual noise trial by trial, and given that there were two trial difficulty levels based on the number of Gabors on screen, a total of 10 sigma values were fit per participants, per model, one for each trial type. It was expected that these values would maintain an internal consistency based on these trial dimensions, with more noisy and difficult trials relating to a higher value level of perceptual noise. However, while the structure of trial information and how it relates to the sigma values were explicitly stated within the modelling process, with particular Gabor numbers and contrast levels being indexed to certain sigma values, the noise levels themselves were not required to be ordered in any way.

The *Thresholds* parameter relates to the placement of the confidence bin thresholds in the confidence space. In this way, it differentiates what values of confidence sit in bin 1 versus bin 2, bin 2 versus bin 3, and so on. These parameter values were generated before being ordered and applied to the simulated confidence reports. Thus, the confidence being fitted by the models are the binned confidence values in the range of bins 1-4. These bins were established as quartiles of the data within each individual allowing for across-participant comparison to be conducted. In order to fit confidence into four separate bins, 3 thresholds were fit, which were later bookended by 0 and 1. For example, if

an individual, under a particular model, divided the meta-cognitive confidence space evenly, those thresholds would be fit at 0.25, 0.5 and 0.75, respectively. This would create four confidence bins the lowest between 0-0.25, and the highest between 0.75-1.

The *Lapse Rate* parameter is used to represent the lapses that occurred on the decision itself. This parameter does not necessitate that the trial is an error but assumes a lapse in attention such that the participant chooses randomly on that trial. Therefore, on a lapsed trial, the choice could be correct or could be made in error. The lapse rate is a value between 0.00001-1. It was implemented as follows: for each decision trial, a random number generator generates a number between $0-1$. If the number generated on that trial is greater than the preset "lapse rate" then the decision will not be a lapse, otherwise, the trial is labelled as a lapse and the resulting decision is a random one that is set using a random integer generator. In this way, the higher the "lapse rate" the more likely a trial is to be labelled as and generated according to the lapse procedure. Therefore, for each participant, under each model, a single lapse rate value was fit.

The $Meta\ Noise$ parameter relates to the presence of metacognitive noise which is applied to the confidence reports. Much like the perceptual noise parameter, this acts as a jitter which offsets the internal representation of evidence strength that guides the confidence report (x_c) from the representation of the evidence used to guide the original decision (x). Importantly, while x_c reflects a representation of the strength of evidence in favour of a decision, it does so

imperfectly as the value of x is itself imprecise. Therefore, even before noise is applied to this value it does not represent the objective likelihood, but rather the subjectively-evaluated likelihood according to the process generating the original decision.

The metacognitive noise parameter is applied before binning occurs such that smaller levels of metacognitive noise would only affect the borderline cases close to the confidence bin thresholds. By comparison, a large distortion of confidence, due to a high level of metacognitive noise, would see confidence binning more generally being affected. Metacognitive noise is added to the posterior ratio of the trials before it is re-calibrated into the 0-1 confidence space. For each participant, under each model, one value of metacognitive noise was fit.

Finally, $Conf\ Lapse$ is a parameter that relates to a confidence report lapse. This parameter is applied similarly to the decision lapse. The lapses are inserted at the confidence binning level, with a lapse resulting in a random confidence bin being chosen. This was to ensure that – unlike metacognitive noise – there is a clear chance that a lapse would change the confidence bins submitted to the model. This reflects a mirroring of the decision lapse parameter which is thought to be a lapse of attention during reporting such that the input value is random and not connected to the actual trial information. Thus, applying the confidence lapse to the confidence bins allows the sense of input error to be more clearly delineated from the distortions caused by metacognitive noise. The confidence lapse and decision lapse were not dependent – they could occur on the same trial, but the simulations and the generative models themselves did

not necessitate this co-occurrence. Equally, the occurrence of a decision lapse did not affect the probability of a confidence lapse under these models. Similar to the fitting of the decision lapse rate, for each participant, under each model a single confidence lapse value was fit.

For each participant and each model, over multiple repetitions, BADS outputs a maximum likelihood of the responses given the data and inferred parameter values. Across these repetitions, BADS iteratively adjusts the parameter values away from the start values but within the bounds established, to reach a set of parameter values that allows IBS to converge most effectively in line with the behavioural data. As IBS functions as an iterative simulation of the log-likelihood, the simulations were optimised, reducing the IBS run time from approximately 6 minutes to sub-6 seconds per model per participant. This allowed IBS to converge more quickly and in turn, improved the computational load of BADS significantly.

The model fitting procedure was set to run ten times per participant per model. A try count was established to see if, and how many times, a model failed to converge on each iteration. Failure to converge would be suggestive of a poor fit of the model to the data such that any subsequent successful convergence would likely be the product of extreme parameter start points, rather than a meaningful fit of the model to the behavioural data. Each trial was estimated 10 times per participant per model with a maximum iteration of 10^6 re-samples per trial estimate within which the successful estimation could occur. These settings gave scope for IBS to work through "difficult" trials where the response

accuracy and confidence were discordant but still ensured iterative multi-fidelity by averaging across the 10 estimates per trial.

As before, 4 models are being fit. As with the original model fitting, these are a normative Bayesian model for complex trials and a rule-based model for simple trials (Model 1 - Proposed Model), and the converse (Model 3 - Rev-proposed Model), as well as a normative Bayesian model (Model 2 - Full-Bayes Model) and a rule-based model (Model 4 - Full-Rule Model).

Model Fitting Results

The output of the BADS and IBS model fitting procedure was in the form of 16 fitted parameter values and one log-likelihood of the model fit for that participant per run per model per participant. The try counts were first checked to ensure that no model-fitting attempts had failed to converge. The results of the extended model fitting are displayed in Fig. 3.14. The Full-Bayes Model (Model 2) was once again identified as the best-fitting model numerically but was once again indistinguishable from the Full-Rule Model (Model 4). However, of note is the replicated failure of Model 1 to adequately capture the patterns of the data despite it being the hypothesised best fit to these data.

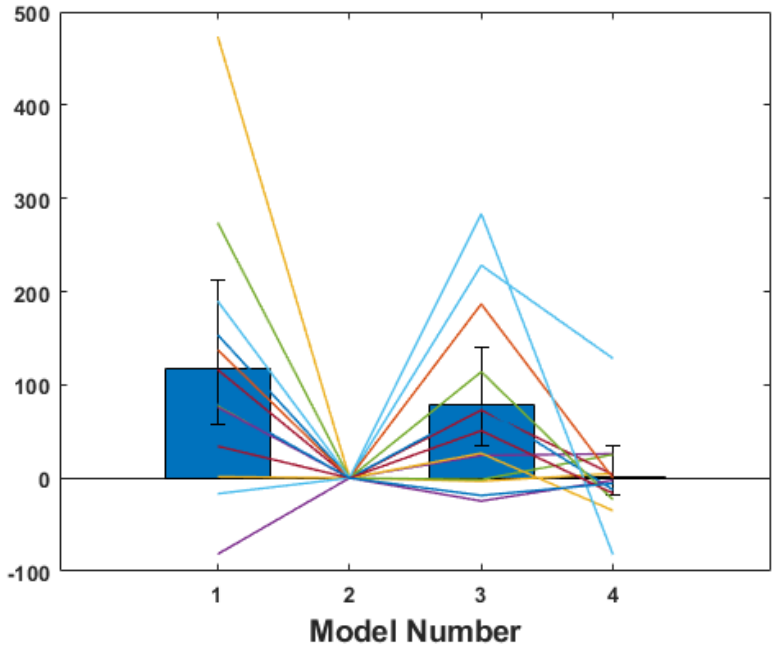


Figure 3.14: The comparative fit of the four models is shown for the extended model simulations. The y-axis indicates the Δ BIC of the models, with lower values indicating a better fit. Models are numbered 1-4 with the Full-Bayes Model (Model 2) being identified as the best fit but once again being indistinguishable from the Full-Rule Model (Model 4), based on the error bar overlapping the zero mark. Individual coloured lines indicate the fits of the alternative models, per participant, relative to the best fitting model, the Full-Bayes Model (Model 2).

Individual best-fitting models, per participant, are shown in Fig. 3.15. Here, the Full-Rule Model (Model 4) appears to be the numerically stronger model with most participants fitting this model best. Almost half of the participants had the Full-Rule Model (Model 4) as their best-fitting model. By contrast, the Full-Bayes Model (Model 2) was only the best-fitting model for 3 participants. While this may seem counter-intuitive given that it was the overall best-fitting model at the group level, the Full-Bayes Model (Model 2) may still

be a stronger fit on average by being the second best-fitting model for the majority of participants without giving a particularly close fit to any individual dataset. Perhaps the clearest result remains that participants appear to maintain a single model of confidence in the task, rather than favour a model which switches representations on different sessions based on task difficulty.

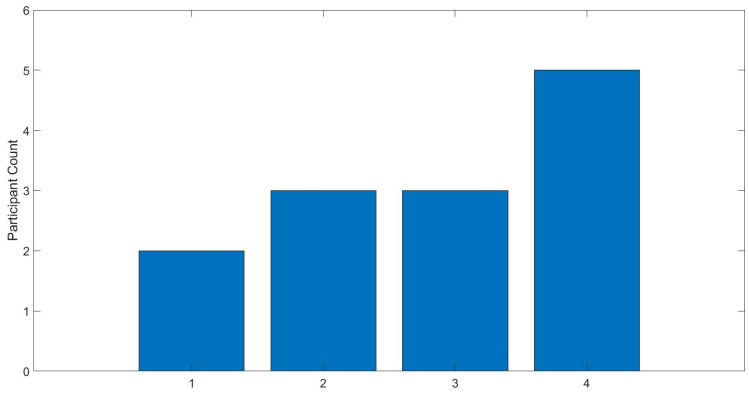


Figure 3.15: Models 1 to 4 are listed on the x-axis, the y-axis is a count of participants for whom that model was the best fitting model. The count of the best-fitting model per participant indicates that the Full-Rule Model (Model 4) was the most widely best-fitting model across participants. By comparison, the Full-Bayes Model (Model 2) was only the best-fitting model for 3 participants.

Following on from these results, to understand the behaviour of the Full-Bayes and Full-Rule models and to see how they might be distinguished, or not, the models were simulated and the model simulations were compared with the behavioural data itself. These investigations explored the responses and confidence of the two competing best-fitting models. When exploring the binned confidence values predicted by the two competing models some key differences can be illustrated, see Fig. 3.16. The fit of the two best-fitting models is depicted

by the shaded error bars over the line plots of the behavioural data. Although both models over-predict confidence in the hardest trials (where the contrast is lowest – see the blue lines in both panels) some specific differentiating features can be observed. Foremost, the values predicted by the Full-Rule Model (Model 4) seem to better match the different contrast levels. This is seen in the way that the model predictions separate out across the stimulus space according to contrast, mirroring the patterns of behavioural data. This is not observed as strongly in the simulations under the Full-Bayes Model (Model 2) which cluster together more strongly except for the trials with the lowest contrast. In particular, note the poor fit of the slope of the Full-Bayes Model (Model 2) which simultaneously over-predicts confidence at both the overlap of the categories (at the zero point of the x-axis) and the category extremes, while under-predicting confidence in the middle of the stimulus space. By comparison, not only does the Full-Rule Model (Model 4) differentiate between the contrast levels better, but it also seems to trace the slope of confidence more accurately. This can be most clearly observed in the plateauing of confidence at the extremes of the data, in line with the behavioural trends.

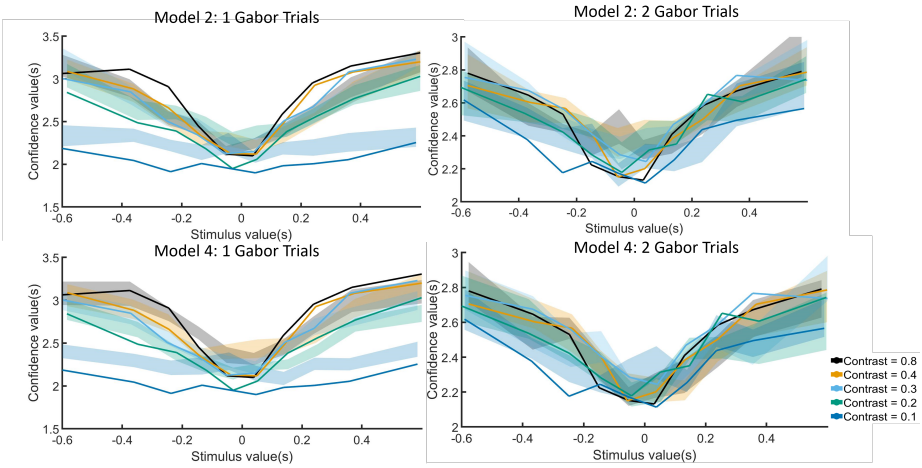


Figure 3.16: Fit of model 2 (above) and model 4 (below) to the data for simple (left) and complex (right) trials, for confidence reports across stimulus values. Data are plotted separately for each contrast level of the stimuli varying from 0.1-0.8. The model-based simulations are represented by error-shaded lines while the actual data are plotted as simple lines. Note the wider error on the model 2 simulation lines in the 2 Gabor case (upper right). Both models show the distinction between contrast in the one Gabor case (left), and both models under-predict confidence for the lowest contrast level for those trials. Conversely, both models match the clustering of confidence reports, across contrasts, in the two Gabor cases. As above, note the under-fitting of model 4 for confidence reports closer to the mid-line at zero.

As noted above, a key new feature of these models, compared to those in my MSc project, was the modelling of response choices themselves. Therefore, the fit of the models can also be evaluated in relation to response patterns (i.e., participants' decisions) as well as their reported confidence. Indeed, when the model simulations were used to explore the response patterns of the two competing best-fitting models, a different pattern was observed, see Fig. 3.17. As before, in this figure, the fit of the two best-fitting models is depicted by the shaded error bars over the line plots of the behavioural data. the y-axis depicts the response probability for one choice or the other, as we move across

the stimulus space on the x-axis the probability of choosing one response over the other flips. This flip point equates to the 0.5 mark on the y-axis, or chance, and coincides with the 0 mark on the x-axis where evidence for each choice is equal as the categories overlap fully. In terms of these response data, the Full-Bayes Model (Model 2) appears to better predict the behavioural data as compared with the Full-Rule Model (Model 4). In contrast to the relatively tight fit of the Full-Bayes Model (Model 2), the Full-Rule Model (Model 4) fails to map the response curve at the extremes of the data. This is seen most clearly in the lowest level of contrast (the hardest trials) demarcated by the blue lines. The simulated values from the Full-Rule Model (Model 4) suggest that the participants should be more definitive in their choices at these points with a stronger trend towards one choice over the other as compared with the behavioural data.

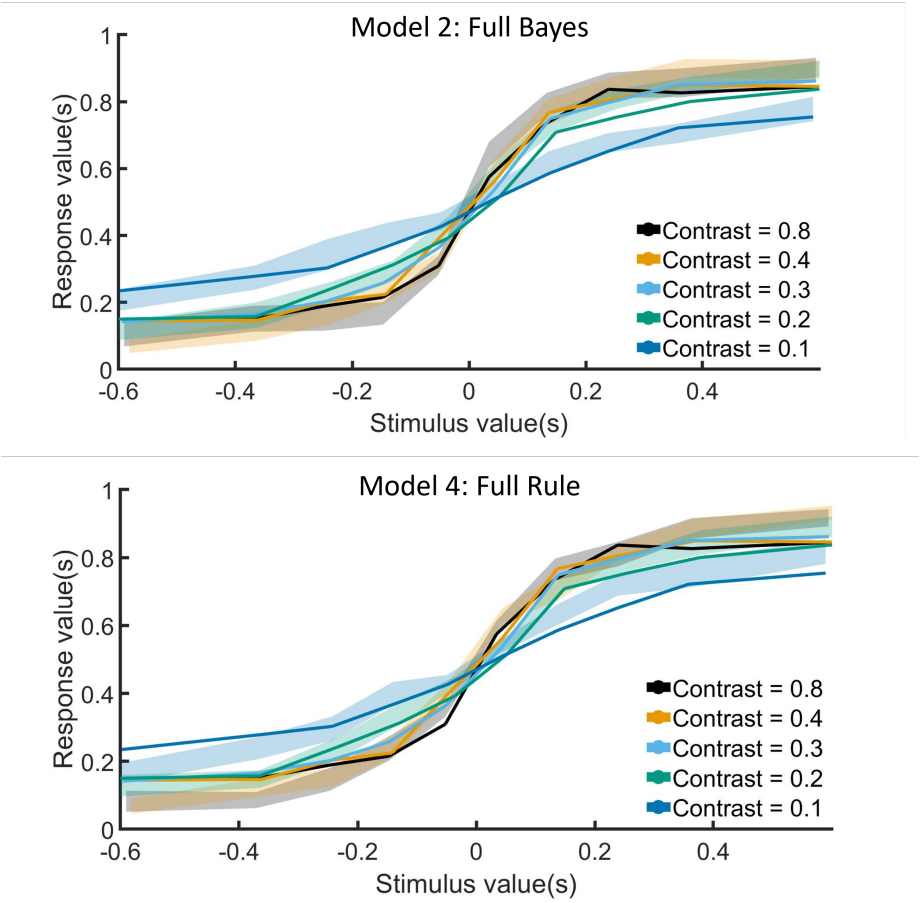


Figure 3.17: The response probability across the stimulus space as predicted by the Full-Bayes Model (Model 2 - upper panel) and the Full-Rule Model (Model 4 - lower panel) are plotted via the shaded error bars. The comparative response data from the behavioural data is denoted by the thin dark lines. The y-axis denotes the probability of one response over the other with 0.5 acting as the decision flip point between the categories. Data are plotted separately for the 5 levels of contrast that vary trial by trial. Note the more accurate fit of the Full-Bayes Model (Model 2) to the behavioural data in the response data.

Given that the model fits remained ambiguous despite the more extensive modelling, reasons for this trend were explored. As the clearest result remains that participants appear to maintain a single model of confidence in the task,

rather than switch representations with changes in task demands (which occurred across different data collection sessions on different days), the possibility of an order-effects in the data were explored. Each participant completed four sessions which had either an A-B-A-B or a B-A-B-A structure. The order in which the participants saw the data was randomised. Here A and B denote the two levels of difficulty of the task, the easy level (A) had a single Gabor patch on screen, and the hard version (B) had a second distractor Gabor on screen to increase the difficulty and dimensionality of the choice. The original prediction was that participants would leverage a different model of confidence in each version of this task and switch from a Full-Bayes model to a Full-Rule model when the task went from hard to easy (and the reverse) – this was Model 1 – the Proposed Model. However, this model (and the Rev-proposed Model – Model 3) were consistently poor fits to the data. It is plausible that the split of model fit between the Full-Bayes Model (Model 2) and the Full-Rule Model (Model 4) seen in both the original and extended model fit process was the product of model perseverance across these sessions. As such, the first type of session (either easy or hard) that a participant was exposed to may have established which model (Full-Rule or Full-Bayes) they used for that session and all subsequent sessions regardless of changes of difficulty.

To investigate this possibility, the dataset was split based on the first session type for each participant and the model simulations were conducted relative to those data subsets. Of the 13 participants, six saw the easy version of the task first which was designed to elicit a decision-confidence response pattern that

aligned with the Full-Rule Model (Model 4), the remaining seven participants were exposed to the more complex version of the task first which was designed to elicit a confidence model aligned with the Full-Bayes Model (Model 2). If the order effect was absolute then the model simulated values for the Full-Rule Model (Model 4) would be a good fit to the "easy first" sub-group of participants, while the model simulated values for the Full-Bayes Model (Model 2) would best fit the "hard first" subgroup of participants. Figure 3.18 illustrates these data and simulations.

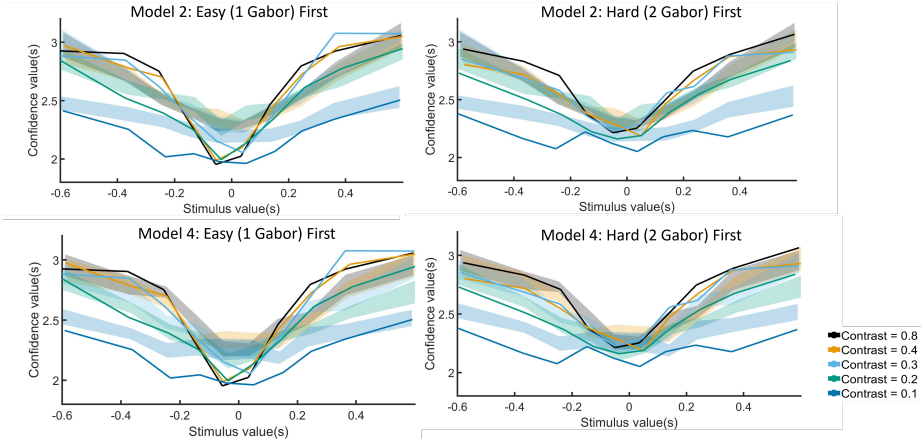


Figure 3.18: The binned confidence values predicted by the Full-Bayes Model (Model 2 - above) and the Full-Rule Model (Model 4 - below) for the easy first subgroup (left panels) versus the hard first subgroup (right panels) are plotted via the shaded error bars. The comparative confidence from the subgroup behavioural data is denoted by the thin dark lines. Data are plotted separately for the 5 levels of contrast that vary trial by trial.

While neither model simulates the behavioural data perfectly as an absolute order effect might predict, note in Fig. 3.18 that the slope of confidence simulated from the Full-Bayes Model (Model 2) is a better fit for the hard (right panel) than the easy (left panel) first subgroup of participants. The differences

between lower panels in Fig. 3.18 are less pronounced – which may relate to the overall better fit of the Full-Rule Model (Model 4) to the confidence as seen in Fig. 3.16.

Looking at the best-fitting model per participant, in the case of the best-fitting models within participants from the "easy first" subgroup, the Full-Rule Model (Model 4) was numerically the winner with 4 of the 6 participants having the Full-Rule Model (Model 4). The remaining 2 participants had the Proposed Model (Model 1) as their best-fitting model across all sessions. By contrast, among the seven participants who experienced the "hard first" condition ordering, the results were more mixed. One participant favoured the Full-Bayes Model (Model 2) as would have been predicted. Two more had the Full-Rule Model (Model 4) as their best fit, and the remaining 4 participants from the "hard first" favoured the Rev-proposed Model (Model 3). Taken together this could be summarised to suggest that when participants saw the easy version of the task first they broadly adopted the Full-Rule Model (Model 4) across all trials while starting with the hard condition meant that participants lacked a single approach to the task – perhaps due to its complexity level which may have prevented the participants from engaging properly with the subsequent sessions as well.

A further exploratory model fitting was conducted on the two order-based subsets of the data separately but the results remained inconclusive between the Full-Bayes Model (Model 2) and the Full-Rule Model (Model 4) in both cases. This could be due because the subgroups are still not appropriately described

by the different models, but it is equally possible that by halving the number of trials in each model fit the power is reduced to such an extent that any true signal cannot be detected.

3.4 | General Discussion

This chapter aimed to model confidence in a simple perceptual categorisation task and to compare a simple rule-based model with a Bayesian model that accounts for detailed features of the underlying category distributions. In order to achieve this, I advanced my prior work by first exploring the original model fits, and then employing more developing an approach to simultaneously model confidence and decisions, using IBS and BADS to enable an efficient search of a large parameter space.

The results presented in this chapter are not clear-cut in terms of favouring one model or theory, either Bayesian or rule-based, over another. What is apparent is that the original hypothesis surrounding model switching under different task demands (based in the work of Wills & Mackintosh (1998)) did not emerge – at least not under the task demands that were established here. The task varied in difficulty across testing sessions with either a single Gabor version (the easy stimuli) or a two Gabor version (the hard stimuli) where one Gabor acted as a constant visual distractor. A consistent theme emerging was that the Proposed Model (Model 1) was not competitive in the comparative fitting of either the original or extended model versions. Instead, participants' behaviour was best captured by a mix of different strategies across individuals – namely the consistent use of either a Full-Bayes (Model 2) or a Full-Rule (Model 4) model throughout all sessions by participants. Seemingly, there might not be a single dominant strategy in this task and different participants may have employed

different but internally consistent task representations to navigate the space. Indeed, the fit of these models was divided across participants and may have roots in order effect based on the session order experienced by the participants. Yet, despite the ambiguity of the results, several key lessons and insights can be gleaned from the exploration of the model fits and model simulated behaviour.

Foremost it is important to acknowledge the overarching similarity of the results of the extended model fitting as compared with the original model fitting. This in itself is interesting for several reasons. In particular, it is interesting that the comparative model fits remained stable across model fitting procedures despite the extension accounting for both the response itself, response lapses, and confidence (and confidence lapses). It was assumed that the inclusion of the response-based data would have constrained the model fits further and provided a challenge for the models to simulate against. The fact that this did not distinguish the models better than the original approach suggests that the original strategy of fitting to confidence alone is a sensible one. This could be argued to be because confidence is such a rich and continuous signal. It is richer than the response data which is a simple binary outcome per trial. Therefore, this similarity of fit despite the inclusion of responses suggests that confidence can alone be a helpful window into the mechanisms responsible for behaviours. This is a vital assumption that underpins all arguments surrounding the validity of metacognitive reports and plays out well here (Adler & Ma, 2018; Aitchison et al., 2015; Navajas et al., 2017).

Secondly, there is strong evidence for substantial variability across parti-

cipants across the two sets of model fits explored here, with some participants showing a pattern that aligned with the predictions of the rule-based model and others having confidence and response patterns that suggested the Bayesian model was a better fit to the data. This may suggest that human decision-makers utilise highly idiosyncratic concept representations, or that participants use confidence reports differently (Navajas et al., 2017). This idea is counter to the common practice of establishing the "winning" models across all participants. Indeed, taking the mean of the BIC values across participants under each model, as is conducted as part of the analyses here, makes the implicit assumption that all participants are using the same model. This is not an assumption that should necessarily be made here. Alternative group statistics such as exceedance probability could be a useful secondary analysis route (Piray et al., 2019; Stephan et al., 2009).

Another possibility of course is that there is a single true model but that the variability is simply due to noise, this cannot be ruled out based on the current data particularly given the difficult and un-intuitive nature of the harder version of the task. This noise may also arise in our data due to people's category representations evolving. This is a strong possibility given that feedback was provided trial-by-trial which could have led to continual changes to the representations that people used. This may also explain why the best-fitting model per participant was more mixed for those participants who viewed the complex task version first. In the "hard first" participants, it is arguable that learning would have occurred in later sessions as the first session may have been too

difficult to conceptualise without first experiencing the easy task version. The current model fitting cannot capture the potential for these dynamic changes as the parameters and models more broadly are assumed to be static over time.

Shifting away from individual differences in the types of representations participants might have used, it is also important to note that there was no clear evidence of different strategies being used across task complexity conditions. This is contrary to the original hypothesis made which suggested that participants would hold more complex task representations – like the Bayesian model – when the task was too complex to verbalise a rule (Livesey & McLaren, 2009; Wills & Mackintosh, 1998). This may be because the manipulation failed to generate the two types of strategy as the complex task was not difficult enough for some participants. It is very possible that even in the complex condition participants managed to verbalise a rule over time, thus clouding the model fitting results (Livesey & McLaren, 2009). Creating a more integrated multi-dimensional stimulus set, or leveraging an artificial category dimension could be one way to approach eliciting these task demand-dependent representations, if they are there to be generated. It may also be true that the experimental design, due to the within-subject design, was actually testing a more specific question about the dynamism versus stability of task representations across difficulty and dimensionality. A between-subjects replication could address this potential confound.

The original aim of this chapter was to explore whether the people's task representations look Bayesian, in terms of using the distributional properties of

categories, in the computation of their confidence reports. The results are not very clear – a full Bayesian model was overall best-fitting at the group level but did not tend to provide the best fit for many individual participants. It is evident that the results from this chapter are ambiguous regarding the type of representations of concepts that people use. However, they do suggest that individuals might be sensitive to task demands and experimental design features and that they could be dynamic in their representations.

It may be possible to extend the approach used here to provide clearer evidence for Bayesian representations. One approach would be to simplify the approach further, to strip back the experimental design to try and better infer the category distribution representations being utilised by participants. This could include removing all experimenter-imposed perceptual noise. For example, in the data explored here, the trial-by-trial noise level varied significantly based on the contrast levels applied to the Gabor patches. This noise could have introduced uncertainty into the confidence reports of participants that obscured the underlying task representations.

However, I opted for a different approach in the subsequent lines of research reported in this thesis: to look for evidence by varying the categorical properties more directly. Creating a new stimulus set, and experimental design, which is designed to highlight the categorical uncertainty – in the absence of significant perceptual uncertainty – could allow for greater clarity about the distribution representations of the participants. In theory, this would allow uncertainty about category membership, as communicated by decision confidence reports,

to map out the shape of the category representations of participants without the noise of varied perceptual uncertainty across the stimulus space from obscuring that mapping. In Chapter 4 this corresponds to manipulating the variability of category distributions and exploring the effects of this manipulation on people's confidence reports.

CHAPTER 4

Category Information & Confidence Sensitivity

Abstract

This chapter marks a methodological shift away from the tasks with high perceptual uncertainty in the previous work. In this chapter, the notion of category distributions, and the underlying models of confidence that map to category judgements, are explored in three related behavioural studies. The experimental designs used in these studies leverage a specific feature of the category distributions – namely their variances – to explore specific predictions about confidence.

In designing experimental categories with unequal variance the predictions made by classic models of confidence – such as distance-to-bound models – are tested and extended. In turn, this allowed for an exploratory modelling exercise to be undertaken whereby the predictions of a Bayesian model (like that of the previous chapter) with access to distribution information, were compared against a simpler comparator model that acted as a proxy for a distance-to-bound variant.

The behavioural and model-based analyses suggest that, while a measure of distance to the category bound is relevant to the computation of confidence, this model is insufficient to fully explain the behaviour of confidence in more complex and variable stimulus spaces. That is to say that metacognitive confidence reports appear to be sensitive to differences in category variance in a way that is not captured in classic distance models of decision confidence.

4.1 | Introduction

Following on from the results of the previous chapter, the experiments in this chapter used manipulations of category variance as a novel way to contrast and explore Bayesian and heuristic confidence computations. This work grew organically from the limitations found and discussed in the dataset of the previous chapter where a combination of perceptual noise and overly influential design elements appeared to obscure the results and create ambiguity in the outcomes. Therefore, in this chapter, the previous methodology was refined in a novel paradigm that reduces the experimentally imposed perceptual noise and explicitly manipulates category distribution variance, with the aim of comparing different models of confidence with greater clarity.

Understanding what information is utilised in the computation of metacognitive decision confidence is important because confidence values track strongly with decision patterns and other behavioural variables. For example, in terms of predicting behaviour, confidence reports have been shown to guide future outcomes such as willingness to gamble on outcomes (Persaud et al., 2007) or the likelihood of information seeking (Desender et al., 2019). Therefore, understanding the origin of self-reported confidence values and their relation to a decision-maker's representation of choice categories is valuable to increasing our understanding of real-world behaviours.

Furthermore, exploring this topic could lead to the creation of targeted interventions to increase metacognitive accuracy in important fields such as health-

care or finance, for example. Indeed, in the field of financial decision-making, confidence in financial knowledge and terminology is known to be often overstated by financial experts (Atir et al., 2015). Similarly, this kind of metacognitive miscalibration can be seen in the overconfident forecasting by chief financial officers in their own firm investment return predictions (Ben-David et al., 2013). These kinds of inaccuracies could have potentially devastating consequences in terms of the provision of overconfident financial advice or execution of financial decisions on behalf of clients without acknowledgement of the uncertainties involved. Beyond the financial sphere, in the field of healthcare, practitioners demonstrate only moderate metacognitive accuracy, with higher rates of miscalibration with greater expertise or experience. For example, the rate of overconfidence has been shown to change with the level of expertise ranging from 25% in students up to 41% in the more experienced residents in one medical decision-making experiment (Friedman et al., 2005). This metacognitive calibration has also been shown to change the likelihood that a medical practitioner will seek further information, with higher calibration of confidence resulting in increased use of secondary diagnostic action when the original diagnosis was incorrect (Clayton et al., 2022). How and whether this advice or new information is taken on board is in itself another potential source for concern, with egocentric advice discounting being a well-founded and robust phenomenon whereby decision-makers rarely follow the advice they are given even to their own detriment (Bonaccio & Dalal, 2006; Yaniv, 2004). Given the importance of high confidence accuracy fidelity in such high-stakes professional fields, as

well as in daily life, exploring how such inaccurate confidence reports may arise and how overconfidence may emerge by modelling the computations themselves becomes paramount.

Categorisation tasks may be uniquely positioned to give insight into confidence reports. One reason for this is that the scope of structures of category spaces is more complex than that of classic perceptual decision tasks, which are often structured with a distinct boundary that divides a one-dimensional uniform stimulus space. In reality, when faced with a categorisation task, the stimulus space and the boundaries that divide it are often more complex. For example, when thinking about category spaces, the possibility of perceptual overlap of conceptually distinct categories is high. That means that there are situations where exemplars of one category could belong to the other with some probability. For example, a cloudy sky might look like it is going to result in rainfall, but in fact does not. Or a milkshake might look like a malted vanilla flavour but in fact, be a salted caramel option. These examples contrast with perceptual decision tasks where a hard boundary often separates uniform distributions unlike these kinds of latent category concepts which hold different kinds of information and uncertainties and which may require different kinds of classification approaches.

Furthermore, the choice distributions of categorisation tasks themselves are rarely as symmetrical as seen in common perceptual tasks. In perceptual choice tasks, there is often a uniform distribution of stimuli with the boundary defined in that space, as distinct from categorisation tasks that establish the underlying

category distributions from which the stimuli are then probabilistically drawn. For example, a naturalistic choice paradigm was illustrated by the example from Chapter 2 in which a dermatologist was navigating medical consultations where benign skin blemishes showed more variability than malignant ones. For a more everyday category judgement example, consider sorting fruit into apples and pears based on colour hue, for example (see Sanborn *et al* (2010) for an example of an interesting multi-dimension fruit sorting task). Apples come in a great variety of colours from green and gold to cherry red. Therefore, the variance would be wide when representing an "apple colour" distribution. By comparison, pears occupy a smaller colour hue band ranging from green to brown with a heavy yellow tone throughout. Thus, the distribution variance of "pear colour" would be relatively low and the resulting distribution relatively narrow. These two distributions would overlap with different means and different variances.

How confidence might operate in these kinds of spaces is unclear. As described in the previous chapter, if categories differ in variance, Bayesian models of confidence – by definition – would utilise the values of the category variances to produce the trial-by-trial confidence reports (Beierholm, 2013). Take, for example, the differences shown in the panels of Fig. 4.1. The top panel of Fig. 4.1 illustrates simulated distributions of two overlapping categories of different means. A probability-based model of confidence, such as a Bayesian model, would use the properties of these distributions to map out confidence based on the probability that a stimulus is drawn from one of these latent concepts. As shown in the second panel of Fig. 4.1, these probabilities (of

P(Category|StimulusValue)) map out the asymmetry of the space well and therefore capture the differences in predicted confidence in a way that encapsulates that asymmetry. In particular, the probability that the underlying category is the one associated with the wide distribution (rain clouds) is higher than the alternative category (non-rain clouds) for stimulus values around the category boundary, reflecting the steeper drop-off in the probability of the narrow distribution around this point relative to the wide distribution probabilities

However, the predictions of a Bayesian model contrast with those of a simple heuristic "distance to bound" model of the kind that have previously been applied to studies of perceptual decision confidence, (Kepecs & Mainen, 2012; Kepecs et al., 2008; Komura et al., 2013). In this model type, as shown in the third and final panel of Fig., 4.1, confidence relates to the distance to the category boundary, or more specifically the distance of the observed evidence from the decision boundary. The boundary can be described as the balance point at which the most likely choice flips from one category to the other, or the midpoint between the two category means, which will be the same point in stimulus space when the categories are symmetrical. Confidence in a given choice is then thought to increase as the stimulus value moves away from that boundary toward the category mean itself and beyond. Therefore, the slope of this confidence trajectory encompasses information about the stimulus, the boundary, and, perhaps, also the category means.

These distance-to-bound models can be likened to the rule-based models of the previous chapter. The rule-based models also established the idea of a

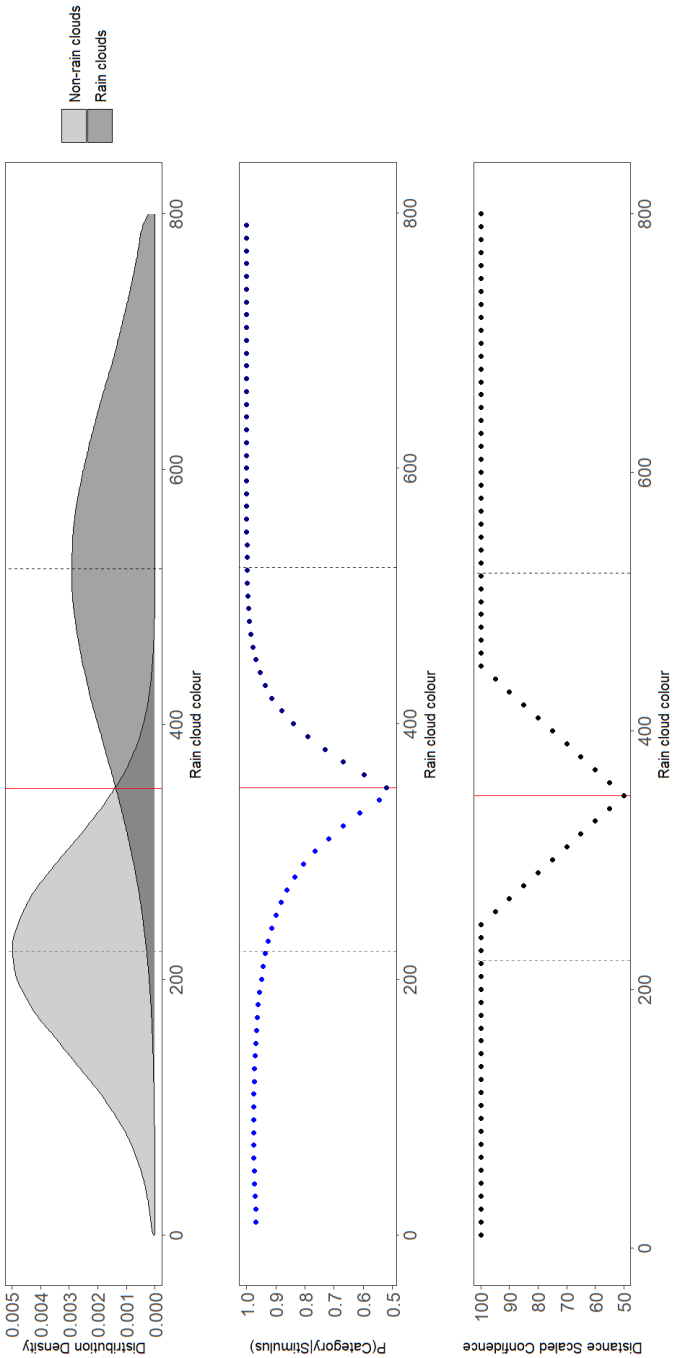


Figure 4.1: The x-axis remains constant across all figure panels and represents the stimulus dimension of cloud colour which differentiates rain-producing and non-rain-producing clouds. Similarly, on each panel the vertical dashed lines map the category means while the red vertical line indicates the point of equi-probability between the means, the balance point at which the response pattern should shift from one category to the other. Top panel: two simulated category distributions of different means. Middle panel: the simulated probability of a stimulus belonging to a specific category (P(Category|Stimulus). Bottom panel: a simplified distance-to-bound model in which confidence increases in a linear manner from the equi-probability point between the means to the means of the categories.

clear decision boundary being constructed. However, in contrast to the rule-based model of the previous chapter, which established the decision rule in an all-or-nothing manner and then scaled confidence based on the trial-by-trial perceptual uncertainty which lent itself to differences in certainty about stimulus identity, distance-to-bound models of confidence posit that confidence will increase moving away from that bound in a monotonic manner (here depicted as linear) (Kepecs et al., 2008; Komura et al., 2013).

Theoretically speaking, this model, therefore, allows for the incorporation of certain information about the category distributions into metacognitive confidence reports, such as the boundary between the two distributions. However, despite its theoretical and methodological utility (Kepecs et al., 2008; Komura et al., 2013; Vickers & Packer, 1982), the extent to which this model describes the underlying computations of confidence, and not just a simple pattern in the data remains unclear. Indeed, much experimental evidence for this model of confidence has relied on empirical work that draws on stimuli generated from a fixed or uniform distribution of stimuli with the boundary defined in a discrete manner in that space. Confidence is a simple "read off" of the distance of the stimulus measurement from the decision criterion. In these instances, the slope of confidence with distance from the boundary is mirrored and symmetric in both directions. This evidence base makes it unclear if or whether confidence reports are sensitive to the variance of the category distribution or simply map a monotonic function from the decision bound to the category means in stimulus space. Furthermore, this symmetry also means that the decision boundary is

simultaneously set on both the choice balance point and the midpoint between the two category means as they are themselves equitable. In this way, it is also unclear if the distance-to-bound is being measured relative to the mid-point between the means or the balance point of the category likelihoods.

This chapter, therefore, focuses on unpicking the sensitivity of metacognitive confidence to category distribution variance. In order to interrogate these questions, a series of experimental studies and complementary computational modelling exercises were conducted and are reported in detail. By creating two overlapping categories with different variances and means along a single stimulus dimension with low perceptual uncertainty, the sensitivity of metacognitive confidence to these distribution properties is tested. If confidence is sensitive to the difference in variance it would be apparent in a difference in the slope of confidence on either side of the decision boundary.

4.2 | Study 2

The first study in this chapter initiated a line of experiments testing the contrasting predictions of Bayesian and distance-to-bound models that emerge from an asymmetry of variance in a decision-choice task. In this study, participants were asked to decide whether to "retrieve" or "zap" packages dropped to earth by two spaceships, and, on each trial, provided a graded confidence judgement in their decision. The stimuli for this study, therefore, took on the form of perceptually well-controlled categories, defined as locations along the horizontal axis. Importantly, the stimulus category distributions were designed to have different means and different variances – one spaceship dropped packages over a wider area than the other. This design choice creates overlapping categories of different spreads. In turn, this allows for a stronger test of the "distance to the bound" model of confidence than previous work using nested categories (Adler & Ma, 2018), where the distance from the boundary predictions and the fit of more involved models of confidence are confounded. The key prediction in this study was that the effect of distance to the category boundary, as seen in the gradient of the confidence reports, would differentiate distributions of different spread, with a steeper gradient and numerically lower confidence for narrow as compared with wide category distribution trials.

4.2.1 | Experimental Questions

1. Do participants adjust confidence according to the distance to bound, with confidence increasing with distance from the category boundary?

2. Do participants show a sensitivity to the variance of the category distribution, with higher confidence in trials drawn from wider distributions?

4.2.2 | Methods

The data set, data cleaning script, confidence simulations, and analysis script for this study are housed on the (OSF). The experimental scripts and relevant materials for running the data collection can be found on (GitHub).

Participant Sample

A total of 60 participants took part in this study. Participants were compensated at a rate of £10.00 per hour for their time. After exclusions (see below), the N=52 participants included in the final sample had an average age of 33 years (range 18-64, standard deviation = 12.5). Self-identified gender splits were female (N=31), male (N=20) and non-binary (N=1). Handedness was reported as right (N=46), left (N=5), and ambidextrous (N=1). The average time taken by participants to complete the experiment was approximately 36 minutes (minimum 24 minutes, maximum 65 minutes, standard deviation 8.7 minutes).

Exclusion Criteria

Participants who on average fell below 2 standard deviations of the total sample mean accuracy in either the associative learning and/or the active test trials were excluded from the analysis. This approximated a requirement to achieve over 80% accuracy in the phase 0 associative learning trials – a reasonable requirement given the general performance levels of participants and the absolute nature of the relationship being learned. 6 participants (10%)) were identified as being below this cut-off point in the associative learning phase. A further 5 participants fell below 2 standard deviations of the sample average accuracy in phase 2 active test trials. The lower bound cut-off in this phase approximated a requirement to be above chance level with a cutoff at 55%, 3 of whom were also below the cutoff in the phase 0 trials. This amounted to 8 participants in total eligible for exclusion based on these criteria. When applied the remaining sample size was N=52. Therefore, all results below are reported with these exclusion criteria applied with N=52.

The local ethics committee approved the study protocol (CUREC: R68879/RE001). Participants were provided with an information sheet and gave consent to participate before beginning the study. Participants were informed that they could withdraw at any point by closing their browser window. Incomplete responses were not analysed. The study used the Prolific platform to recruit participants and participants were compensated for their time in accordance with the platform guidelines. Data collection was performed using a custom

javascript experiment script hosted locally on the lab server. Integration testing using Cypress was conducted to ensure reliability across browsers.

4.2.3 | Experimental Design

Participants' main task was to learn the relative drop locations of two spaceships, a "good" spaceship that dropped coins and a "bad" one that dropped bombs. They had to choose whether to retrieve or zap packages accordingly. Prior to the main experiment, participants went through a short specified training to memorise the identity (colour-based) of the good and bad spaceships. The experimental blocks were then divided into learning and testing phases. During learning, they saw spaceships (that were identifiable by their colour) drop packages across different locations; during testing, they saw a package dropped, but the identity of the spaceship was hidden (by cloud cover). Therefore, they had to infer whether to retrieve or zap the package according to the location of the dropped package. Participants received no feedback on their specific responses during the test phases. This design feature means that their confidence should reflect their representation of categories, rather than any reinforcement history associated with responses to stimuli in particular locations.

This experiment was therefore designed to explicitly test the learning of categories (here defined as the different spaceships) when the observable properties of those categories (the drop locations of the packages) are drawn from overlapping distributions with different variances. More details of each phase of the experiment will follow in the subsequent sections of this chapter.

Phase 0: Associative Learning

Two spaceship stimuli were used throughout the study – a "good" and a "bad" spaceship. Each spaceship was identifiable by its colour (blue or orange, with the assignment to good/bad randomised across participants). At the beginning of the study, participants were introduced to these spaceships and learned about each of them. The "good" spaceship dropped coins that the participants were told they were trying to collect. The "bad" spaceships dropped bombs which would cause them to lose coins from their coin purse if they collected them. Thus, the participant aimed to optimise their gains by collecting only those packages that were dropped by the "good" spaceship. To facilitate this, this phase of the game was constructed to ensure the participants had memorised the assignment of colours for the "good" and "bad" spaceships.

To enable this learning, the first stage of the experimental protocol was a 20-trial associative learning phase in which the image of one of the spaceships appeared on the screen followed quickly by the "package" it dropped. The package was an image of a cardboard box with unknown content. The package appeared alone and in the centre of the screen – learning of the category information (in the form of the locations of the package) came later. Participants were asked to decide to either "zap" or "retrieve" the package using the response buttons and visual feedback was provided for each trial, see Fig.4.2. Feedback was given on every trial in the form of an image of the content of the package that they had just responded to. This feedback was again presented centrally

onscreen. Note that if the "zap" response button was used the feedback image was overlaid with the word "zap" to indicate that it had not been collected, in the case of the coins. Similarly, in the case of bombs, the "zap" response was shown to neutralise the negative effect of the bomb on the outcome. The ideal response pattern was, therefore, to only retrieve the coins and to only zap the bombs.

In this way, through the feedback, the participants should memorise which spaceship was good and which was bad. The relationship between the spaceship and the package contents was deterministic. The stimulus was on screen for 1.5 s and the subsequent visual feedback lasted 2 s. There was a time-out function of 10 s for a response on each trial. The associations learned in this phase were held across the entire study for each participant. Therefore, this associative learning phase was only experienced once by the participants, at the start of the study.

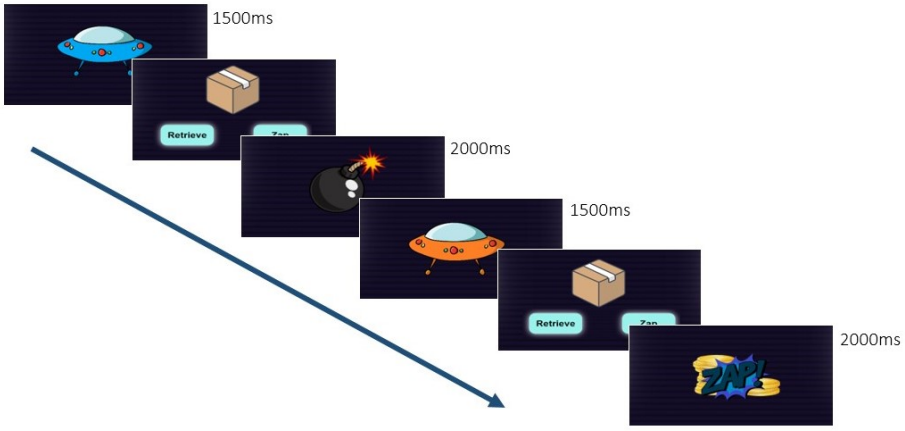


Figure 4.2: The timeline of the associative learning phase. The images and response buttons are depicted alongside the visual feedback provided for the different response combinations.

Phase 1: Passive Learning

Each of the four experimental blocks began with a 60-trial passive learning phase, that gave participants information about the relative drop locations of the two spaceships that would hold for the duration of that block. On each trial, one of the two spaceships was shown to move onto the screen and drop a package at a particular location before flying away, Fig. 4.4. Across trials, participants could therefore learn the spatial distribution of drop locations associated with each spaceship. The contents of the packages were not revealed, but the associated package contents for each spaceship had been previously established in the associative learning phase, Phase 0, of the study. The animation was continuous with no interaction required by the participant. The order in which the two spaceships were presented was interleaved randomly. Overall, the trials

were split evenly between each spaceship (30 trials per spaceship). This phase lasted approximately 2 s per trial, with each trial representing a single package drop animation cycle.

The drop locations for these trials were established using predefined Gaussian distributions with set means and variances. The values for these distribution properties were chosen based on simulations constrained by the 0-800 pixel space (see Appendix for full details). Based on these simulations, and centring the boundary of the two distributions on 350 pixels, the mean for the narrow distribution was set to 222 and the wide to 522, for blocks in which the wide distribution fell to the right of the screen (to centre the boundary on the balance point P(A)=P(B)=0.5 point). When the wide distribution was flipped to sit on the left of the screen, the boundary was re-centred on 450 pixels and the means were set to 278 for the wide and 578 for the narrow distributions respectively, (again to centre the boundary on the balance point P(A)=P(B)=0.5 point). Here the word category boundary is used to describe the point at which P(A|x)=P(B|x)=0.5. There is no fixed boundary as in the case of perceptual decision-making tasks where a hard boundary is inserted into the stimulus space.

These values were shown to produce measurable differences in "confidence gradients" on either side of the category boundary. See Fig. 4.3 for simulated confidence values in this space. Confidence here is approximated as the likelihood of a given distribution trial at each location based on drawing 10000 samples from those distributions. The simulated confidence in Fig. 4.3, combined with a study of Fig. 4.1, highlights a key prediction of higher confidence in

trials drawn from the wide distribution, on average. This derives from the fact that, simultaneously, the likelihood of a given location value for the *less likely* category is generally much higher for the wide distribution than the narrow. Thus, the key prediction is that if participants are sensitive to the differing variance of the two distributions, they should be more confident when classifying stimuli as being from the category with the wider (versus narrower) distribution, particularly in the middle ranges of values between the boundary and the category means.

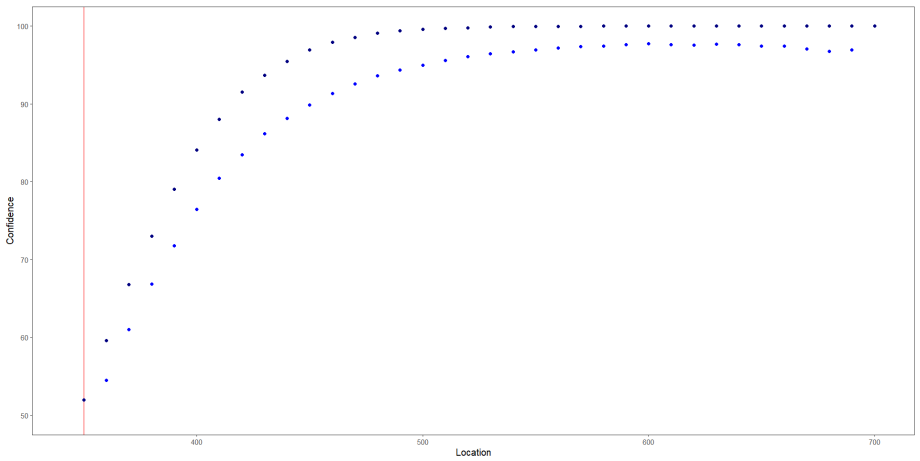


Figure 4.3: Simulated expected confidence values for the narrow (blue) and wide (navy) distributions moving away from the category boundary (vertical red line at 350) based on the chosen values for the distribution means and variances. For ease of comparison, the values for the narrow distribution have been flipped across the boundary and plotted alongside the wide distribution trials moving away from that boundary. The small jitter of confidence off 50 at the boundary relates to the sampling and averaging method used to simulate this figure.

Within the Phase 1: Passive Learning blocks, the specific locations seen were pseudo-randomly generated from those Gaussian distributions anew for each block, with the constraint that the spread of values across the 30 trials

per spaceship approximated both the mean [within a max of 5.5 pixels within participants] and the variance [within a max of 4 pixels within participants] of the distribution assigned to that spaceship. This forcing function ensured that the participants were exposed to an array of trials with a "wide" spread in one instance, and a "narrow" spread for the other.



Figure 4.4: Screenshot of the animated passive learning phase whereby the spaceships flew on screen, dropped a brown paper package down to the earth and flew away.

Phase 2: Active Testing

Following the blockwise training, each testing block consisted of 60 trials. In test trials, the package was seen to drop to Earth without the identity of the spaceship being visible. A cloud bank was seen on the screen where previously the spaceships had been visible. Thus, the content of the packages could be inferred using only the drop location. Based on this, the participants then

decided whether to retrieve or zap the package. After each retrieve/zap decision, participants rated their confidence in that decision by moving a slider on a scale that ranged from "totally UNSURE" to "sure CORRECT". The slider thumb started at a random position on every trial. The slider required interaction to prevent the passive use of the slider's start position. The confidence report was submitted by pressing the confirm button. For further details of the participant experience in the Phase 2 active testing trials see Fig. 4.6.

Drop locations for the test trials were generated in a different manner from the passive learning trials. Instead of sampling from the underlying distributions randomly, the on-screen locations were divided into preset location bins to facilitate analysis of participants' decisions and confidence ratings as a function of stimulus location relative to the category boundary, see Fig. 4.5. The objective category boundary of the distributions was taken to be the centre of the middle bin with the boundaries of the location bin specified as being 20 pixels on either side of that bound. On either side of the middle bin, four bins of 40 pixels width were then established (i.e., binning down the space from 170-530 pixels in the 350 pixel category boundary case). A fifth "outlier" location bin then spanned the rest of the location space up to 10 pixels from the screen boundary (e.g., 10-170 pixels and 530-790 pixels in the 350 pixel category boundary case). In each bin, a certain percentage of trials were generated. In the middle bin, and four bins to each side 10% of trials were generated (translating to N=6 trials per bin, per block) from a uniform distribution across the possible location values of that location bin. In the "outlier" bins, at the tails of the distributions, 5%

of trials (N=3 trials per bin, per block) were generated from a uniform distribution across the possible location values of that location bin. Therefore, every bin represents 10% of trials and together the "outlier" bins also represented 10% of trials, see Fig. 4.5.

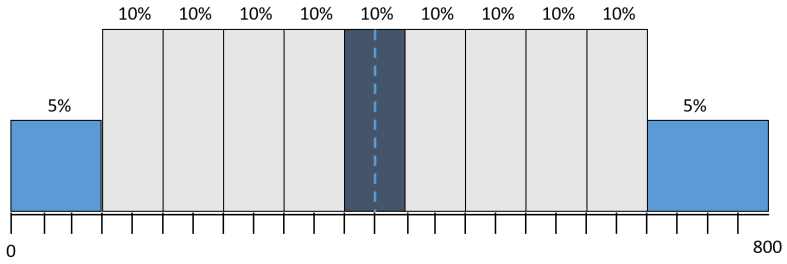


Figure 4.5: Chart demonstrating the preset location bins within which trial drop locations are generated. Note the smaller percentage of trials in the outlier bins at the extremes of the stimulus space. Also, note the difference in width of the outlier bin on the right compared with the left. This was reversed for blocks in which the location of the wide and narrow distributions was reversed.

The accuracy of responses was defined based on the most likely distribution per location bin. The middle bin was therefore dummy coded with half of the trials being ascribed as "bombs" and the other half being designated as "coins" as the likelihood of each distribution was almost equal across the bin space. All other bins were labelled as narrow or wide, according to the side of the screen on which the wide or narrow distribution was situated in a given block. A trial was coded as being correct if the participants chose to retrieve a package that should have contained coins or to zap a package that should have contained a bomb, according to these likelihoods. Participants received 3 coins for retrieving coin packages but lost 3 coins for retrieving a bomb. Zapping packages resulted in no score change. This stimulus-response contingency created a balanced reward

matrix. Importantly, there was no trial-by-trial feedback given in this phase of the experiment, participants were told their cumulative score at the end of each block and encouraged to try to improve their score in the next section. There was a time-out of 10 s on the initial choice. Once a choice was submitted there was no time-out on the confidence slider screen.

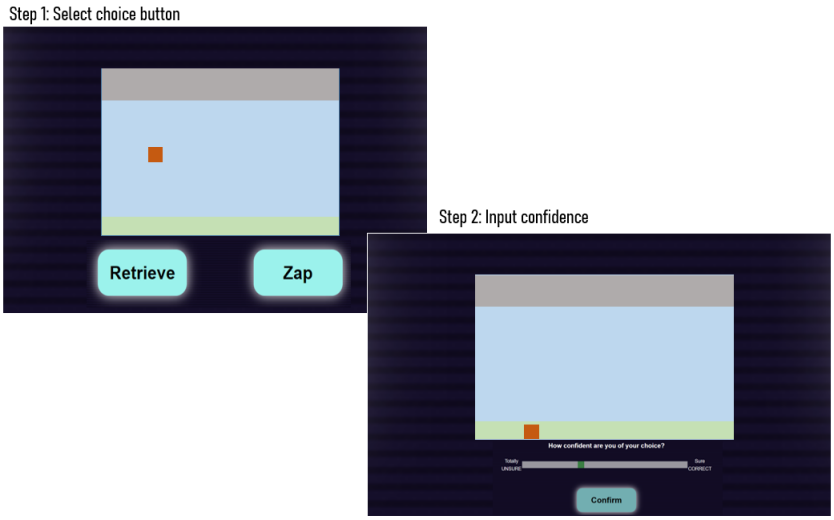


Figure 4.6: Screen shots of the response and confidence input screens in the active testing phase of the study. Note that the spaceships are not visible to the participant. The response and confidence must be selected based on just the drop zone of the unidentified package.

Distributions & Randomisation

The locations of the package drops were drawn from predefined normal distributions differing in means and variances. As the experimental screen was 800 pixels wide and based on the simulation results, the means were set at either 222 pixels (narrow) or 522 pixels (wide) when the wide distribution was on the right-hand side of the screen. This offset the objective category boundary

between the categories away from the natural 400 px mid-point to 350 px to prevent a simplification of the task via the use of the screen mid-line. When the wide distribution was on the left-hand side of the screen the distribution means were shifted 278 pixels (wide) and 578 pixels (narrow), shifting the category boundary to 450 px giving more space for the wide distribution samples to prevent a high proportion of trials from falling out of range.

In accordance with the simulations, the variances were set to either 6400 (narrow) or 19600 (wide) (standard deviations of approx. 80-pixel units and 140-pixel units respectively). These category distribution features were balanced across blocks and spaceships such that the wide spaceship and narrow spaceship were seen to occupy different sides of the screen in different blocks. The order in which the distributions were allocated to the sides of the screen and the spaceships was semi-randomised across blocks and participants. The ordering was constrained such that each participant saw all four possible combinations of spaceship colour and distribution width assigned to each location.

4.2.4 | Results

Phase 0: Associative Learning Results

After exclusions, the average accuracy in the associative learning trials was 93.5%. Most participants (49 out of 52) achieved perfect performance in the second half of the learning block, as expected given the simplicity of the task. This suggests that participants understood the link between the spaceships and the packages required to succeed in the task.

Phase 2: Active Testing Results

The average accuracy on test trials was 84%. Across all four blocks, most participants tended to achieve above 75% accuracy. As there was a new location mapping to learn in each block, this indicates that participants were attentive and performed well across the experiment as a whole. Analysis also indicated that within participants on average accuracy increased with binned distance from the category boundary, $F(2.68, 136.51) = 412.44$, $p < 0.001$, see Fig. 4.7. These data indicate that the participants were sensitive to the category boundary and to distance relative to that bound.

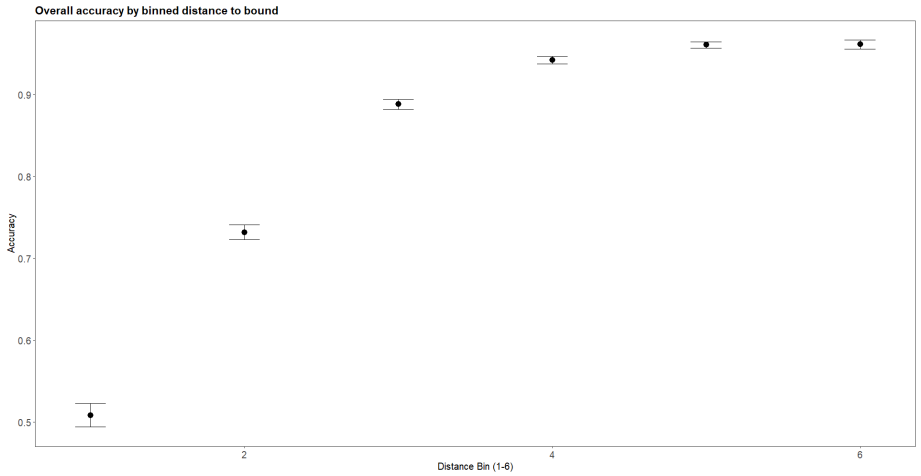


Figure 4.7: Mean accuracy per distance bin from 1 to 6 - moving away from the category bound, across all participant data (with error bars corresponding to SEM). Note that the correct category identity is dummy-coded for Bin 1, so that accuracy will necessarily approach 0.5 with a correspondingly wide error bar.

Phase 2: Confidence Analysis

Looking at confidence reports, participants appeared to use the range of values available on the scale well and differentiated qualitatively between narrow and wide distribution trials in their confidence reports, particularly at high confidence levels, see Fig. 4.8.

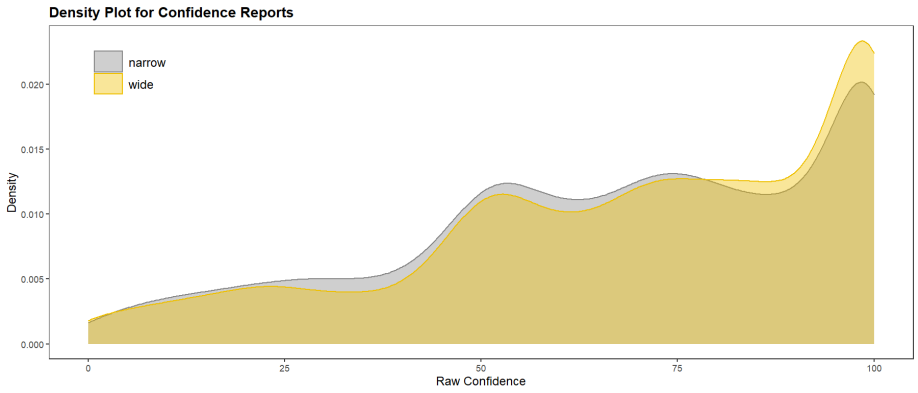


Figure 4.8: Density plot capturing the confidence scale usage by all participants, split by trial distribution (wide versus narrow).

Confidence reports also varied significantly with the accuracy of trials with mean confidence for correct trials versus mean confidence for incorrect trials within participants being significantly different (M = 71 for correct trials, M = 50 for error trials; $t(51) = 15.81$, $p < 0.001$). Mean reaction times for making confidence reports in correct and incorrect trials did not differ significantly ($t(51) = -1.65$, $p = 0.1$). Focusing on trials where the participants chose in line with the most likely distribution (i.e., correct trials in bins 2-5) the grand confidence means in each bin split by distribution can be seen below in Fig. 4.9.

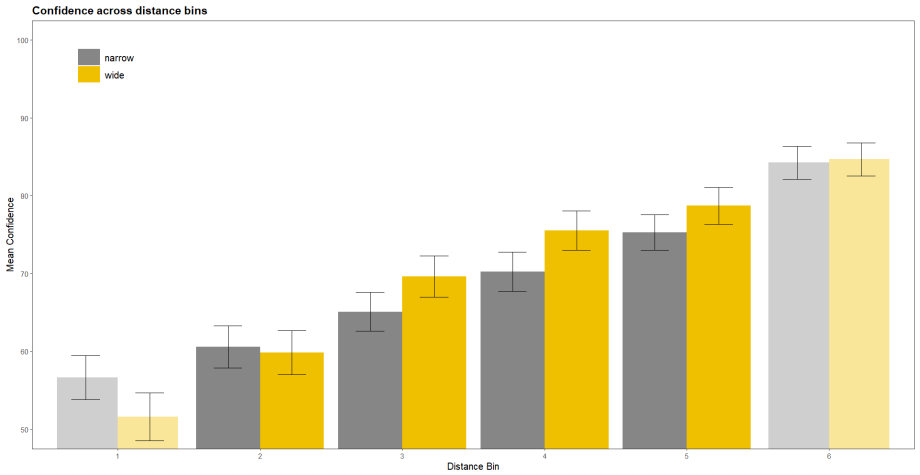


Figure 4.9: This bar plot visualises data from Study 2 relative to the objective boundary. It indicates the mean confidence on the y-axis (with error bars corresponding to SEM) and distance bin, here running increasing from left to right from 1-6 away from the objective category boundary at 1, on the x-axis relative to the objective category boundary. The confidence reports are split based on distribution, whether trials are drawn from a wide or a narrow distribution. The stimulus space was folded across the boundary to collapse from 11 location bins down to 6. Bins 2-5 were included in an ANOVA (see below). Bins 1 and 6 were not included in the ANOVA as they contained fewer trials and Bin 1, the middle bin which contained the objective boundary, was dummy-coded for accuracy so confidence could not be reliably decoded. Mean confidence in Bins 1 and 6, in each distribution, are plotted with a lighter shading to indicate that they are not included in the ANOVA.

The critical prediction tested in this study is that confidence would vary differently as a function of distance-to-bound for the narrow vs. wide category distributions. To test this prediction, a two-way ANOVA was carried out on mean confidence ratings as a function of location bin (Bins 2-5 only) and distribution (narrow vs. wide), with Greenhouse Geisser correction applied to correct for violations of sphericity. Mean confidence was numerically higher for categorisations of the wide distribution, but the main effect of distribution was not significant [$F(1, 51) = 3.33$, $p = 0.074$]. However, the main effect of distance

(here represented by location bin) [$F(1.4, 71.43) = 102.46$, $p < 0.001$] and the interaction between distance and distribution [$F(2.29, 116.9) = 7.47$, $p < 0.001$] were both statistically significant. Participants were more confident when deciding about locations further from (vs. closer to) the category boundary, as would be expected, but critically the relationship between confidence and distribution varied consistently as a function of the binned distance to the bound: this can be attributed to the higher confidence that emerges in the wide distribution in distance Bins 3-5.

Follow-up t-tests indicated that Bins 3-5 all had significantly different confidence reports for wide versus narrow distribution trials ($p < 0.05$ in all cases). The confidence values in Bin 2 were not significantly different based on distribution $t(51) = 0.382$, $p = 0.7$). Importantly, the significant interaction effect suggests that participants were not just generally more confident when making decisions for the "wide" category. Indeed, that is not the case in Bin 2, where they were numerically more confident for "narrow" decisions. This trend of lower confidence in "wide" decisions is also seen in the Bin 1 reports and then participants were highly confident for both decisions in Bin 6. The results in Bin 6 are arguably particularly interesting. The confidence reports are numerically very similar for both distributions, even though objectively the points sampled in this space would be on average closer to the category boundary for the "narrow" trials than the "wide". Given that the confidence values have not reached the ceiling (averaging around 85) the reason for this is not clear. It could perhaps be due to the lower number of trials in these bins leading to erro-

neous averages or, perhaps, it is indicative of a difference in the representation of the two distributions relative to their objective reality by the participants. Taken together, this analysis of confidence suggests that participants are not only sensitive to the metric of distance-from-bound but instead, as predicted by Bayesian accounts, seem to be sensitive to the distribution variance.

Phase 2: Subjective Response Boundaries

Given the sensitivity of response patterns to the reward contingencies (see Appendix for details) and the potential for individual differences in learning or modelling of the stimulus space, the data were reanalysed by looking at the inter-individual differences in response curves. This was necessary as the main analysis was performed relative to the objective category boundary as implemented by the experimental design, but it is possible that the confidence differences just reflected participants consistently establishing the boundary in a different location. For example, participants may have inserted the boundary closer to the narrow distribution mean than expected, such that each "wide" bin lay further from the category boundary in this subjective space than the corresponding "narrow" bin. This would lend itself to the calculation of an inflated confidence difference between the two distributions when analysed in the objective boundary space.

Within each participant, looking at the response patterns as a percentage per binned location, it is clear that participants showed variability in the mapping of their choices onto the stimulus space. See Fig. 4.10 for a detailed illustration

of these differences. Note for example that some participants have a steeper response curve than others. Furthermore, note that some participants reached an absolute response rate in the extreme bins (i.e., bins 1 or 11 showing a response proportion of 0 and 1 respectively), while others did not. For some participants, the objective boundary did not seem to capture the balance point of their response curves, as shown by the deviation of their response curves from the vertical red dashed line when y = 0.5.

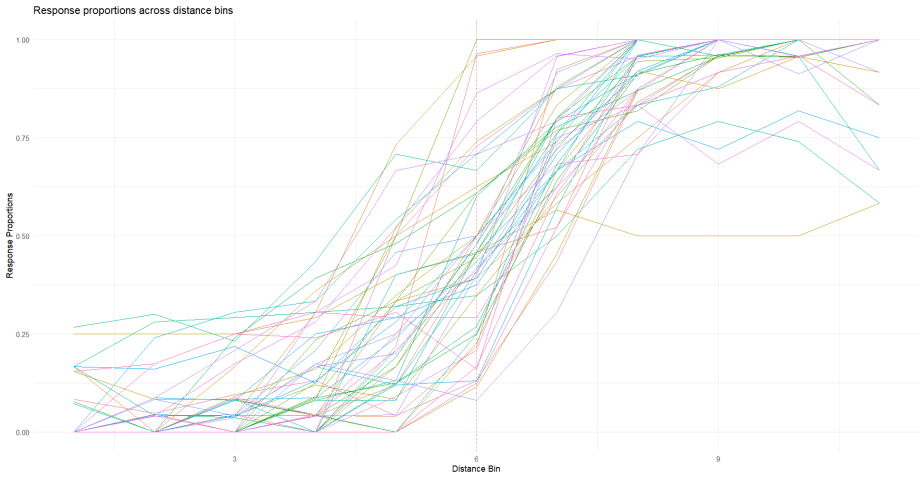


Figure 4.10: Individual participant-based response curves indicating where each participant was responding as though trials were drawn from the wide distribution. The x-axis indicates the 11 location bins, with the objective category boundary falling at bin 6 of 11 - demarcated with the vertical dashed red line. The y-axis indicates the proportion of responses that were for the "wide" distribution in that bin. Note the variability displayed.

Based on this variability, and on the understanding that people's confidence is strongly related to their response patterns, data were analysed at the participant level to find their subjective cross-over point. This is the point at which participants flipped from responding as though trials had been drawn from one

distribution to responding as though trials had emerged from the other. Looking at Fig. 4.10 it is clear that this point varies across location bins for different participants. Within each participant, and within each of these 11 location bins, the average location of the bin (as experienced by that participant) was calculated. When plotted against either wide or narrow response proportions to an individual s-shaped curve of response proportions across the stimulus space. To these points, within each participant, a sigmoidal curve was then fit using a non-linear least-squares (NLS) approach. This fitting took as parameters a mid-point on the x-axis, an upper and lower asymptote, and a slope or growth rate of the curve. The start points were set equitably across all participants as follows: lower asymptote 0.1, upper asymptote 0.9, slope β (utilised in the form -1/β) of 30, and midpoint on the x-axis of 405. The data being fit were transformed such that all values relating to the wide bin fell on the right-hand side of the location space and the narrow distribution values were on the left-hand side. Therefore, a 405 midpoint value is a slightly conservative starting point for that parameter which suggests that participants may hold a representation of the distributions that is more symmetrical than the objective reality with a wider narrow distribution and a narrower wide distribution than actually exists. The purpose of this fitting was to find the best estimate of the mid-point on the x-axis parameter which would be used as a proxy for the subjective decision boundary within each participant and therefore allow for a reassessment of the distance-to-bound model accounting for this inter-individual variability.

This fitting procedure and these start points enabled the data of all but one

participant to be successfully fit using the NLS functions. One participant, Participant 60 could not converge, most likely due to an observed dip and rebound in response proportions around Bin 6. Participant 60 was therefore fit manually by interpolating a line between bins 5 and 7 choosing the mid-point of that line as the subjective category boundary. Within each participant, the data were then rebinned around that category boundary such that the binning mirrored the objectively defined bin size and layout, but centred on the subjective boundary point. This allowed the ANOVA to be re-run using the subjective category boundary as the distance marker to see if this increased the signal-to-noise ratio. The aggregated results of this rebinning are displayed in Fig. 4.11.

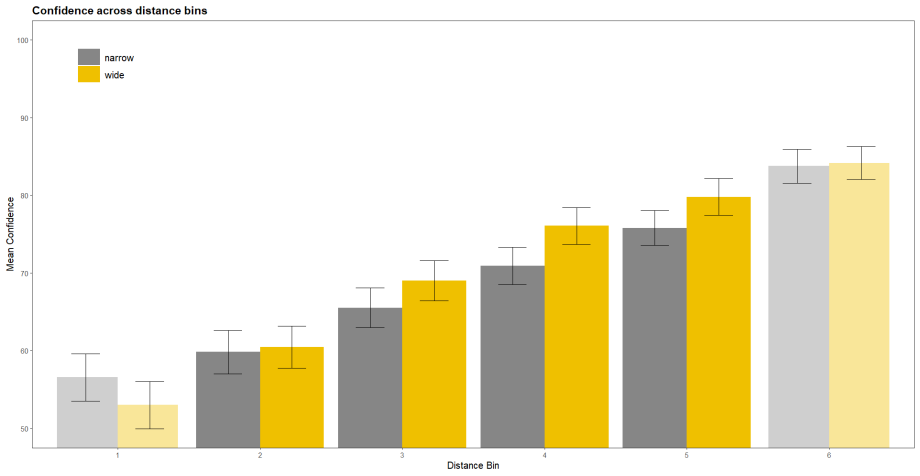


Figure 4.11: This bar plot visualises data from Study 2 relative to the within-participant subjective boundary. It indicates the aggregated mean confidence per distance bin as binned within individuals relative to the subjective category boundary of their responses (error bars indicate SEM). Bars 1 and 6 are shaded out to indicate that they were not included in the ANOVA.

The main experimental prediction in this study, that confidence would vary differently as a function of distance-to-bound for the narrow vs. wide category

distributions, was re-tested using a two-way ANOVA carried out on mean confidence ratings as a function of subjective location bin (bins 2-5 only) and distribution (narrow vs. wide). Once again, confidence was averaged per participant within each subjective distance bin for each distribution trial type. Greenhouse Geisser corrections were applied to correct for violations of sphericity. The main effect of distribution [$F(1, 51) = 8.28$, $p = 0.006$], the main effect of distance [$F(1.62, 82.77) = 86.96$, $p < 0.001$], and the interaction between distance and distribution [$F(2.69, 136.96) = 4.96$, $p = 0.004$] were all statistically significant. Participants were more confident when deciding about locations further from (vs. closer to) the category boundary, and in wide versus narrow trials. The relationship between confidence and distribution varied consistently as a function of the binned distance to the bound, being smaller in Bin 2 than in the location bins further from the subjective category boundary. Follow-up t-tests indicated that there was no significant difference between confidence reports based on the generating distribution in Bin 2 ($t(51) = -0.52$, $p = 0.59$) but the difference in confidence reports by distribution was significant for Bins 3-5 (all $t(51) < -2.8$, all $p < 0.05$). These results once again are consistent with the initial experimental predictions that confidence would be affected by the category distribution variance such that confidence would be higher for trials where variance was higher, i.e., in the wide distribution trials.

4.2.5 | Discussion

This study aimed to investigate the sensitivity of confidence to category distribution variance by exploring the behaviour of confidence relative to a category boundary when the variances of the two experimentally imposed category distributions were different. The main analyses of confidence indicated that confidence increased with distance to the bound, however, there was a difference in the gradient of confidence values for the "wide" versus the "narrow" distribution trials. Importantly, the significant interaction that emerged in both the objective boundary and subjective boundary analyses effects suggests that participants were not just generally more confident when making decisions for the "wide" category.

Collectively, these analyses suggest that participants are not only sensitive to the metric of distance-from-bound but instead, as predicted by Bayesian accounts, seem to be sensitive to the distribution variance as well. To ensure the robustness of these findings a direct preregistered replication was conducted and is detailed in the following section.

4.3 | Study 2b: Direct Replication

To assess the reliability of the above-described results, I conducted a direct preregistered replication of Study 2. The preregistered research hypotheses for Study 2b were as follows: Participants will report higher confidence on trials drawn from a wider, as compared with a narrower, category distribution and that sensitivity to distribution variance will be modulated by distance to the category boundary. Then, based on the results of Study 2, the following specific predictions were made for the preregistered ANOVA. It was expected that the main effect of distance to bound would be significant with confidence increasing with distance to the category bound. Another key prediction was that the interaction effect between distance and distribution would be significant with the effect being driven by differences in confidence reports in distance bins 3 and 4 (middle distance bins), with greater confidence for trials on the side of the wide distribution. No specific predictions were made about the main effect of distribution which was not significant in Study 2.

4.3.1 | Methods

The data set, data cleaning script, confidence simulations, pre-registration, and analysis script for this study are housed on the (OSF). The experimental scripts and relevant materials for running the data collection can be found on (GitHub).

Participant Sample

Participants were sampled in batches of 10 until at least 55 participants passed the exclusion criteria, up to a maximum of 10 batches. Participants were compensated at a rate of £10 per hour for their time. Data were checked against the exclusion criteria from batch 6 onward ($N = 60$) to prevent oversampling. Here, a total of $N = 70$ participants were collected, leaving $N = 58$, post exclusions.

After exclusions (see below), the $N = 58$ participants had an average age of 25 years (range $18 - 54$, standard deviation = 7.6). Self-identified gender splits were female ($N = 32$), male ($N = 26$) and non-binary ($N = 0$). Handedness was reported as right ($N = 51$), left ($N = 7$), and ambidextrous ($N = 0$). The average time taken by participants to complete the experiment was approximately 35 minutes (minimum 21 minutes, maximum 71 minutes, standard deviation 8.3 minutes).

Exclusion Criteria

Exclusion criteria were defined based on the values established in Study 2 as requiring accuracy above 75% in the phase 0 associative learning trials and above 55% in the phase 2 test trials. Furthermore, participants were excluded if they used confidence reports in a way that signalled non-compliance, defined as using fewer than 3 unique values across all confidence trials. Of the sample of $N = 70$, 9 participants were identified as having a score in phase 0 that was below the preregistered exclusion cutoff of 75%. 8 participants were identified

as having accuracy below the phase 3 cut-off of 55%, 5 of whom overlapped with the previous 9, leaving 12 candidates for exclusion. No participants used confidence in a way that signalled non-compliance. Thus, the final exclusion list was $N = 12$, leaving a final sample size of $N = 58$, post exclusions.

Experimental Design

The same study design, trial numbers, and stimuli were utilised in this replication. Therefore, the experiment had a within-subject design with conditions counterbalanced and randomly assigned across blocks. The stimuli were location-based along the horizontal axis and were generated by two distributions with unequal means (wide or narrow). The values of the means and variances of these distributions matched those of Study 2. Once again, the conditions were based on the position of the category distribution means on-screen (left or right) and the stimulus type (good or bad) associated with the distributions. The sampling and exclusion criteria were preregistered on the basis of the values observed in Study 2.

4.3.2 | Results

The data processing pipeline from Study 2 was reproduced for Study 2b. In terms of accuracy and task performance, after exclusions, the average accuracy in the associative learning trials was 92.6%. The average accuracy on test trials was 84.5%. The analysis also indicated that on average accuracy increased with binned distance from the category boundary, $F(2.86,163.27) = 417.99$, $p <$

0.001. Looking at confidence reports, confidence was higher ($t(57) = 12.75$, $p <$ 0.001 – mean confidence for correct trials = 75.1, mean confidence for incorrect trials = 50.7) and faster ($t(57)$ = -5.47, $p < 0.001$ – mean confidence reaction time for correct trials = 1835.7ms, mean confidence reaction time for incorrect trials = 2035ms) for correct versus incorrect trials.

As before, to test the key predictions of the study, a two-way ANOVA was carried out on mean confidence ratings as a function of distance bin (bins 2-5 only) and distribution (narrow vs. wide). The direction and strength of the effects were similar to those of Study 2. Greenhouse Geisser corrections were applied to correct for violations of sphericity. Unlike Study 2, the main effect of distribution [$F(1, 57) = 24.14$, $p < 0.001$] was significant, with higher confidence in trials drawn from the "wide" as compared with the "narrow" distribution. As before, the main effect of distance [$F(1.43, 81.38) = 107.63$, $p <0.001$] and the interaction between distance and distribution [$F(1.95, 111.37) = 3.1$, $p = 0.050$] were both statistically significant. Pairwise t-tests indicated that all Bins 2-5 showed significant differences in confidence based on distribution, with higher confidence in the "wide" than the "narrow" trails (all $p < 0.05$).

These results are in line with the preregistered predictions and replicate the main findings of Study 2, see Fig. 4.12 for details. Some key differences to note in these data, relative to the analogous results in Study 2 include the fact that the main effect of distribution was significant here – possibly driven by the higher confidence reports in Bin 2 as compared with Study 2. However, as in Study 1, confidence was not simply higher for all "wide" responses – as shown by

Bin 1 and Bin 6 patterns where confidence was numerically higher for "narrow" trials. This evidence provides further support for the idea that confidence is sensitive to distribution properties like variance, beyond the distance-to-bound.

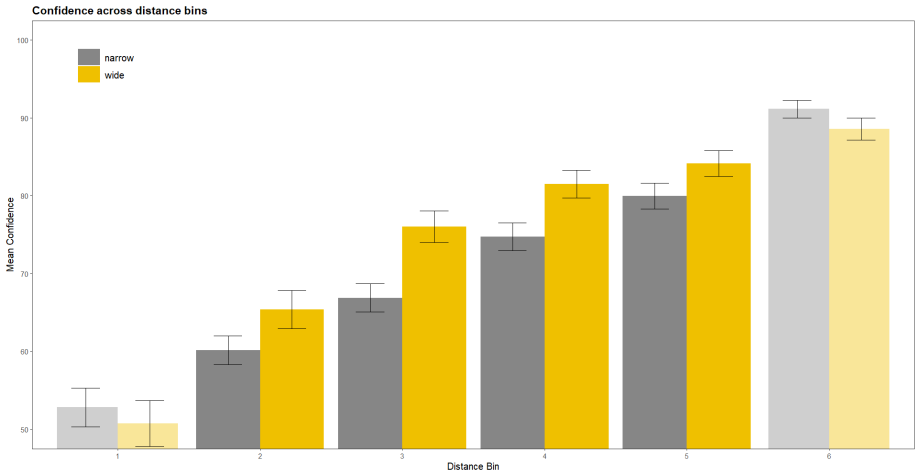


Figure 4.12: This bar plot visualises data from Study 2b relative to the objective boundary. It indicates the mean confidence on the y-axis (with error bars corresponding to SEM) and distance bin, here running increasing from left to right from 1-6 away from the objective category boundary at 1, on the x-axis relative to the objective category boundary. The confidence reports are split based on distribution, whether trials are drawn from a wide or a narrow distribution. The stimulus space was folded across the boundary to collapse from 11 location bins down to 6. Bins 2-5 were included in an ANOVA (see below). Bins 1 and 6 were not included in the ANOVA as they contained fewer trials and Bin 1, the middle bin which contained the objective boundary, was dummy-coded for accuracy so confidence could not be reliably decoded. Mean confidence in Bins 1 and 6, in each distribution, are plotted with a lighter shading to indicate that they are not included in the ANOVA.

As in Study 2, to check that these results are not a function of discordance between the experimenter-imposed boundary and the participant's subjective category boundaries, for each participant, the response-based subjective boundary was again established in this data set as part of the preregistered exploratory analyses. The boundary fitting process was mirrored from Study 2. The sub-

jective decision boundary of each participant was defined using response-based data – where participants responded as though a trial was coming from the wide as compared with the narrow distribution. An average location and percent-based response rate per bin per participant was then calculated and a sigmoid curve was fit to the resulting points. The point at which the sigmoid curve crosses 0.5 on the y-axis was established as the response cross-over point and the corresponding x-axis value describes the subjective decision boundary for that participant. Specifically, using NLS curve fitting, the start values were re-used from Study 2 barring a small adjustment to the start value for the halfway point which was shifted from 405 down to 400 to improve the number of convergences. The start values were therefore set as follows: lower asymptote 0.1, upper asymptote 0.9, slope β (utilised in the form -1/β) of 30, and midpoint on the x-axis of 400. These values fit all but 3 participants who had their mid-points established via linear interpolation as before in Study 2.

With these subjective midpoints established, the locations were re-binned around the subjective mid-point for each participant, and the confidence by distance bin and distribution ANOVA was re-run. Greenhouse Geisser corrections were applied to correct for violations of sphericity. The main effect of distribution [$F(1, 57) = 24.43$, $p < 0.001$] and the main effect of distance [$F(1.39, 79.29) = 91.86$, $p < 0.001$] were significant. The interaction between distance and distribution [$F(2.26, 128.85) = 1.4$, $p = 0.249$] was not statistically significant after the subjective, response-based re-binning of distance. These trends are detailed in Fig. 4.13. As in the previous study, the results are driven by

greater confidence in the "wide" trials with larger numerical trends seen in the mid-distance Bins 3-5. As before there is not a blanket higher confidence for "wide" in all bins, as seen in Bins 1 and 6.

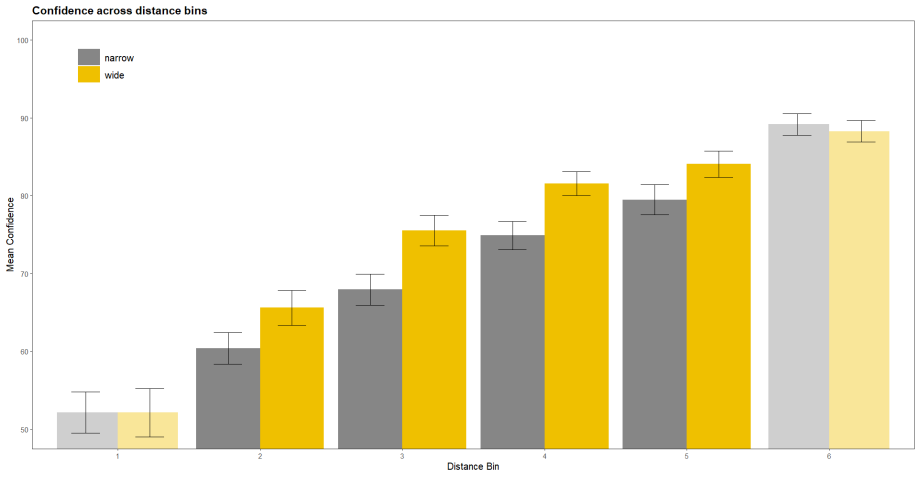


Figure 4.13: This bar plot visualizes data from Study 2b relative to the within-participant subjective boundary. It indicates the aggregated mean confidence per distance bin as binned within individuals relative to the subjective category boundary of their responses (error bars indicate SEM). Bars 1 and 6 are shaded out to indicate that they were not included in the ANOVA.

4.3.3 | Discussion

Overall, in line with the preregistered predictions, these data replicated the key patterns described in Study 2. Confidence increased with distance to both an objective and subjective version of the category boundary and was also sensitive to the variance of the distribution from which a trial was drawn. Interestingly, the interaction effect was less clear in these data, particularly when analysed with a subjective response boundary, with the trends being driven by the main effect of distribution. Consistent with the previous studies there was an overall

greater confidence for "wide" trials, particularly in the mid-distance bins 3-5. However, as before, this was not just a general increase in confidence for all "wide" trials, as categorisations of stimuli as being drawn from this distribution were not made more confidently in bins at the category boundary and very distant from it. Based on these findings, and in order to assess these effects outside of this stimulus set, a conceptual preregistered replication was carried out with details below.

4.4 | Study 3: Conceptual Replication

To understand these effects on confidence, beyond the scope of the original stimulus set, a preregistered conceptual replication with a new stimulus set was also conducted. Instead of sorting packages from spaceships, here the participants played a different kind of arcade-style game where they took on the role of marine biologists catching fish in different lakes. This maintained the interactivity and experience of the gamification but allowed a new stimulus set to be explored. Once again this study hypothesised that participants would report higher confidence on trials drawn from a wider, as compared with a narrower, category distribution and that sensitivity to distribution variance would be modulated by distance to the category boundary.

4.4.1 | Methods

The data set, data cleaning script, preregistration, and analysis script for this study are housed on the (OSF). The experimental scripts and relevant materials for running the data collection can be found on (GitHub).

Participant Sample

After exclusions were applied, a sample of $N = 71$ participants remained. Participants were compensated at a rate of £10.00 per hour for their time. After exclusions (see below), the $N = 71$ participants had an average age of 39 years (range $18 - 66$, standard deviation $= 12.8$). Self-identified gender splits were

female ($N = 32$), male ($N = 39$) and non-binary ($N = 0$). Handedness was reported as right ($N = 58$), left ($N = 10$), and ambidextrous ($N = 3$). The average time taken by participants to complete the experiment was approximately 32 minutes (minimum 18 minutes, maximum 81 minutes, standard deviation 10.7 minutes).

Exclusion Criteria

The exclusion criteria and sampling method matched that of Study 2b with the exception of an attention check at the start of the experiment which if failed twice, excluded participants prior to study completion. These participants are not represented in the final sample size: Their data was not saved as the attention check occurred at the beginning of the experiment to ensure the rules of the game were understood correctly.

A total of $N = 76$ participants passed the attention check. Two of these participant datasets were incomplete due to a server issue, leaving $N = 74$ prior to the application of the preregistered exclusion criteria. As there was no Phase 0, the exclusion criteria were set based on the accuracy of Phase 2 active testing trials and the use of the confidence slider. No participants were identified as using the confidence slider in a way which suggested non-compliance, here defined as using less than 3 unique values for confidence across all trials. Just 3 participants had accuracy below the 55% cut-off in the phase 2 trials and were excluded.

Therefore, after exclusions were applied, a sample total of $N = 71$ remained.

This sample was higher than the preregistered requirement which was set to 55 with data again collected in batches of 10 up to a maximum of 100. The over-sampling occurred due to an undetected change in the Prolific re-sampling procedure for participants who did not complete the study. Due to this error, no analysis was done prior total stoppage of the data collection and the final sample size was larger than expected.

Experimental Design

The new stimulus set consisted of categories of differently sized fish that needed to be sorted by the participant into either a native or an invasive species type. The species type was defined based on the size of the fish from end to end of the ovular body drawn on the screen, excluding the tail element which was the same on all stimuli. Rather than avoiding bombs, participants were asked to avoid returning invasive species to a lake where they would cause damage to the ecosystem of the native species. The reward matrix was established in a similar way to Study 2 and 2b, with the participant winning 3 coins for correctly returning a native fish, losing 3 coins for incorrectly returning an invasive fish and no coin changes on trials where native fish were incorrectly caught or correctly catching an invasive fish.



Figure 4.14: Visual of the screens viewed by the participant in Study 3. Note the 2D fish stimuli of a certain length which should either be caught and removed from the lake or returned to the lake. Note also the change in response button labels and the consistent use of the confidence scale and scale labels on the confidence reporting screen.

The study design proceeded in a way that largely replicated the methods of Study 2 and 2b. The participants were asked first to read some instructions and, in lieu of the Phase 0 associative learning seen previously, a simple attention check was carried out to ensure the participants knew the correct stimulus-response pairing for the game. This attention check took the form of a four-trial mini-block where participants were shown a fish on screen with the label "native" or "invasive" present on screen alongside them. The participants had to use the correct response button on every trial to achieve a pass on the attention block. If they failed, the response pairings and instructions were reiterated and participants had a second chance to complete the attention check. If they failed a second time they were removed from the study, as per Prolific guidelines. If

they passed at either stage, Phase 1 and Phase 2 of the study then proceeded as before.

Instead of watching spaceships dropping packages on screen, Phase 1 consisted of a series of fish stimuli appearing along with a label of either "native" or "invasive" printed onscreen. The participants were passive observers of this phase, just like the Phase 1 passive learning in Studies 2 and 2b. In this way, the participants learned which size of fish was associated with which species type in that block. For example, in one block the native species might be bigger on average but with a narrow variance in size across exemplars, and when the participants moved to the next block, as distinguished by being told they were moving to a new lake, the native species might be smaller with a wide variance. The order of the blocks and stimulus-distribution-response matrix was counterbalanced and pseudo-randomised within and across participants to prevent order effects. In this way, all participants worked through the four possible combinations of the distribution variance, stimulus size, and category – but in a randomised order.

The means of the distributions were shifted in space compared with Study 2 and Study 2b. The total screen size was 1200 pixels, of which 1000 pixels was the maximum stimulus size desired. This means that the category boundary should represent a stimulus that is half the maximum size if that 100-pixel buffer was to be maintained on either side. To allow for optimal use of the screen space, given the centrally presented stimulus on each trial, the category boundary between the distributions was therefore set to 500 pixels. Accordingly, the

means were set to 329 and 629, or 371 and 671, depending on the distribution variance allocation as being for the larger or smaller species. In the active testing phase (Phase 2) the participants viewed the fish stimuli without the species labelling and had to respond to them according to their understanding of the distributions and provide a confidence report on each trial. Thus, as in the previous studies, participants' confidence reports should depend on the distributions observed in the initial passive learning phase, with no opportunity to learn via feedback about particular categorisation responses. The underlying distribution variances, reward matrix, feedback frequency, trial number, and confidence reporting were all the same as in Study 2 and 2b.

4.4.2 | Results

Broadly speaking, the results were in the same direction and of similar strength to studies 2 and 2b. In terms of accuracy and task performance, after exclusions, the average accuracy on test trials was 83.5%. The analysis also indicated that average accuracy increased with binned distance from the category boundary, F(2.65,185.39) = 357.26, $p < 0.001$. Looking at confidence reports, confidence was higher ($t(70) = 13.56$, $p < 0.001$ – mean confidence for correct trials = 77, mean confidence for incorrect trials = 60) for correct versus incorrect trials. However, correct trials were not significantly faster in this data set ($t(70) =$ -1.29, $p = 0.19$ – mean confidence reaction time for correct trials = 2086ms, mean confidence reaction time for incorrect trials = 2139ms). These results suggest that confidence ratings were calibrated in a sensible manner in the new

task space and on the perceptual dimensions of the stimulus set.

To test the critical preregistered predictions, a two-way ANOVA was carried out on mean confidence ratings as a function of distance to bound bin (bins 2-5 only) and distribution (narrow vs. wide). Greenhouse Geisser corrections were applied to correct for violations of sphericity. The main effect of distribution [$F(1, 70) = 6.02$, $p = 0.017$] and the main effect of distance [$F(1.41, 98.46) = 100.02$, $p < 0.001$] were both statistically significant. This replicates the distance-to-bound effect and also confirms the main prediction that confidence should be higher in trials drawn from the wide versus the narrow distribution. However, unlike Studies 2 and 2b, the interaction between distance and distribution was not significant [$F(2.56, 179.03) = 2.56$, $p = 0.425$]. This is contrary to the preregistered prediction of finding a significant modulatory effect of distribution on confidence. Furthermore, looking at Fig. 4.15, numerically the differences between the wide and narrow bars are also less than in the previous studies. It is possible that the shift in dimension from the location on the screen to fish size made the distribution properties less salient to the participants.

Furthermore, numerically, the increase in confidence for "wide" categorical decisions was similar across all bins and was even apparent in Bins 1 and 6. This contrasts with the previous data where the extreme bins (1 and 6) showed higher confidence in the “narrow” trials than the “wide” on average. In sum, these results are broadly in line with the preregistered predictions and replicate the main findings of Study 2 and 2b, however the lack of interaction effect was contrary to the study predictions, see Fig. 4.15.

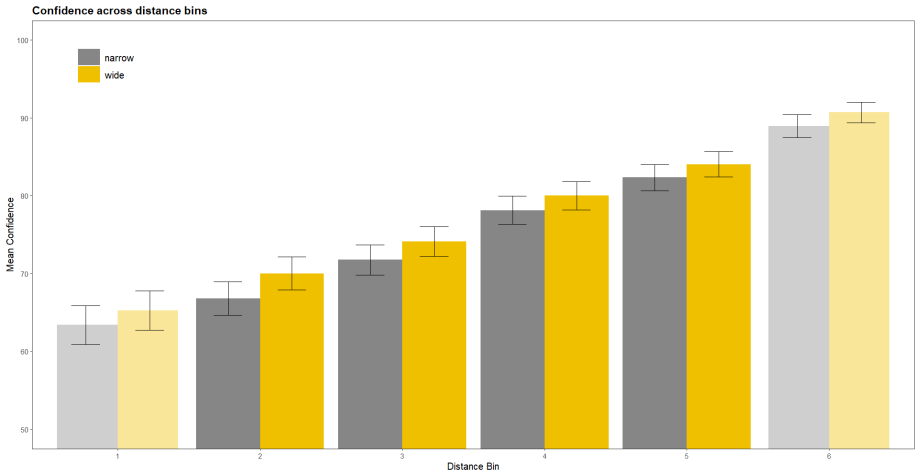


Figure 4.15: This bar plot visualises data from Study 3 relative to the objective boundary. It indicates the mean confidence on the y-axis (with error bars corresponding to SEM) and distance bin, here running increasing from left to right from 1-6 away from the objective category boundary at 1, on the x-axis relative to the objective category boundary. The confidence reports are split based on distribution, whether trials are drawn from a wide or a narrow distribution. The stimulus space was folded across the boundary to collapse from 11 location bins down to 6. Bins 2-5 were included in an ANOVA (see below). Bins 1 and 6 were not included in the ANOVA as they contained fewer trials and Bin 1, the middle bin which contained the objective boundary, was dummy-coded for accuracy so confidence could not be reliably decoded. Mean confidence in Bins 1 and 6, in each distribution, are plotted with a lighter shading to indicate that they are not included in the ANOVA.

The response-based subjective boundary was again established in this data set as part of the pre-registered exploratory analyses. The boundary fitting process was mirrored from Studies 2 and 2b. These values fit all but 3 participants who had their mid-points established via linear interpolations as before in Study 2. With these subjective midpoints established, the locations were re-binned around that mid-point within each participant and the confidence by distance bin and distribution ANOVA was re-run. Greenhouse Geisser corrections were applied to correct for violations of sphericity. The main effect

of distribution [$F(1, 57) = 24.43$, $p < 0.001$] and the main effect of distance [$F(1.39, 79.29) = 91.86$, $p < 0.001$] were significant. The interaction between distance and distribution [$F(2.26, 128.85) = 1.4$, $p = 0.249$] was not statistically significant after the subjective, response-based re-binning of distance. The numerical trends were in line with the previous ANOVA on the objective category boundary, however, the difference in confidence reports was greater between the "wide" and "narrow" trials in the subjective space, see Fig. 4.16 for details.

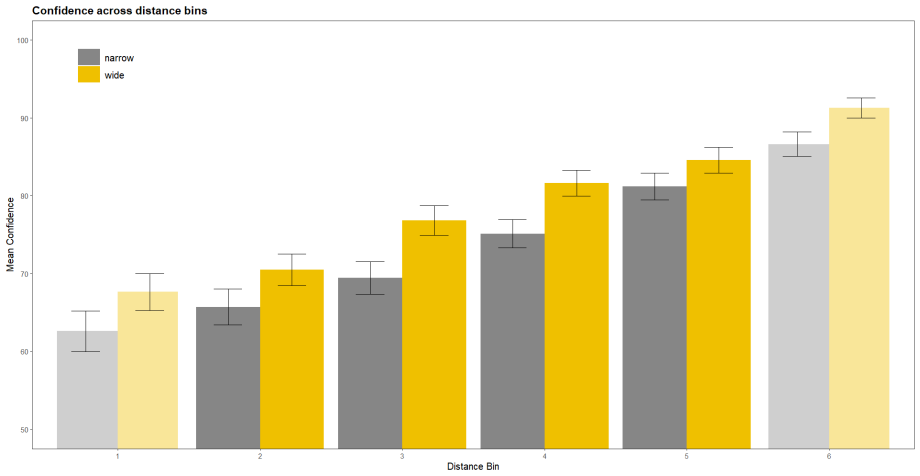


Figure 4.16: This bar plot visualises data from Study 3 relative to the within-participant subjective boundary. It indicates the aggregated mean confidence per distance bin as binned within individuals relative to the subjective category boundary of their responses (error bars indicate SEM). Bars 1 and 6 are shaded out to indicate that they were not included in the ANOVA.

4.4.3 | Discussion

This study aimed to replicate the findings of Study 2 and 2b in a new stimulus set. Moving away from categories defined by location on the screen, here categories were made up of fish stimuli defined by the length of their body.

Overall, the confidence reports showed sensitivity to both distance-to-bound and the variance of the distribution from which a trial was drawn. However, unlike previous studies, there was a blanket increase in confidence in "wide" distribution trials across all distance bins and no interaction effect was found in the mid-distance bins used for the ANOVA. It is possible this change is related to the shift in category dimension which may have made the distribution properties less salient to the participants.

4.5 | Model-based Analysis

As a final exploration of these data, I returned to the modelling approaches leveraged in Chapter 3. Based on the robust confidence differences elicited by the experimental designs, it is expected that the data should clearly differentiate models. To this end, I compared a Bayesian model, which had access to full and accurate distribution information, against a simpler comparator model that is likened to a distance-to-bound model.

The main Bayesian model included information about the true distribution properties including the means and unique variances which identified the "wide" and "narrow" distributions in this chapter. Due to the complexity of the resulting posterior, the model pipeline established in Chapter 3 became a useful parallel to capture the behaviour of this model. A comparison model which assumed that the participant representation of the stimulus space was comprised of a pair of symmetrical distributions centred on the distribution midpoints was also created. The idea was to capture the relative sensitivity in the participant data to the asymmetry of the true underlying distributions and to understand to what extent the participants simplified their representations. By creating a symmetry in the comparator model, the slope of the predicted confidence became akin to the predictions of a distance-to-bound model of confidence.

4.5.1 | Model Derivation

The model derivation of confidence mirrored that of the normative model in Chapter 3. However, the computation of confidence was expanded to allow for different variances to be captured by the main model.

The models were derived as follows. Assuming first that the observers have the same model of measurement, x, as in Chapter 3, a category judgement must also be made. That is, on a given trial, an observer must also decide whether the stimulus belongs to Category 1 or Category 2. If the observer has a perfect perceptual system the observer must then decide whether s is an example of category 1 stimuli or an example of category 2 stimuli, both of which follow a normal distribution centred on their respective means. Importantly, unlike in Chapter 3 - these distributions have a different value of σ_2 to mirror the design features of this chapter.

That is,

$$p(s|C=1) = N(s; -a, \sigma_{s_1}^2) \tag{4.1}$$

$$p(s|C=2) = N(s; a, \sigma_{s_2}^2) \tag{4.2}$$

$$p(C=1) = p(C=2) = \frac{1}{2} \tag{4.3}$$

Given that, by the laws of probability, we have:

$$\frac{p(C=1|s)}{p(C=2|s)} = \frac{p(s|C=1)}{p(s|C=2)} \tag{4.4}$$

$$= \frac{N(s;-a,\sigma^2_{s_1})}{N(s;a,\sigma^2_{s_2})} \tag{4.5}$$

Therefore, expanding out the equation for the normal distribution and taking the log of both sides we get:

$$\log\frac{p(C=1|s)}{p(C=2|s)} = \log\frac{\frac{1}{\sqrt{2\pi\sigma^2_{s_1}}}e^{-\frac{(x+s)^2}{2\sigma^2_{s_1}}}}{\frac{1}{\sqrt{2\pi\sigma^2_{s_2}}}e^{-\frac{(x-s)^2}{2\sigma^2_{s_2}}}} \tag{4.6}$$

However, an observer may face both unreliable evidence and unreliable perception. So the observer only has a noisy measurement, x, of the stimulus, s. Furthermore, the stimulus, s, might be only partially informative of the category. This can be described as having distributions around the measurement of the stimulus, and the stimulus itself:

$$p(x|s) = N(x;s,\sigma^2) \tag{4.7}$$

$$p(s|C=1) = N(s;\mu_{s_1},\sigma^2_{s_1}) \tag{4.8}$$

$$p(s|C=2) = N(s;\mu_{s_2},\sigma^2_{s_2}) \tag{4.9}$$

Given these noisy distributions, we must therefore work to compute the

posterior probability of $C = 1$ and $C = 2$ given a noisy measurement x. That is,

$$\frac{p(C=1|x)}{p(C=2|x)} = \frac{p(x|C=1)}{p(x|C=2)} \tag{4.10}$$

Given that;

$$p(z) = \int p(z|y)p(y)dy \tag{4.11}$$

$$\frac{p(C=1|x)}{p(C=2|x)} = \frac{p(x|C=1)}{p(x|C=2)} \tag{4.12}$$

$$= \frac{\int p(x|s,C=1)p(s|C=1)ds}{\int p(x|s,C=2),p(s|C=2)ds} \tag{4.13}$$

As x is conditionally independent of C, given s, $p(x|s,C=1)$ is identical to $p(x|s)$, (4.7), (4.8) and (4.9) can be substituted into the formula.

$$\frac{p(C=1|x)}{p(C=2|x)} = \frac{\int p(x|s)p(s|C=1)ds}{\int p(x|s),p(s|C=2)ds} \tag{4.14}$$

$$= \frac{\int N(x;s,\sigma^2)N(s;\mu_{s_1},\sigma^2_{s_1})ds}{\int N(x;s,\sigma^2)N(s;\mu_{s_2},\sigma^2_{s_2})ds} \tag{4.15}$$

Computing the product of two normal distributions results in a denominator of,

$$\frac{S_{s_{2,1}}}{\sqrt{2\pi}\sigma_{s_{2,1}}} \exp - \frac{(x-\mu_{s_{2,1}})^2}{2\sigma^2_{s_{2,1}}} \tag{4.16}$$

Where $\mu_{s_{2,1}}$ is the mean of the resulting normal distribution, and $\sigma_{s_{2,1}}$ is the standard deviation of the resulting normal distribution. $S_{s_{2,1}}$ is a scaling factor and is itself a Gaussian PDF

and a numerator of,

$$N(x; s, \sigma^2)N(s; \mu_{s_1}, \sigma^2_{s_1}) = \tag{4.17}$$

$$\frac{S_{s_{1,1}}}{\sqrt{2\pi}\sigma_{s_{1,1}}} \exp - \frac{(x-\mu_{s_{1,1}})^2}{2\sigma^2_{s_{1,1}}} \tag{4.18}$$

Where the formula for $\sigma_{s_{x,1}}$ is,

$$\sigma_{s_{x,1}} = \sqrt{\frac{\sigma^2_{s_x}\sigma^2}{\sigma^2_{s_x} + \sigma^2}} \tag{4.19}$$

and the formula for $\mu_{s_{x,1}}$ is,

$$\mu_{s_{x,1}} = \frac{\mu_{s_x}\sigma^2 + \mu\sigma^2_{s_x}}{\sigma^2_{s_x} + \sigma^2} \tag{4.20}$$

and the scaling factor S is a Guassian PDF on μ_{s_x} and μ with standard

deviation $\sqrt{\sigma_{s_x} + \sigma}$,

$$S_{s_x,1} = \frac{1}{\sqrt{2\pi(\sigma^2_{s_x} + \sigma^2)}} exp\big[- \frac{1}{2} \frac{(\mu_{s_x} - \mu)^2}{\sigma^2_{s_x}\sigma^2} \sigma^2_{s_x,1}\big] \tag{4.21}$$

This results in the following:

$$\frac{p(C=1|x)}{p(C=2|x)} = \frac{\int \frac{S_{s_1,1}}{\sqrt{2\pi}\sigma_{s_1,1}} \exp - \frac{(x-\mu_{s_1,1})^2}{2\sigma^2_{s_1,1}} ds}{\int \frac{S_{s_2,1}}{\sqrt{2\pi}\sigma_{s_2,1}} \exp - \frac{(x-\mu^2_{s_2,1}}{2\sigma^2_{s_2,1}} ds} \tag{4.22}$$

This posterior ratio allows for the model to hold different values for the mean and variance of each distribution. However, it does not require this and a second posterior ratio could be derived which holds the variances to be symmetrical and the means to be mirrored around the midpoint and equidistant from one another. This allows for a simple model comparison – where the participant representations and resulting confidence values are either inclusive of the asymmetry of the distribution properties or they are insensitive to these differences. This posterior ratio is then related to confidence in the same way as in Chapter 3.

4.5.2 | Model Fitting & Results

In order to fit the two models to the data the slow simulations-based pipeline was replicated from Chapter 3. The code for these analyses can be found on this GitHub Repository. These simulations were conducted using Inverse Bi-

nomial Sampling (IBS; (van Opheusden et al., 2020)) in MATLAB (MATLAB R2020b) to simulate the log-likelihood on each trial. The features of the model extension detailed in Chapter 3, such as fitting to the response and confidence simultaneously, and the inclusion of parameters like metacognitive noise and lapse rates were upheld. The simulations were adjusted to suitably represent the current stimulus set by removing the graded experimenter-imposed stimulus noise, changing the values of the generated stimuli, and moving the expected perceptual noise out of the space of circle statistics. The data set from Study 2 was used in this model-fitting exercise. The confidence reports were binned into 5 bins prior to exporting the dataset into MATLAB.

As before, in Chapter 3, the model fitting procedure was conducted using the complementary Bayesian Adaptive Direct Search (BADS; (Acerbi & Ma, 2017). The parameters to be fitted and the values of the bounds set on the parameter spaces are shown below, see Table 4.1.

Free Parameter Initialisation & Bounds			
Parameter	Start Point	Upper | Lower Bounds	Plausible Bounds
$Sigma_X$	Random [PLB:PUB]	100 - 1/1000	10 - 14
$Thresholds$	Random [PLB:PUB]	$0.99 - 0.01$	$0.9 - 0.1$
$LapseRate$	Random [PLB:PUB]	$1 - 0.01$	$0.45 - 0.01$
$MetaNoise$	Random [PLB:PUB]	$4 - 0.0001$	$2 - 0.0087$
$ConfLapse$	Random [PLB:PUB]	$1 - 0.0001$	$0.1 - 0.01$

Table 4.1: Parameter space established for BADS to search for the maximum likelihood.

The output of the BADS and IBS model fitting procedure was in the form of 16 fitted parameter values and one log-likelihood of the model fit for that participant per run per model per participant. The try counts were first checked

to ensure that no model-fitting attempts had failed to converge. The results of the extended model fitting are displayed in Fig.4.17. The plot shows the Δ BIC values for each model relative to the overall best-fitting model - Model 1. The Bayesian model accounted for the asymmetry of the distribution variances (Model 1) and appeared to fit the patterns of the data better than the simplified comparator model (Model 2). Individual best-fitting models, per participant, were also calculated. Of the N = 52 participants that were fit to these two models N = 45 were best fit by Model 1 and only N = 7 were best fit by Model 2.

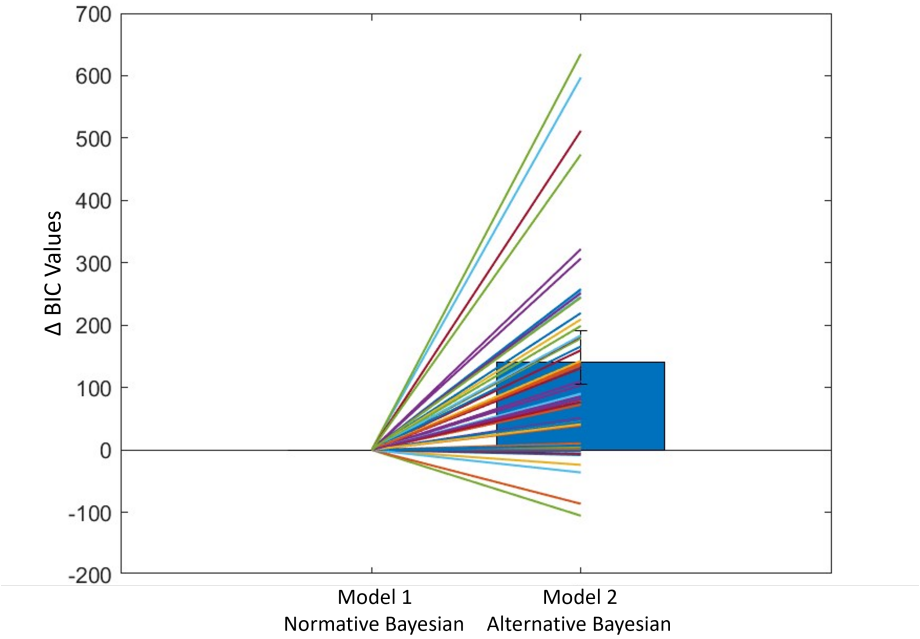


Figure 4.17: The comparative fit of the two models is shown. The y-axis indicates the Δ BIC of the models, with lower values indicating a better fit. Models are numbered 1-2 with the Bayesian model that accounted for the asymmetry of the distribution variances (Model 1) being identified as the best fit as compared with the simplified comparator model (Model 2), which approximated a simple distance to bound representation of the task space. Individual coloured lines indicate the fits of the alternative models, per participant, relative to the best-fitting model.

4.5.3 | Discussion

This model-fitting exercise served two purposes. Foremost it tried to systematically pull apart the predictions made by a normative Bayesian and a simplified model that approximated a distance-to-bound approach to confidence computation. Secondly, it served as a test of the model pipeline used in Chapter 3 in a stimulus set with lower perceptual noise and whose category values had been

selected to maximise differences in confidence as predicted by different models.

The results indicate that the normative Bayesian model which leveraged information about the asymmetry of the category variances was the best-fitting model both in terms of the BIC values and in terms of the number of participants from whom the model was the best fit. This is not to say that the distance-to-bound model is overtly incorrect, but it seems that the sensitivity shown by the normative model to the variances of the distributions is an important feature of the confidence reports in the dataset. This indicates that metacognitive confidence reports can encompass information about category distributions and reflect the learned category space with some accuracy. This maps well to the behavioural results of all three of the studies in this chapter which demonstrated a sensitivity to both the distance and the distribution variance in participant confidence reports. These ideas will be explored further below.

4.6 | General Discussion

The studies in this chapter used manipulations of category variance as a targeted way to contrast and explore Bayesian and heuristic confidence computations. The structure of category distributions, and the relevance to underlying models of confidence that map to category judgements, was explored in three related behavioural studies and accompanying model-based analyses. The experimental designs used in these studies leveraged a difference in the category distribution variances to explore specific predictions about confidence.

It was predicted that the effect of distance to the category boundary, as seen in the gradient of the confidence reports, would differentiate distributions of different spread, with a steeper gradient and numerically lower confidence for narrow as compared with wide category distribution trials. In three studies, an initial exploratory behavioural study, a direct preregistered replication, and a preregistered conceptual replication, these predictions were evidenced in numerous complementary analyses. Importantly, the significant interaction that emerged in both the objective boundary and subjective boundary analyses effects of the first two studies, as well as the corresponding numerical trends, suggested that participants were not just generally more confident when making decisions for the "wide" category.

In terms of the theoretical implications of this work, as described before, Bayesian models of confidence inherently utilise the values of the category variances to produce the trial-by-trial confidence reports (Beierholm, 2013). How-

ever, the predictions of a Bayesian model contrast with those of a simple heuristic "distance to bound" model of the kind that have previously been applied to studies of perceptual decision confidence, (Kepecs & Mainen, 2012; Kepecs et al., 2008; Komura et al., 2013). Distance-to-bound models do not necessitate that the computations of confidence account for differences in distribution variances. Collectively, the above-described results suggest that participants are not only sensitive to the metric of distance-from-bound but instead, as predicted by Bayesian accounts of confidence and decision-making, seem to be sensitive to the distribution variance as well. This conclusion was further supported by the results of the model-based analyses where a normative Bayesian model with accurate information about the asymmetry of the distribution variances was a better fit to the data than a simplified comparator model which assumed a symmetry of variances. However, these results do not suggest that participants are necessarily ideal Bayesian observers. It is entirely possible, and indeed plausible, that the underlying mechanisms of these computations are not truly Bayesian in the strictest senses (Geisler & Ringach, 2009).

Importantly, these results also do not suggest that a distance-to-bound model is a poor descriptor of confidence in perceptual decision-making or categorical judgement tasks per se. However, they do suggest that a model which includes the distribution properties such as variance better describes the confidence reports of the participants. That is to say that the computation of metacognitive confidence appears to be sensitive to, have access to, or include the use of, these distribution properties. In turn, this means that confidence reports may be a

useful lens for investigating the representations held by participants of the categories and concepts that they navigate in the world. This finding is relevant to ideas regarding the improvement of calibration in real-world decisions (such as in finance (Atir et al., 2015) and healthcare (Friedman et al., 2005)) – if we can infer from confidence reports how experts are representing (or misrepresenting) the uncertainty that is intrinsic to the distribution properties they are working with then we can begin to address those misalignments. Importantly, in terms of fieldwork in settings such as healthcare centres when studying medical decision-making, confidence reports are also an intuitive and internally reliable variable that can be collected in real time. It is arguably more reliable than variables such as decision reaction time when considering the longer time scale of, for example, the diagnostic process of our dermatologist examining a dubious freckle on a patient's forearm (for a specific applied example of this kind of decision-making and confidence reports see Hasan *et al* (2023)).

Also of note in this chapter is the success of the analyses in distinguishing different task representations relative to the approach taken in Chapter 3. By directly manipulating distributions and eliciting related confidence reports, the extent to which the participants aligned their representations to the objective categories could be tested. This approach appeared to be more successful than trying to infer differences in representation from behavioural data that emerged from a more standard task. The structure of the task itself allowed the models of confidence to be tested explicitly.

This approach also highlights the utility of engaging with categorical de-

cision spaces more broadly. As noted in the introduction, the scope of structures of category spaces is more complex than that of classic perceptual decision tasks which are often structured with a distinct boundary that divides a one-dimensional uniform stimulus space. These possible structures include categories with discrete boundaries, asymmetrical classification rules, and multi-dimensional spaces. These kinds of spaces, and the implications they have for both distance-to-bound and Bayesian models of confidence, are explored in the next chapter of this thesis.

CHAPTER 5

Discrete Categories

Abstract

This chapter investigates decision confidence patterns in discrete multi-dimensional category spaces. In doing so, this chapter aims to test the distance-to-bound and normative theories of confidence in spaces where the optimal way to navigate ideas of distance and decision rules is more complex. The stimulus space leveraged in the tasks in this chapter is asymmetrical in both a literal and figurative sense. This means that not only are the stimulus values not symmetrical around the boundary, but also the facets of the stimuli that the participant should attend to vary across the space too. Using this space, an initial exploratory study and direct confirmatory replication were conducted, together with exploratory model-based analyses of "distance" metrics in the task space.

The findings of this chapter suggest that confidence reports are sensitive not only to the decision boundary placement in a discrete categorisation task but also to the different types of classification strategies that observers employ in multi-dimensional spaces – to integrate dimensions or to classify stimuli using a single-dimension when appropriate. Importantly, these differences in strategy appear to play out within individual participants across the category space.

5.1 | Introduction

Contrary to the tasks used in the previous experimental chapters, in the real world, categories are often multi-dimensional. That is, they are usually defined as being distinct from one another on the basis of more than one dimension. For example, stepping back to the idea of a dermatologist working to categorise a blemish as malignant or benign, it would be unusual for a decision to be made on the basis of a single factor alone, such as colour. Instead, the decision to seek further information, or create a treatment plan, would emerge from a combination of dimensions including, for example, the size, shape, and symmetry as well as the colour of the blemish.

Another example would be simply trying to categorise items of fruit as either apples or pears based on both the shape and colour of each item. Apples are usually more reddish and roundish, though they can also be green. Pears are often more greenish and ovular, but they are rarely reddish in hue. The two dimensions in combination can be used to more accurately categorise the items than a single dimension could. Even our prior example of rain clouds is again pertinent if we think about both cloud colour and cloud density or cloud colour and cloud height in the atmosphere as bearing relevance to the outcome of rain or dry weather. Perhaps dark, dense, and low-hanging clouds are more likely to result in rain than the opposite. But, perhaps, the height of the cloud weighs more heavily on the outcome than either the density or colour – which in themselves might be highly correlated. By taking this information together,

and combining the category dimensions, a better judgement can be made about whether to bring an umbrella to work or to risk leaving it at home! Indeed, across most daily categorisation tasks, there exist multiple intersecting dimensions that should be used in tandem to help define one category as distinct from another.

We can distinguish these everyday examples further by thinking about multi-dimensional categories where there are both continuous and distinct dimensions at play. For example, we could look at categorisation when differentiating vessels as being either mugs or bowls. On the one hand, we have continuous variation in the circumference and height of these vessels with bowls tending to be shorter and wider than mugs. On the other hand, it is a fairly universal rule to suggest that if a vessel has a single handle, it is no longer considered a bowl – it is a mug. This example serves to illustrate two points. Foremost, it demonstrates that there are categories where there exists a discrete bound between them. This stands in contrast to many perceptual tasks, including those designed and deployed in the previous chapters of this thesis. Secondly, this example shows that there can exist an asymmetry or hierarchy between and across category dimensions and definitions. In the case of mugs and bowls, while it is normally true that the wider the diameter the more likely it is to be a bowl, however, once a handle is present the diameter dimension can be ignored and the correct assessment is clear – it is a mug.

Many decision-making tasks leverage perceptual categories that are defined by a single dimension, however, these tasks do not capture the richness that exists in everyday multi-dimensional categories. By contrast, within the cat-

egorisation literature, multi-dimensional categories have been utilised to distinguish between different models of decision-making and categorisation. In prior research, these kinds of stimulus spaces – with multiple dimensions, discrete bounds, and an asymmetry of decision rule application – have been used to explore logical rule models of categorisation (Little et al., 2011).

Classic models of categorisation have centred on the learning and application of logical rules (Bourne, 1970). With the shift towards encompassing the potential multi-dimensionality of categories, further model types emerged such as prototype (Posner & Keele, 1968, 1970) or exemplar (Medin & Schaffer, 1978). However, with the integration of reaction time into rule-based models, as part of understanding logical models in isolation or as part of larger cognitive systems (Little et al., 2011), these models became a useful tool to parse out different types of processing and categorisation strategies by focusing on the time-course of each decision (Fific et al., 2010).

The broad idea behind these logical rule models is that decision-makers may complete a series of decision judgements along each dimension of a category and then combine these decisions together using logical rules to make a final categorisation judgement. Traditionally reaction time has been modelled in these spaces to judge the serial or parallel nature of the independent decision application (Fific et al., 2010; Little et al., 2011). Importantly, in these category spaces, observers may use either exhaustive or self-terminating stopping rules when deciding whether the logical rule has been satisfied, that is they may not necessarily evaluate all decision paths across all dimensions (and their

combination) when evaluating the decision rules (Little et al., 2011).

An example of a stimulus space used to explore these logical models in multidimensional categories is detailed in Fig. 5.1. These stimuli were used to test the way in which observers were processing the dimensions of the stimuli in order to classify them as either belonging to category A or B. Classification was based on two spatially separated dimensions, the width of the base and the curve of the top of each lamp stimulus. Importantly, upon examination, to classify a stimulus as A, it is necessary to pay attention to both dimensions (wider base AND curved top piece), but to identify a stimulus as belonging to B it is enough to check if it has a narrow base OR a flat top piece. The evidence suggested that individuals processed the dimensions in a serial manner but that the processing was self-terminating with the decision being made once the logical rule was satisfied rather than once all possible dimensions were analysed. This was seen in the success of the serial-self-terminating processing models fit the reaction times of the participants in these spatially separated dimensions, resulting in faster mean reaction times for trials drawn from Category B than A. In an exploratory study that replicated the categorisation task using a stimulus set that had spatially overlapping dimensions (rectangles in shades of red with black vertical lines in different positions), a self-terminating process was still favoured by participants but a mixture of serial and parallel decision processes were found to fit the data (Little et al., 2011).

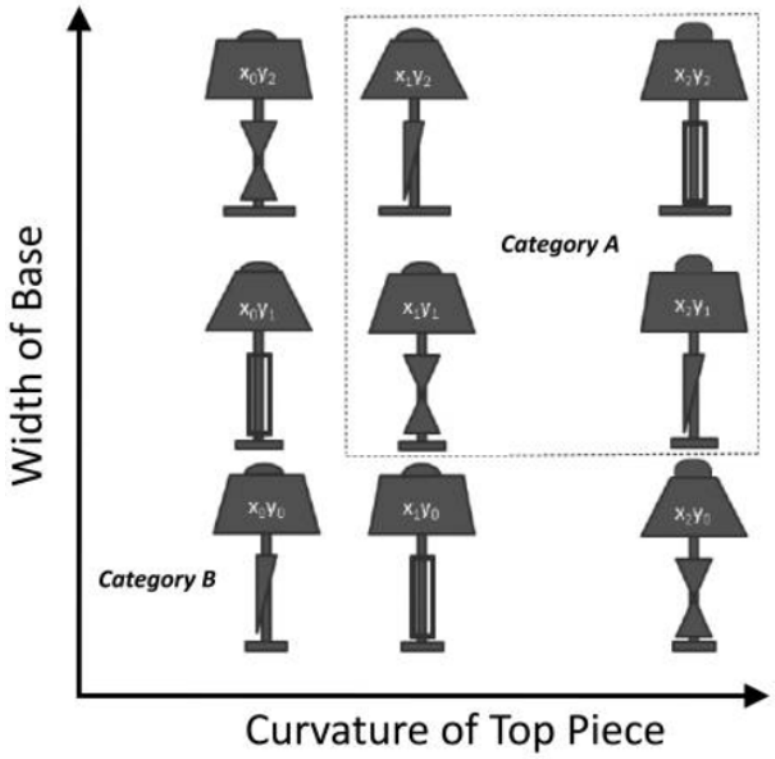


Figure 5.1: Reproduction of Figure 4 from (Little et al., 2011) used with permission from the APA Journal of Experimental Psychology: Learning, Memory, and Cognition (License Number: 5541830268190). It illustrates the stimuli used in previous studies which were mapped across two dimensions – lamp curvature and base width. The two stimulus categories, A and B, were defined along an L-shaped boundary in the space which divides the space in a non-intuitive way.

How these findings translate to theories and computations of metacognitive decision confidence is less clear. In these spaces, the definition of distance in the distance-to-bound model explored in previous work is perhaps insufficient. While a reaction-time-distance model, much like the distance-to-bound confidence model, applies well in a single bounded space, a more complex model may be required in spaces with multiple bounds. However, there is scope to adapt the distance-to-bound model here too. The question is what is the distance

that we should use to judge as a proxy of confidence? In a two, rather than a one, dimensional space it is unclear if the "distance" unit is best defined as the maximum, minimum, or combined distance to the boundaries on each stimulus dimension. Similarly, how confidence might behave if there is a discrete boundary between the categories, as opposed to a graded perceptual overlap, is also unclear. Would it be expected that confidence would reflect the discontinuity in the stimulus space, and can this be used to help map out the cognitive representation of the category space effectively? In order to address some of these uncertainties, an exploratory study was designed. This study aimed to replicate the stimulus space used in previous research, but now asking for confidence reports in addition to categorisation decisions.

This chapter and the experimental data within it seek to address some of these queries and relate the findings to our current understanding of the distance-to-bound model of confidence. Therefore, building on the previous experimental work of this thesis, this chapter aims to explore and map confidence in a discrete category space where two discrete categories are defined based on a distinct and objectively implemented decision boundary. This allows for a reduction in the amount of noise present in confidence reports near the boundary and may enable a clearer understanding modelling of confidence to occur. Furthermore, in this work, the stimuli are defined across two different perceptual dimensions. This means that the decision boundary is established based on the intersection of those two dimensions. This allows for a more complex model of confidence in categorical choices to be explored that is closer to the

kinds of multi-dimensional categories that are encountered in daily life. This complexity also gives way to a set of testable predictions about how and when each dimension, or their combination, is used to help decide with some level of certainty. Finally, by manipulating the geometry of this space we can build an asymmetry into the stimulus space which challenges the distance-to-bound model of confidence in a way that the one-dimensional categories do not. This notion will be explored in more detail when the stimulus space is explained in depth.

5.2 | Study 4

5.2.1 | Introduction

Study 4 leverages the category spaces used by Little *et al.*, (2011) to investigate the behaviour of metacognitive decision confidence in a discrete multi-dimensional category space. Importantly, in this space, the nature of categorisation is very different. Not only are the boundaries between the categories distinct, but there is also more than one bound, and their intersection, to navigate.

As with the experimental design of Study 3, fish stimuli were utilised in this study. However, unlike Study 3, here, in Study 4, the stimuli were drawn from a stimulus feature space that was defined across two dimensions. The fish were defined based on their size (small to large body length) and colour (red to blue), see Fig. 5.2 for details. On this continuous feature dimension space, a three-by-three grid-like sampling structure was superimposed. In each grid box, the stimuli were sampled in a uniform manner. Thus, while the total category space can be subdivided into grid boxes for sampling and analysis, this division is artificial as the participant experience was continuous in nature. A discrete category boundary is defined in that space as an "L-shaped" boundary that separates out a square of grid boxes from an L-shaped category along that border. For clarity, throughout this chapter, the L-shaped category space is named the "L-category" and the corner category is the "□-category".

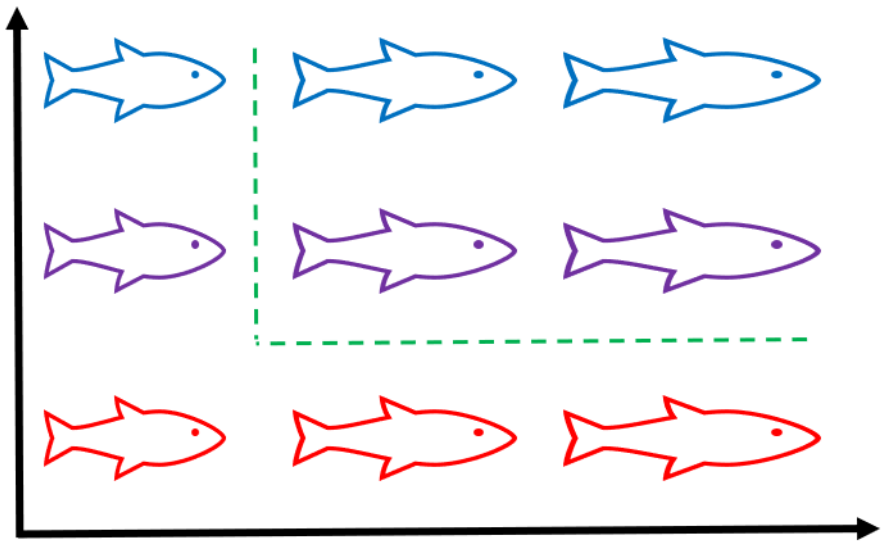


Figure 5.2: The stimulus space for this study is mapped out here in a way that parallels the stimuli of (Little et al., 2011). The fish stimuli change in terms of their body length along the x-axis and their colour hue along the y-axis. Nine exemplar fish stimuli are shown with each one being representative of a particular grid-box space, but note that stimulus values shown in the actual experiment were sampled from the full continuous dimensions. The category boundary is defined by the L-shaped green dashed line which separates four grid boxes in the top right (the □-category) and the five grid boxes that make up the L-shaped category on the left (the L-category).

Within this task space, the evolution of the dimensionality of the stimuli, the sampling regime, and the category boundary definition all allowed for more complex questions to be asked about the computations underlying decision confidence in categorical decision spaces and their relationship to the distance to the category bound.

Research Questions

The behaviour of metacognitive confidence in a category space like that of Little *et al.*, (2011) is not entirely clear, and directional predictions or hypotheses are difficult to specify. Thus, Study 4 stands as an exploratory experiment which is without confirmatory hypotheses or predictions. This allows for multiple possible patterns of confidence to be speculated based on broader research questions.

For example, some basic exploratory questions might include the following:

1. How does confidence behave in this space?
2. Is confidence graded across the stimulus space along and between the bounds as we would expect from a distance-to-bounds approach?
3. Is there more complex processing occurring that means confidence behaves differently in these spaces such that it cannot be described as relating to a distance from the bound in some way?

However, taking these ideas further, the structure of this experimental design lends itself to some more specific research questions that are linked directly to the task space.

To begin with, a first exploratory question could be whether decision confidence is higher or lower in the L-category than in the □-category. If confidence is defined based on a simple distance-to-bound model, then it would follow that confidence should, on average, be higher in the □-category because the points

are, on average, further from the bounds than in the L-category. However, if decision confidence is sensitive to the category structure, the average confidence might be higher in the L-category. This is because, in a rule-based version of this task, a decision-maker can extrapolate that once a trial stimulus is small enough the colour hue does not matter as it will fit into the L-category. Similarly, once it is clear that the stimulus is "red enough" it will also be an L-shaped category trial. By contrast in the □-category confidence might be lower because there is no single dimension classification which confirms category membership for stimuli from this category. Instead, these stimuli are classed based on the integration or serial processing of the two dimensions, a more drawn out and involved process which may influence the participant's subjective uncertainty - their confidence. Indeed, depending on the cognitive representation of the task held by the decision-maker, and the ability of the participants to extrapolate logical rules from the space the expected confidence reports may vary significantly. The patterns of confidence values across the space should allow for these predictions to be tested and these competing hypotheses to be addressed.

A second exploratory question involves the definition of "distance" in this type of space. That is to say, it is unclear whether confidence should depend on the distance to the nearest bound, the furthest bound, or a combination (e.g., a sum or an average) of those distances. This definition is important as it changes how we might expect certain patterns of confidence to play out across the stimulus space.

Some of these patterns, particularly within the □-category, would be rel-

atively unaffected by the definition of the distance metric. For example, it is expected that trials based on stimuli drawn from the grid box further from the boundaries (i.e., the top right-hand corner of the □-category) will have higher self-reported confidence as compared to those trials drawn from closer to the boundaries, with the lowest confidence emerging in trials drawn from the grid box that is on the bottom left of the □-category. No matter how distance is scaled this pattern seems plausible.

However, within the L-category things become more complex. This category highlights another key question for exploration – how the two dimensions are being used, integrated, disaggregated, or ignored by participants. For example, we could suggest that stimuli drawn from the bottom left corner of the total space, the small red fish, will yield the highest confidence of the L-category trials as the dimensions are aligned in being low on both axes. One might then expect that there will be a gradient of confidence leading to that grid box along the arms of the L-category such that the smaller and redder the stimuli become the more confident the decision-maker becomes. This would be true if the processing of decision confidence was based on the integration of the two dimensions relative to the most apparent L-category stimuli that are generated in the lower left corner.

Yet, if a serial self-terminating probing of the dimensions is being leveraged a contrary set of predictions could also be made which suggest that the exterior grid boxes on the outer edges of the arms of the L-category should yield high-confidence reports due to the possibility of those stimuli being defined based on

a single-stimulus dimension.

This question, and resulting tension in the contrary predictions, highlights the in-built asymmetry that exists across the two categories. In the □-category, the stimuli are all defined based on both the colour *and* the size dimension, with both dimensions having to be above the boundaries for the stimulus to hold membership to the □-category. By contrast, the L-category membership can be assessed based on the value of either the size *or* the colour. If participants, and their metacognitive confidence reports, are sensitive to this asymmetry an asymmetrical pattern of confidence can be anticipated on either side of the category boundary. This would distinguish the processing in this stimulus space from the simple predictions made by distance-to-bound models of confidence.

5.2.2 | Methods

The data set, data cleaning script, and analysis script for this study are housed on the (OSF). The experimental scripts and relevant materials for running the data collection can be found on (GitHub).

Participant Sample

Prior to exclusions, a total of 60 participants took part in this study of whom N = 58 had data saved successfully to the server. Participants had an average age of 34 years (range 19-57, standard deviation = 10.3). Self-identified gender splits were female (N=30), male (N=28) and non-binary (N=0). Handedness was reported as right (N=54), left (N=2), and ambidextrous (N=2). The average time taken by participants to complete the experiment was approximately 25 minutes (minimum 14 minutes, maximum 49 minutes, standard deviation 7.9 minutes).

Exclusion Criteria

Due to the exploratory nature of the experimental questions, and the novelty of the task being used, there were no predefined exclusion criteria for this study. All participants had to have completed the study fully and consented to the study prior to completion in order to be included in the final sample. N = 60 participants were collected on Prolific, of which N = 2 had incomplete data due to data saving errors on the server and were excluded leaving N = 58.

Once the data was collected a basic set of quality checks was applied to the data. Based on requiring average task accuracy to be above two standard deviations below the mean a post-hoc cut-off for inclusion for analysis was established at an average accuracy of 54% within participants. Of the 58 participants, 3 candidates for exclusion were identified. Of these N = 2 had an average task performance that fell below chance. All 3 participants were removed from the dataset prior to any analyses leaving a final total of N = 55 for analysis.

The local ethics committee approved the study protocol (CUREC: R68879/RE001). Participants were provided with an information sheet and gave their consent to participate before beginning the study. The study was conducted online using the Prolific platform for recruitment and participants were compensated at a rate of £10.00 per hour for their time.

Experimental Design

In this study, participants were asked to learn about the features of two species of fish, a "native" species that should be returned to the lake when caught and an "invasive" one that should be removed from the lake to protect the ecosystem. Thus, on each trial, they had to choose whether to catch or return the fish that appeared on the screen. To facilitate this learning, the experimental blocks were then divided into training and testing phases. During training, they saw fish (that varied in both size and colour) labelled as either native or invasive species; during testing, they saw more fish, but the species was not explicitly labelled. Therefore, they had to infer whether to catch or return the fish to the lake based

on the size and colour of the fish on screen.

In this study, the defining features of the categories of native or invasive fish, the size and colour, were held constant across all blocks and all participants. Therefore, for all participants, the native fish and invasive fish looked the same across all blocks. As before, participants received no feedback on their specific responses during the test phases. Instead, they received a cumulative score at the end of each block. this means that, similar to the work of Chapter 4, participants had no opportunity to learn confidence via reinforcement history of particular responses.

The stimulus categories used were discrete, with no perceptual overlap. The category boundary between the stimuli was implemented such that there was an asymmetry between the stimulus categories in terms of the need to integrate across the two dimensions to define category membership. In one case the values of both the size and colour of the fish were needed to confirm category membership, in the other the value on one dimension alone could be used to confirm category membership. This created a rich decision space for exploratory work on the nature of confidence reporting.

Therefore, this experiment was designed to explicitly test the learning of categories (the different fish species) when the observable properties of those categories (the colour and size of the fish) are drawn from discrete distributions with an in-built asymmetry in the category membership criteria. More details of each phase of the experiment will follow in the subsequent sections of this chapter.

Experimental Stimuli

As in Study 3, the stimuli used in Study 4 were images of fish that were built up from individual geometric shapes such as ovals and triangles that were systematically sampled from controlled category distributions. These stimuli were utilised as they had been successfully deployed as part of a single bounded categorisation task (Study 3) which involved categorising the stimuli based on the length of the fish, as defined by the length of the main oval of the fish body onscreen. Here, this dimension was manipulated again, in the same way. Therefore, the first dimension of the stimulus space in this study relates to the length of the oval body of the fish in pixel units. The second dimension of the stimuli was defined on the basis of the colour hue of that oval body shape. This hue value was drawn from an equiluminant colour gradient that spanned from pure blue to pure red. This colour hue was then applied to the whole body and tail of the fish excluding the eyes and outlines which were defined in black and white. In this way, the fish stimulus space in Study 4 was defined by two dimensions, total body length on the x-axis and body-colour hue on the y-axis.

For the purposes of stimulus sampling and later data analysis, this continuous two-dimensional space was divided into equal-sized grid boxes. The total space was divided into 9 grid boxes in a three-by-three formation. On the x-axis, each of the grid spaces was defined such that it spanned a space equivalent to 200 pixels which began with a start value of 200 pixels and increased up to a maximum value of 800 pixels. This means that small fish bodies fell between

200-400 pixels, the middle stimuli on the x-axis held lengths that varied between 400-600 pixels, and the grid boxes that depicted the largest fish spanned between 600-800 pixels.

On the y-axis, the colour gradient was determined in an equiluminant colour space spanning from pure red to pure blue. This colour dimension was defined in the red-green-blue (RGB) colour space. Importantly, the green value was held constant for all stimuli at 0 allowing the perceptual variation to focus on the red-blue gradient. Concurrently, the proportion of red and blue varied symmetrically. As each value (for R and B respectively) could range from 0-1 in the colour space with 1 being high, for a strong red hue an RGB of (1,0,0) was used. Similarly, a strong blue hue was defined with (0,0,1). These values acted as the bookmarks for the stimulus colour hue dimension. Thus, the gradient of the colour dimension was defined by assigning a value between 0-1 for R - the Red component – and defining the Blue component as (1-R). In this way, the colour dimension of the stimuli stretches along the 0-1 space along a single line from less to more red, or conversely more to less blue. On the y-axis, each of the grid spaces was defined such that it spanned a space equivalent to 1/3 of the hue dimension in the RGB space which began with a start value of 0 and increased up to a maximum value of 1. This means that red-ish fish fell between RGB(1, 0, 0) and RGB(0.66, 0, 0.33), the middle purple-ish fish varied between RGB(0.66, 0, 0.33) and RGB(0.33, 0, 0.66), and the grid boxes that depicted the blue-ish fish spanned between RGB(0.33, 0, 0.66) and RGB(0, 0, 1).

Using this approach, individual differences in colour perception across par-

ticipants' visual systems and monitors are not accounted for. I do not try to assess or account for these differences as the differences in confidence will be explored within individuals rather than between. Furthermore, any variability arising from differences in computer screen resolution or perceptual differences in the visual system is assumed to be relatively stable within the individual.

By combining these two dimensions, for each individual stimulus that was sampled, the x and y values were drawn such that a certain sized fish of a certain colour was built up from the values established on the x and y axes.

The Category Space

Using this three-by-three grid of possible stimuli, the category boundary was defined in this two-dimensional space such that there is a complex and discrete separation of the two categories. Using the same narrative from Study 3 in Chapter 4, the categories are labelled as being either native or invasive species of fish and participants are asked to respond appropriately based on which species of fish they believe to be presented on screen. As described before, the category bound was drawn at right angles such that it separates a box of four grids in the corner of the grid space from an L-shaped space that spans five grid boxes. This space is a replication of the stimulus curation methodology used by Little and Nosofsky to explore reaction time data as a proxy for category learning (Little et al., 2011), see Fig. 5.1 above. The structure is such that L-category decisions can be defined as an "OR" operation (if a stimulus is very red *or* very small) whereas □-category decisions are an "AND" (a stimulus has to be blue *and*

large) classification. In the current study which leverages this stimulus space, the location of the L-shaped category boundary means that the categories do not occupy an equal area in the stimulus space, see Fig.5.2 (not to scale): 5/9ths of the space is occupied by the L-category, while the □-category occupies the remaining 4/9ths. This inequality is accounted for in the trial numbers drawn from each space, but in this study, the space itself is not warped to equalise the mental representation of the categories.

Experimental Procedure

As in Study 3, in this study, the participants played an online arcade-style game where they took on the role of marine biologists catching fish in different lakes. As in Study 3, participants were asked to learn about the two kinds of species that could be found in a lake – native and invasive. They were then asked to learn to avoid returning invasive species to a lake where they would cause damage to the native species and ecosystem. The reward matrix matched Study 3, with the participant winning 3 coins for correctly returning a native fish, losing 3 coins for incorrectly returning an invasive fish and no coin changes on trials where native fish were incorrectly caught or correctly catching an invasive fish.

The participants were asked first to read some instructions and a simple attention check was carried out to ensure the participants knew the correct stimulus-response pairing for the game. This attention check was a four-trial mini-block where participants were shown a fish on screen with the label "native" or "invasive" present on screen alongside them. The participants had to respond

accurately on every trial to pass. In line with the instructions, returning the native fish to the water, and catching the invasive species for removal counted as accurate responses. If they failed, the response pairings and instructions were reiterated and participants had a second chance to complete the attention check. If they failed a second time they were removed from the study as per the Prolific guidelines.

Following the attention check, participants were introduced to and trained on the stimuli and stimulus categories. This training took the form of a passive learning phase which included no feedback as there was no interaction on the part of the participant. The stimuli were shown on screen for 1500 milliseconds and labelled as belonging to their respective category either native or invasive fish species. This phase was comprised of $N = 40$ trials, $N = 20$ for each category. This translates to 4 trials per grid box in the 5-box L-category space and 5 trials per grid box in the 4-box □-category area. This varied trial number of trials per grid box was adopted to account for the difference in the number of grid boxes per category based on the difference in the area represented by each category space. These training trials were produced based on random samples from a uniform distribution in each grid space which were then shuffled together into a randomised trial list. The order in which grids are sampled is, therefore, random across and within participants.

The participants then completed a test block in which they were asked to categorise the stimuli one by one with no trial-by-trial feedback. The test block had an equal number of trials per total category space. This was again set to

20 trials per category, divided across the grid spaces as per the training trials for a total of N = 40 test trials. After each decision, participants reported their decision confidence on the same sliding scale as before (see Study 2, 2b, and 3). The key dependent variable of interest was self-reported decision confidence defined on a continuous scale from "totally unsure" to "sure correct". The participants did not receive any trial-by-trial feedback on the accuracy of their decisions, instead, they received blockwise feedback on the cumulative number of coins earned throughout that block.

In total, the participants completed 4 rounds of training and testing with all parameter dimensions, category boundaries, and trial numbers for the categories held constant in each block. The values and associations were not altered between blocks as the space was considered to be a challenging space to learn for participants without attempting to counterbalance which type of species occupied which area of the stimulus space. This means that across the test space, each participant completed 160 trials with 160 associated confidence reports in the study.

5.2.3 | Results

Task Performance

Given the novelty of the task, and the difficulty of the category space, which requires the participant to not only integrate across multiple dimensions but also learn a complex category boundary, it was relevant to assess the learnability of the task. Thus, participant task performance was assessed and the details are reported here.

Participants appeared to be able to learn the task over time with most participants performing above chance on average. Across all participants, the average accuracy across all blocks was approximately 78%. The minimum accuracy was approx. 57%. After the exclusions had been applied, no participants performed below the chance level on average. Looking at the trends in accuracy across blocks, most participants improved across the four blocks, indicating an ongoing learning curve throughout the study as was expected in this more complex category space. See Fig. 5.3 for details.

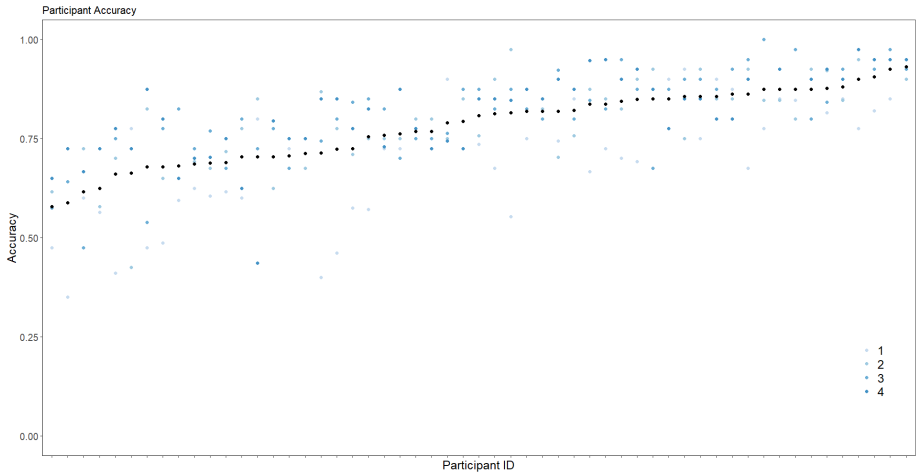


Figure 5.3: Blockwise and overall accuracy per participant in the 4 blocks of test trials. The legend indicates the per block (1-4) accuracy per participant. The black points indicate the overall average accuracy of each participant. The x-axis is ordered based on ascending overall accuracy within participants.

Confidence Analysis

Participants appeared to use the confidence measure in a sensible way. As before, the confidence scale was a sliding scale from 0-100 through which participants adjusted to their preferred expression of certainty on each trial. Foremost, using the data quality checks employed in studies 2b and 3, it was found that no participants used an unusually low number of points on the continuous scale, i.e., just the extremes of the scale. The minimum number of unique confidence values expressed by a participant was 13 with the average number of unique confidence reports sitting at approximately 51 across a possible 100 unique values. This suggests that confidence reports were being used in a nuanced and responsive manner by the participants in this dataset. This assertion is further supported qualitatively by observation of the smoothed averaged

confidence across the entire stimulus space which indicates a contour of lower confidence that roughly maps to the objectively imposed category boundary, see Fig. 5.4. This figure was created by first subdividing the total stimulus space into 36 equal grid boxes and calculating the average confidence within each individual within each of these grid boxes. The smoother was then applied across the total dataset to smooth between those 36 mini-grid boxes per participant and map out the whole space appropriately.

While confidence appears to generally track with the task structure in Fig. 5.4 some nuances are worth noting. Foremost, due to the variability in individual use of the confidence scale, by taking the total average the smoothing function actually underestimates the variation in the space. Secondly, while a sweeping L-shaped contour can be seen, there is clear uncertainty around the upright arm of the L-category as denoted by the wide blue contour in the upper left segment of the plot. This suggests that there is an asymmetry in the dimensions and how they are being attended to, or processed, by the participants. This would correspond to participants being more confident when categorising in the L-category when trials are based on colour rather than size. This corresponds to patterns of error and correct responses, with the highest error level corresponding to trials drawn from the upright arm of the L-category on average.

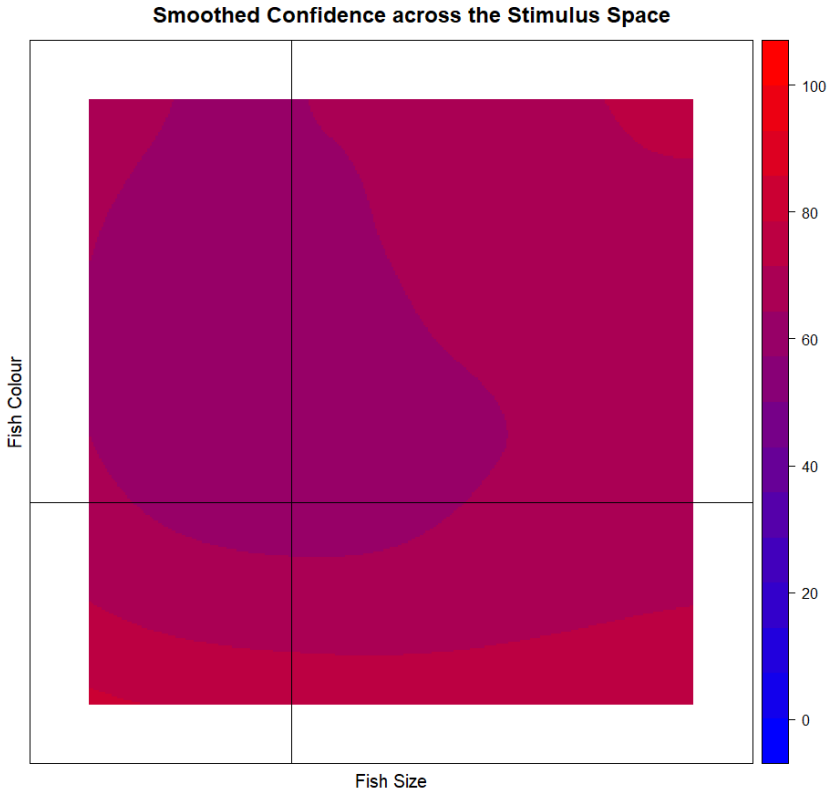


Figure 5.4: Average confidence for all participants when smoothed across the total stimulus space. The more blue the colour the lower the confidence, the more red the colour the higher the confidence. Note the sweeping L-shaped contour of the visualisation which echoes the objectively imposed decision boundary.

Confidence values also passed further data quality checks including demonstrating higher values on average, within participants, for correct (M = 70.8) rather than error trials (M = 57.6), $t(54) = 9.535$, $p < 0.001$. Similarly, confidence reports in correct versus error trials showed a difference in response times with confidence reports in correct trials being faster than error trials (2035 vs. 2180 ms respectively), $t(54) = -2.138$, $p = 0.037$. Furthermore, on average, confidence values varied across the space within individuals. The way in which

this variation occurred was different for different individuals. Some overarching patterns can be seen, however. For example, there were no differences in average confidence reports in trials where stimuli were correctly identified as being native (□-category) as compared with invasive (L-category), between the total correct □-category or L-category trials, $t(54)$ = -1.59, p = 0.1. There was also little numerical difference in the error rates between the two categories – of 8707 valid confidence report trials, 6796 were correct responses of which 3383 belonged to the L-category and 3413 to the □-category. Thus, while participants were marginally more accurate in the □-category this difference was not marked.

Yet, looking at the individual grid boxes when artificially imposed onto the stimulus space there is a significant variation across the different sections of the space. See Fig. 5.5 for average values within and spread of confidence across each grid box. Looking at these values allows for the assessment of specific and directional predictions. For example, within the □-category we can assess simple distance-to-bound effects by comparing confidence from the upper right (grid box [3,3]) as compared with the lower left (grid box [2,2]). Indeed, within the □-category (the top four right-hand grid boxes in Fig. 5.5), there is significantly higher confidence, on average, in the correct trials that were drawn from the upper right (grid box [3,3]) as compared with the lower left (grid box [2,2]) of the category space, $t(54)$ = 7.759, $p < 0.001$. This suggests that confidence is sensitive to a metric of distance in this category space. However, of equal interest is the behaviour of confidence across the category boundary, for example, grid

boxes [1,1] and [2,2] which stand as the representative "closest" grid boxes to the intersection of the bound in each category. A distance model might suggest that, depending on the metric of distance being used, the confidence difference should be negligible between these two grid boxes. However, the confidence in grid box [3,3] is significantly higher than the confidence in grid box [2,2], $t(54) = 5.37$, $p < 0.001$. The large difference in confidence in these two grid boxes suggests that confidence is not just based on a simple distance measure and instead might point to a different computation for confidence, or a difference in classification in grid box [1,1] as compared with grid box [2,2] that is being conveyed in the confidence reports.

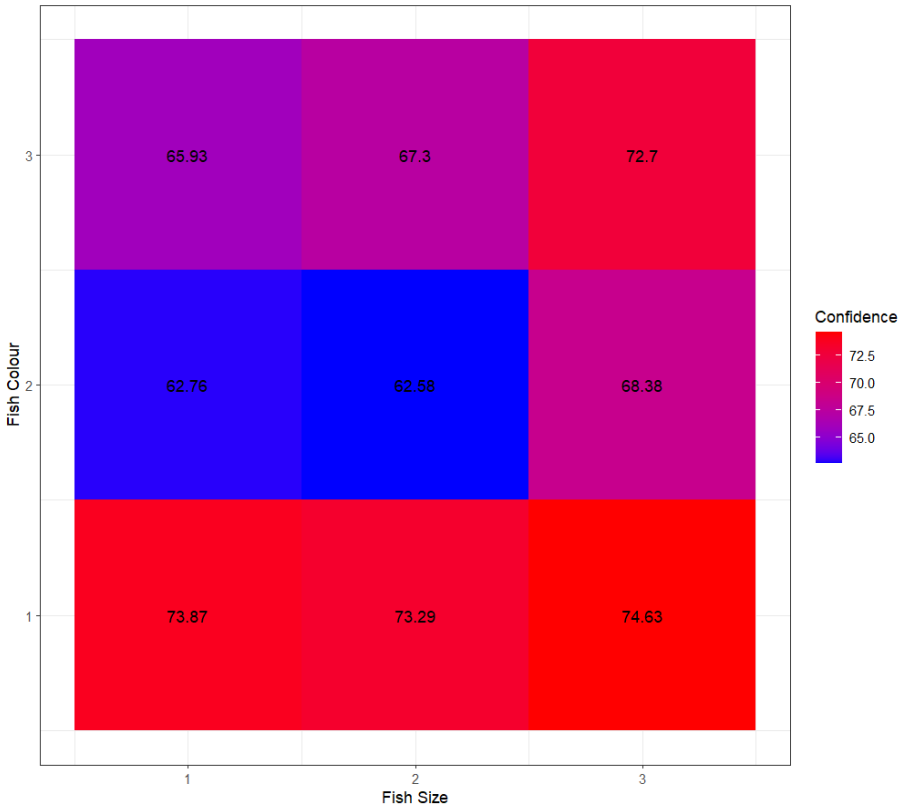


Figure 5.5: This figure shows the grand mean confidence across all participants when averaged for each of the 9 grid boxes used to sample the stimulus space. The more blue the colour the lower the confidence, the more red the colour the higher the confidence. The numbers correspond to the rounded average confidence value. Note that this is an unsmoothed version of Fig. 5.4.

In terms of the L-category, one suggested outcome was that the grid box [1,1] in the bottom left of the space would have the highest confidence. While this prediction did not strictly emerge, numerically the average confidence reports from this subsection of the L-category stimulus space were higher than 3/4 of the other grid boxes in the L-category with a grand average confidence value of 73.87. The only exception was the average confidence reported for the bottom

right grid box [3,1] of the L-category which came to 74.68. Attempts to compare a potential gradient along the two arms of the L-category by collapsing the vertical and horizontal arms together would obscure the clear numerical difference between each arm - with higher confidence reported in stimuli that could be delineated based on their colour alone ("red fish") as compared with those defined by their size dimension ("small fish"), see Fig. 5.5. This echoes the ideas discussed as relevant to Fig 5.4 regarding the asymmetry in the weighting or ease of use of the two category-relevant dimensions. This is reinforced by the fact that grid [1,1] represents significantly higher confidence values than grid [1,3] ($t(53) = 4.33$, $p < 0.001$ – the degrees of freedom are lower by 1 as one participant had no correct trials in grid box [1,3] and so were excluded from this analysis) but not as compared with grid [3,1] ($t(54) = -0.76$, $p = 0.45$).

Based on these initial comparisons, a few key findings can be summarised. Foremost, on average there were no overall confidence differences between the L-category and the □-category. However, looking at these spaces more closely specific differences can be seen. The gradient of confidence found in the □-category suggests that confidence behaves in line with a distance-to-bound model in that space. By comparison, there is a less clear structure in the L-category with dimension differences suggesting a single-dimension classification approach might best describe the behaviour here, or that one dimension is simply more difficult to parse or less salient than the other. However, looking at the broader variability across and within participants in terms of their confidence reports in this space, it became clear that a more nuanced approach to understanding and

mapping out confidence in this new two-dimensional space would be required.

Model-based Confidence Analysis

Therefore, to better understand how confidence was being computed by the participants in the space a systematic model-based analysis approach was used. The geometry of the space was considered further and it became clear that the "distance" to the category bound itself had to be re-conceptualised. As compared with the one-dimension stimulus spaces where the distance to a boundary is established when a boundary is set, what constitutes the distance to the category boundary in this two-dimensional space is less obvious. It is therefore unclear if the predictions made for this study were partially unsubstantiated due to misunderstandings of how confidence might be computed across the space, or if different individuals might be construing the space itself in different ways to aid and/or simplify their task performance. Therefore, to better explore the space, to understand how confidence might be computed in this space and to track better how participants were mapping out the space a number of unique models of confidence and distance in this two-dimensional space were created.

The first model that was established was the minimum distance model (minModel). This model estimated confidence based on a scaled distance from the nearest boundary value – estimating the minimum distance from either boundary in the space. As every given stimulus value is defined by an X and Y coordinate – the minimum distance model would check how far the Y value was from the horizontal boundary value and how far the X value was from the ver-

tical boundary value. The shorter distance of these two comparisons was then used to approximate confidence. The maximum distance model was a complementary approximation to this, it looked for the furthest distance to either of the boundary values in the space, standing as the direct opposition to the minimum model (maxModel). Finally, a third model looked at the summed values for both boundary-relevant distances - adding up the distance of the y value from the horizontal boundary and the distance of the x value from the vertical boundary (sumModel). These models were collectively named the "pure" models to indicate that they were based purely on a single distance approximation across both categories and all areas of the stimulus space. The "distances" calculated from these models in the stimulus space were then scaled between 0-1 to represent a proxy for confidence.

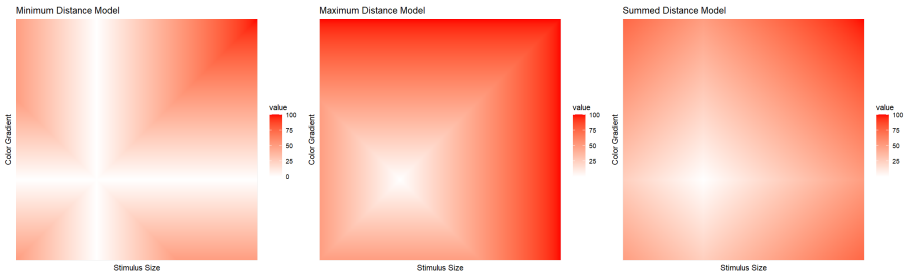


Figure 5.6: The three pure models are shown. Left to right they display as follows: the minimum, maximum, and summed distance models. Across all points of the grid space, and all possible combinations of dimension values, confidence was computed based on these different approximations of "distance". The confidence level at each point in space is illustrated based on the strength of colour from light to dark red with darker hues indicating higher confidence. Note the low confidence contours in the space as predicted by these models.

Following these three models, two further classes of models were also established. These model families varied from the "pure" models in that they had different approximations of "distance" to the bound in each category - the L-category and the □-category. These models hoped to begin to capture the idea that the processing of the stimuli for classification might be happening in different ways on either side of the boundary. This approach is data-driven, as it maps to the immediate patterns identified in the experimental data, but is also based on the findings of Little *et al* (2011) who found that in continuous dimensional processing a mixture of serial and parallel processes best described the operation of this kind of categorisation.

The first of these model sets is dubbed the "asymmetrical" models class. In this class of models, the distance-based confidence values in the □-category were defined either using the minimum, maximum or summed distance – as above in the "pure" models. In the L-category the calculation captured distance based on the closest of the two bounds, as would match the "*or*" logic of the space that was previously discussed. The formal definition of the "distance" in the L-category is then defined as follows: when the value of the x-axis dimension was greater than the value at which the vertical boundary cut the x-axis, only the Y dimension was considered to compute distance. Similarly, when the value on the y-axis was greater than the value at which the horizontal boundary cut the y-axis, only the X dimension was considered. This essentially meant that the min, max, and summed distance in 4/5 of the grid areas of the L-category were defined by a single dimension and so their values were equitable in each case (minimum, maximum, or summed distances were all equitable). In the case where the L-category sits diagonal to the intersection of the two decision boundaries, the distance is defined as the hypotenuse of the right angle triangle made from the intersection point down to the trial stimulus. These models are therefore asymmetrical in that confidence in the L-category is stable across models and does not describe distance in the same way as the □-category. The □-category can then vary between using the minimum, maximum, or summed distance values to proxy for confidence. In this way, a total of three models of the asymmetrical class were defined and are described as minAsym, maxAsym, and sumAsym.

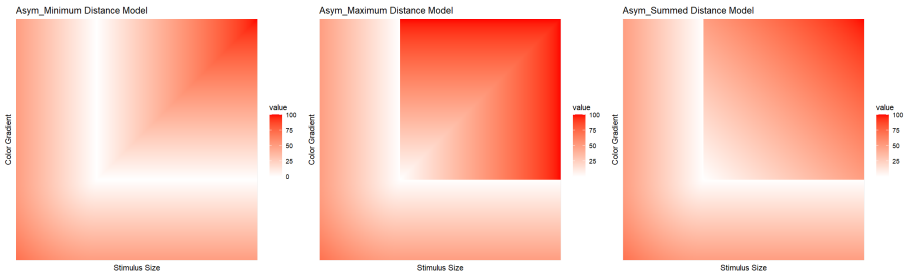


Figure 5.7: The three asymmetrical models are shown. Left to right they display as follows: the minimum, maximum, and summed distance models. The confidence level at each point in space is illustrated based on the strength of colour from light to dark red with darker hues indicating higher confidence. Note that the L-category values are the same across all three models as the approximation of distance is held constant for the L-category in all models.

Finally a third class of models known as the "scale" models were created. These models were scalar in the sense that they capture the probability that a stimulus is drawn from either the L-category or the □-category (1- P(L-category)) as scaled between 0-1 based on the stimulus values. Intuitively this attempts to capture the "*or*" nature of the categorisation task in the L-category where the position relative to the boundaries is less important than how clearly small or red the fish stimulus was. As with the asymmetrical models, this class of models used the minimum, maximum, or summed distance to the bounds to define confidence for stimuli from the □-category space. However, the simulated L-category confidence values were calculated differently as compared with those in the asymmetrical models. Instead of computing confidence based on one of

the bounds as before, in the "scale" models the values in the L-category space were converted into the stimulus space and scaled accordingly. This essentially means that the L-category notion of distance was calculated relative to, or as a function of, the stimulus values themselves not accounting for the decision bounds themselves: confidence is greater the more red or the smaller is the stimulus. This was achieved by converting the stimulus values into a probability space and calculating the probability of a stimulus of a certain x and y value being from the L-category within the L-category space. Combining these L-category values with either the minimum, maximum, or summed distance □-category values, led to the creation of the minScale, maxScale, and sumScale models.

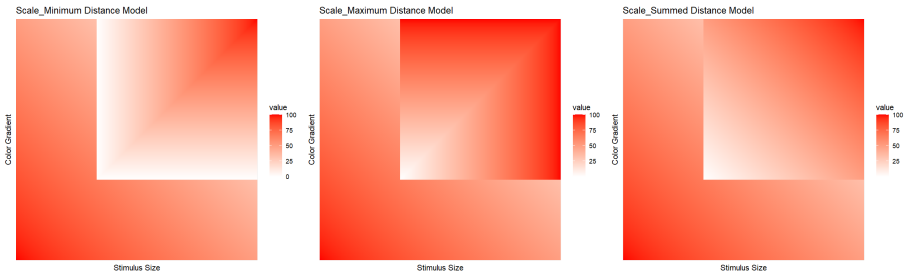


Figure 5.8: The three asymmetrical models are shown. Left to right they display as follows: the minimum, maximum, and summed distance models. The confidence level at each point in space is illustrated based on the strength of colour from light to dark red with darker hues indicating higher confidence. Note that the L-category values are mirrored across all three models as the approximation of distance is held constant for the L-category in all models - however, this approximation is different from that of the L-category values in the asymmetrical models.

In total, this created a 3x3 matrix of models that were either minimum, maximum, or summed in terms of their distance and either "pure", "asymmetrical", or "scale" in terms of their class. For every point in the stimulus space, the values for each of the nine models were simulated.

For each trial, a simulated value of confidence under each model was compared with the actual value of the reported confidence on that trial. The correlations between the simulated model values and the real data were used as a proxy for the goodness of model fit. Second-order statistics were then conducted to complete a basic model comparison procedure. The R-values of the correlations of the nine models were extracted and assessed in a 3*3 ANOVA

examining the type of model (pure, asymmetrical, or scaled) and the model distance (minimum, maximum, or sum). This analysis indicated that both the main effect of model distance [$F(1.04, 56.14) = 5.23$, $p = 0.02$] and the interaction between model distance and model type [$F(1.17, 63.36) = 5.59$, $p = 0.01$] had significant effects on the strength of the correlations emerging from these analyses. From Fig. 5.9, which plots average correlation coefficients across models, a few key findings can be drawn out. Foremost, the performance of the "pure" family of models is weaker than the other variants that model asymmetrical processing across the category boundary. This can be seen in the lower correlations between the model-simulated and true confidence reports of the pure version of every distance model relative to those in the other two model families.

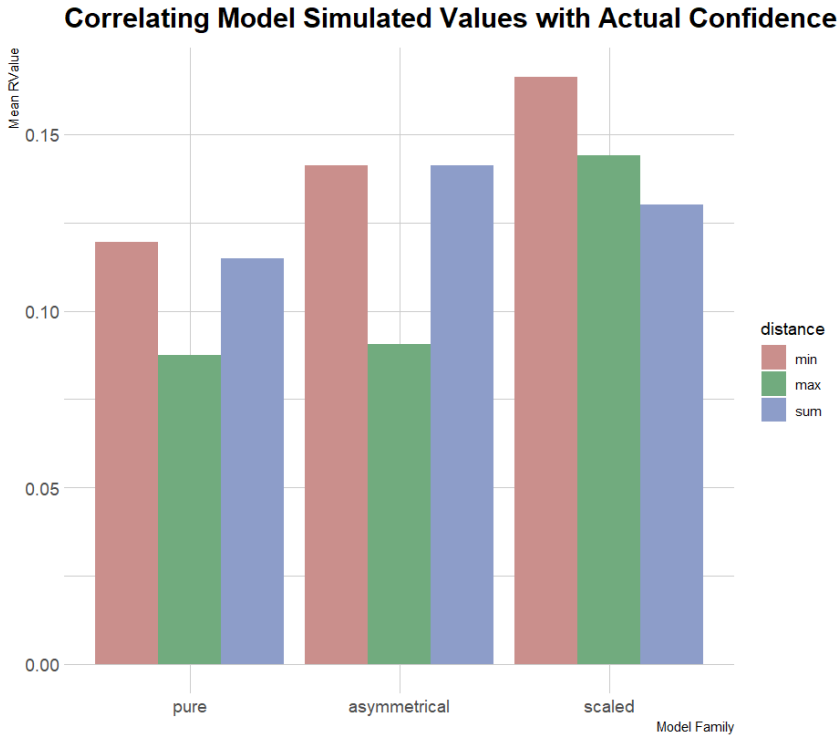


Figure 5.9: The mean R-values for each model's predictions when correlated with participants' confidence reports (as applied to the confidence data for Study 4 using objectively defined boundaries). The bars are clustered based on the model type, while the colours indicate the approximation of distance captured by the individual model.

However, distinguishing the best individual model beyond that is difficult. While on average the "scale" family of models seem to do best the results are far from clear. In order to distinguish these models further, another look at the participant behavioural data is required. Indeed, it is important to note that "distance" in these spaces was originally defined based on the distance from the trial stimulus value and the objectively created category boundaries. In reality,

this boundary may be noisy and may depend on the individual's mapping of the stimulus space.

Therefore, using the response data from the participants, a subjective "response boundary" was established within each participant. This was calculated by computing a grid search of the entire stimulus space to explore what values on the X and Y axis could be estimated for the category boundary such that the participant's error rate would be minimised. That is to say, if we assume the participant was using a logical stimulus-response mapping, where did each participant think the category boundary was in this task based on how they responded across the space?

In the search space, the response boundary values were constrained to prevent boundary placement which would mean that no data points were located in any given quadrant of the space. This normalised the response boundaries and prevented extreme values from being chosen based purely on the fact that no data were seen at a certain point. Such a sampling pattern could occur based on the uniform random sample in the grid boxes and if this was not restricted within the grid search the error count could be minimised by placing the boundary there. A minimum of three data points was required to exist in each quadrant, within each individual participant's dataset, for the values of the dimension bounds to be considered valid – an arbitrary but conservative cut-off point.

These response boundaries were required to maintain the right-angled intersection of the category dimensions but the placement values of the two bound-

aries were free to vary to vary. By re-fitting the nine "distance" models to these boundaries, I aimed to reduce the noise of the fit of confidence by assuming that participants used their confidence judgements and scaled them relative to the boundaries they learned and implemented rather than necessarily to the true category boundary of which they may only have a noisy approximation.

Fig. 5.10 shows the individual fitted boundaries of the participants. In this figure, the colour of each point relates to the confidence reported in that trial ranging from low to high (blue to red). The shape of the points represents the response given by the participants, whether or not that was correct. The fitted subjective boundaries are illustrated by the intersecting black lines on each panel. Foremost, in Fig. 5.10, it is useful to note the individual variability in the fit of the response boundaries across participants. While some participants respond approximately in line with the objective reality of the category bounds (e.g. Participant ID (PID) 1) others appear to struggle to respond effectively to the structure of the space (e.g. PID 50) and others again seem to focus on one dimension exclusively (e.g. PID 23).

Once the subjective response boundaries were established for each participant, the model-simulated values of confidence were re-simulated within each participant relative to those boundary values. A small number of trials (N = 71) were removed at this stage as, due to the extreme values of the response boundaries, the re-scaling of confidence into the 0-100 space in a uniform way was not possible when those trials were included.

Based on these re-simulated model values calculated using the subjective

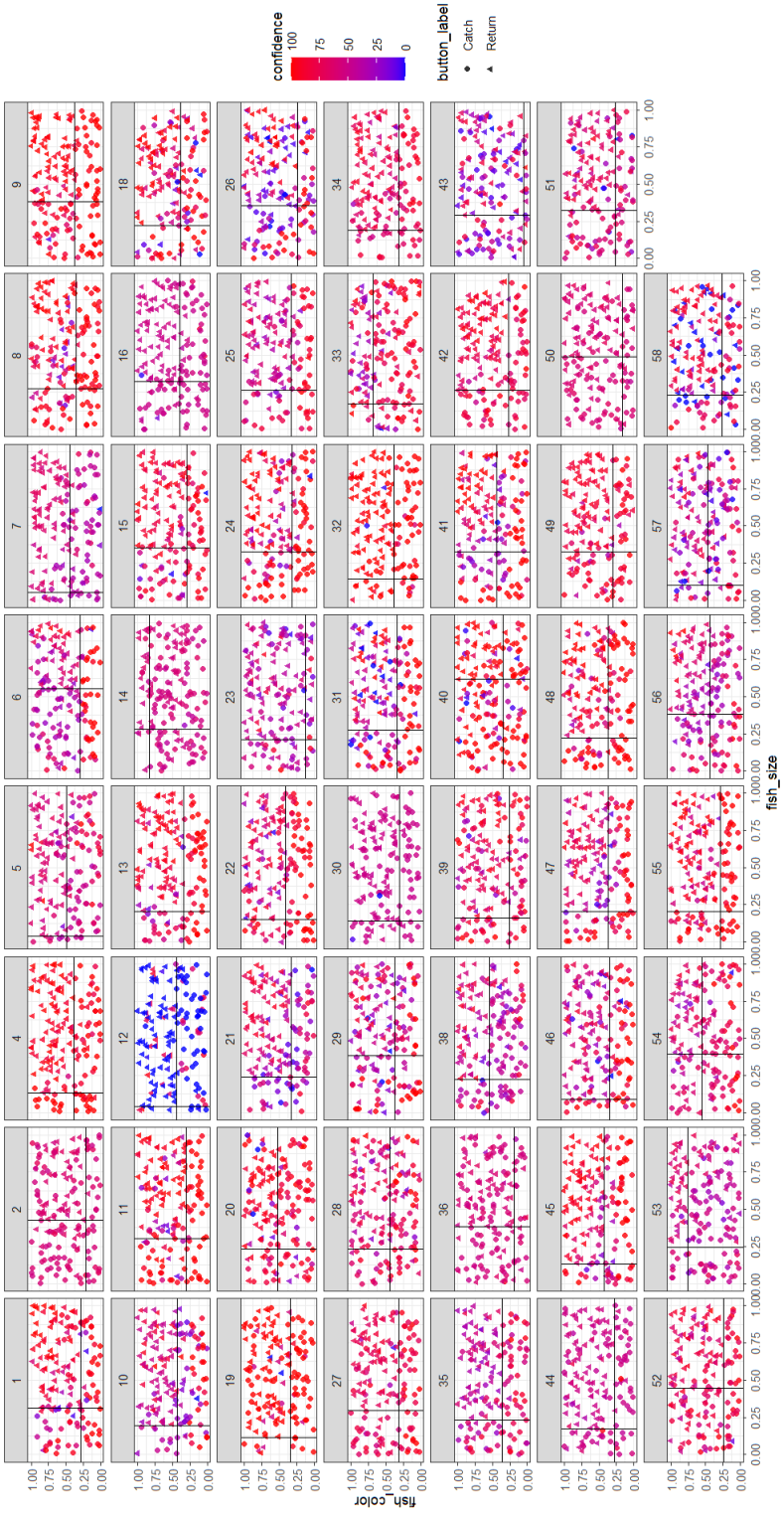


Figure 5.10: Each panel represents the response data, coloured by confidence in that response, per participant. The shape of each point represents which response the participants gave – i.e., what category they thought the trial belonged to. The fitted subjective boundaries are illustrated by the intersecting black lines on each panel. Note the variation in the category representations being used by the participants as inferred from these response-based boundaries.

response boundaries, the correlations with the true confidence values were re-calculated. As before an ANOVA was conducted on the R-values of these correlations acting as a second-order statistical model comparison procedure. The ANOVA indicated that, as with the analysis conducted on the objective category boundary, the interaction of model type and model distance had a significant effect on the strength of the correlations [$F(2.1, 113.2) = 11.97$, $p < 0.001$]. This suggests that not only are the models differentiable but that their difference in fit to the data can be distinguished best by looking at both the model distance and type. However, unlike in the objective boundary analysis, the main effect of distance was not significant in the subjective boundary analysis [$F(1.51, 81.29) = 3.15$, $p = 0.06$]. Plotting out the mean R-values across all participants for the models gives some indication both of the variations in model fit and of the comparative strength of particular models, see Fig. 5.11.

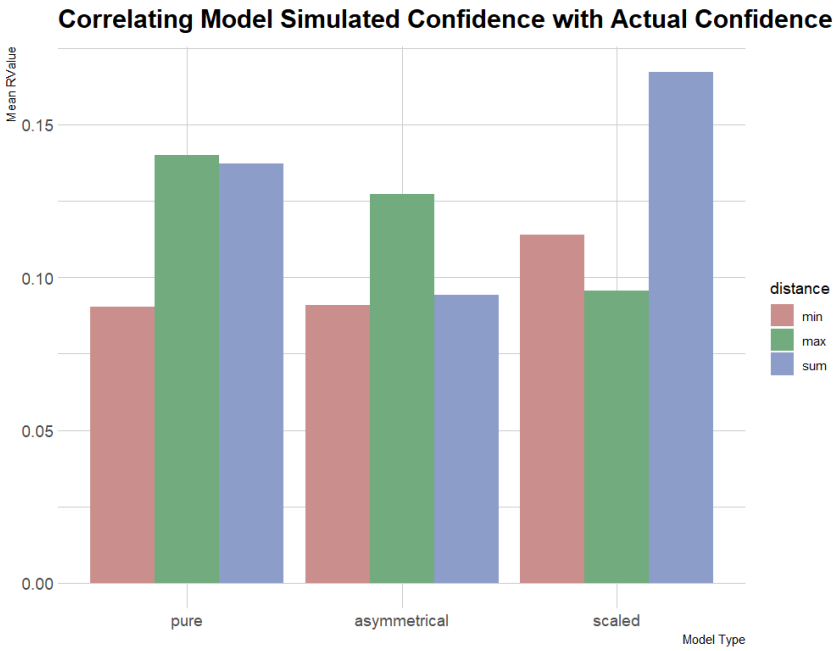


Figure 5.11: The mean R-values for each model are plotted here, as applied to the confidence data for Study 4 using subjectively defined boundaries. The bars are clustered based on the model type, while the colours indicate the approximation of distance captured by the individual model.

Some broad observations can be made from this visualisation. For example, comparatively speaking, the minimum distance model performed the worst of all distance approximations across all three model types. Similarly, in general, the asymmetrical model family did not perform well across the board as compared with the other model types. Overall, the scale-summed model showed the strongest correlation with the actual confidence reports. this model implies that confidence is functioning asymmetrically around the boundary and is sensitive both to a combined distance metric in the □-category and a scaled

interpretation of the stimulus strength expressed in the probability space in the L-category. However, the R-values from the pure summed and pure maximum distance models were also reasonable. These results differ from those of the objective analysis seen in Fig. 5.9 where the pure model family was the poorest fit the data and the scaled-max model was the best fitting model. It is unclear if the improved fit of the pure model family, which is based purely on classic distance-to-bound confidence computations in both categories, is because the distance metrics are now calculated relative to the subjective bounds used by the participants or if the correlations are just picking up noise.

Based on these results is it difficult to make a conclusive statement about the behaviour of confidence in this space. In order to better understand these data I, therefore, conducted some further exploratory model-free analyses. To understand if the patterns derived in the model-based analyses are based on meaningful signals a pre-registered replication is also run in the form of Study 5 below.

Model-free Confidence Analysis

To complement these model-based analyses, I also conducted a more nuanced look at patterns of confidence to better explore how confidence functions on either side of the category boundaries of both dimensions.

To achieve this, the grid boxes on either side of the boundary were divided in half, creating a set of rectangular stimulus spaces that sat nearer or further from the boundary on either side of the bounds. This was conducted to allow

for a closer look at the effect of distance on confidence in mirrored spaces on either side of the boundary in each category. It allowed some of the queries regarding the symmetry or asymmetry of confidence across each category to be investigated in more detail. See Fig. 5.12 for a clearer view of what these subdivisions looked like in practice, each grid box next to either the vertical or horizontal boundary line was halved to create an "inner" and "outer" rectangle, denoted by the light and darker colours respectively.

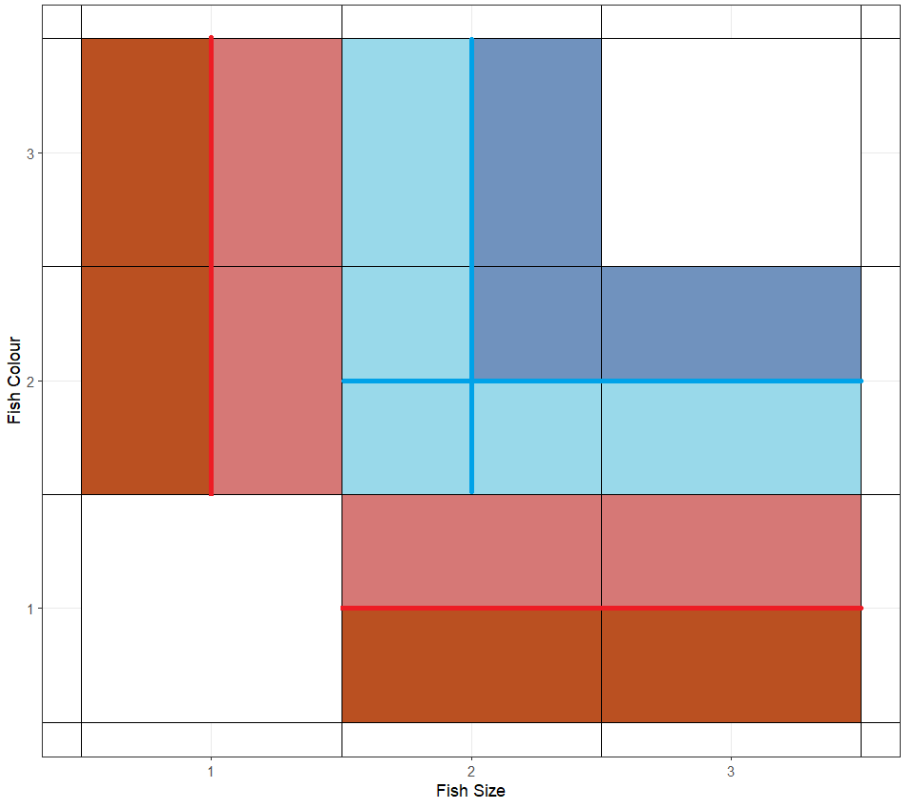


Figure 5.12: The grid areas next to the boundaries on each dimension were subdivided into rectangular spaces near (light colours) and far (dark colours) both in the L-category (red tones) and the □-category (blue tones). Note that the grid box [2,2] in the L-category was double sampled by necessity to compare both arms of the boundary.

Within each of these rectangles, within each individual, the average confidence in correct trials was calculated. These values were analysed using a three-way ANOVA examining the distance (inner or outer sub-grid location, either closer to or further form the bound) of the rectangle relative to the bound, the category membership (L or □-category), and the arm of the boundary (vertical or horizontal) within participants. One participant was dropped from this analysis due to their having no correct trials in the grid box [1,3], leaving N = 54 for analysis.

The ANOVA indicated that the main effects of distance [$F(1, 53) = 50.15$, $p < 0.001$], arm [$F(1, 53) = 48.02$, $p < 0.001$], and category [$F(1, 53) = 4.25$, $p = 0.044$] were all significant at the 0.05 alpha level. The two-way interactions between distance and category [$F(1, 53) = 12.44$, $p < 0.001$], and arm and category [$F(1, 53) = 52.91$, $p < 0.001$] were also significant. The three-way interaction [$F(1, 53) = 2.99$, $p = 0.09$] and the two-way interaction between distance and arm was not significant[$F(1, 53) = 2.93$, $p = 0.093$]. These results suggest that confidence is sensitive to the structure of the categories and the position of the bounds, as well as the different dimensions used to define them. A key finding is that confidence changes more quickly as a function of distance from bound in the L-category than in the □-category, as shown by the two-way interaction between distance and category. While the different dimensions (as indicated by the effect on "arm") seemed to have different effects on confidence in each category, as shown by the category-arm interaction – this effect was not evident in the three-way interaction which included distance. That is the

different dimensions did not influence how confidence changed as a function of distance in the two categories.

A feature of the full three-way ANOVA is that some data points contribute to more than one cell of the factorial design – specifically, data points in the lower left quadrant of [2,2] appear in the "near" distance bins of both arms of the □-category. To ensure the two-way interaction between the distance and category – i.e., the effect that suggests that distance is operating differently in each category – was not the product of this double-dipping of data in the grid box [2,2], two separate 2-way ANOVAs were conducted on each arm of the category bound. The interaction between distance and category was significant in the analysis of the vertical arm of the data [$F(1, 53) = 12.65$, $p < 0.001$]) but not in that of the horizontal arm [$F(1, 54) = 3.18$, $p = 0.08$] – suggesting that the effect was not driven purely by the double sampling that occurred in the three-way ANOVA. This variability in the interaction effect also suggested that there could be a difference in the processing of the two dimensions such that distance was interacting with the category membership in a different way on each dimension bound. This result is highly relevant and pertinent to the predictions of the distance-to-bound model which would not account for these variations. Indeed, the distance-to-bound model cannot easily capture or account the variability of confidence gradients on either side of the category bound that are evidenced by these interaction terms.

Qualitative Data Exploration

A final analysis related participants' classification performance to the free-response text that participants reported on their task strategies at the end of the experiment. Indeed, based on the quantitative analyses there seems to be a difference in salience of the two dimensions in the L-category. So as a final look at the data, I explored whether this difference was apparent in participants' reports of their strategies. These responses were prompted by the following prompt: *"Please describe what strategy you were using to reach your decision about whether or not to catch a fish."*. A response was not required but all N = 55 participants included in the analysis chose to complete the text box with varying levels of detail.

Interestingly, of the N=49 valid responses, N = 23 broadly described using both dimensions to approach the classification of both categories e.g. *"I tried to remember the combination of colour and shape that indicated which species it was from the learning phase" (Participant 21)* or *"Trying to remember the combinations of colours and sizes that were native/invasive." (Participant 9).* Interestingly, the dimension of colour was mentioned first in 19 of these 23 responses suggesting its relative salience for these participants.

The dominance of the colour dimension was also clear when only one dimension was referenced by the participants. When both categories were referred to by a single dimension it was always in relation to colour (N = 3) for example for Participant 28 their strategy involved trying to see *"which colour fish were*

invasive and which were native". When only one category was mentioned in one-dimensional classifications (N = 4), the majority (3/4) were based on colour e.g. *"I knew that red ones were invasive and the others I guessed" (Participant 25).*

Looking at more specific approaches, N = 3 participants were classified as attending to the individual dimensions in a serial manner in both categories e.g. *"I initially went by colours and then shapes." (Participant 5).* Some participants had a clear way to articulate the "OR" structure that could be used in the L-category (the invasive species of fish) e.g. *"Trying to find common thin[g]s between fish of the same category, i.e. when red or redish -> invasive -> catch or when round -> invasive -> catch. I decided to only do this for the invasive fish, hence all that did not fall into one of my categories I classified as native" (Participant 47)* with N = 9 participants making this type of connection explicitly for the L-category (with varying degrees of specificity about the relevant dimensions for the □-category. Some participants, N = 3, identified only colour as being explicitly relevant in the classification of the L-category and both dimensions being needed for the □-category *"I looked at how many fishes were native and had similar size. Then I tried to see which colors were present the most in this size. I also recognized the color of the fish, that were invasive most of the time and remembered this aswell[sic]. But here I didn't look at the size, only the color." (Participant 48).*

Looking at these data holistically a few themes emerge. First, most participants were aware of and used both dimensions to complete the task. Secondly, some participants identified a difference between the approaches they used for

the two categories as relevant to the dimensions. Finally, a clear difference in the dimensions in terms of salience was identified both in the number of times colour was mentioned as relative to size, but also in the amount it was used as a single dimension classifier for the L-category.

5.2.4 | Discussion

This study aimed to investigate the behaviour of confidence, and its relationship to the distance-to-bound models of confidence, in a multi-dimensional category space with a logical rule structure apparent in one of the two category spaces. This exploratory study, and its findings, have a complex and rich set of results. From the more basic confidence analyses to the modelling of the space relative to the subjective response-based bounds of the participants, several areas of interest were identified.

The asymmetry between the L-category and the □-category, that the L-category stimuli can be classified using a logical OR rule whereas the □-category requires both dimensions to be evaluated for classification (an AND rule), is a key feature of this task.

This asymmetry may be driving the different operations of confidence in relation to distance in each category. This difference was seen both in the basic analyses of confidence across the space and more involved comparisons of confidence as related to the boundaries. For example, the difference in the suggested operation of confidence based on the comparisons between grid boxes [1,1] and [2,2] and grid boxes [2,2] and [3,3] highlight this. The difference in confidence within the □-category comparison (grid boxes [2,2] and [3,3]) suggested that confidence was responding relative to the distance to the bound. However, the difference in confidence over the decision bound (at grid boxes [1,1] and [2,2]) suggested that this gradient function for confidence was not equitable across

the bound in the two categories. To complement these results, this asymmetry also appears in the participant confidence reports via the differential effect of distance in each category along both dimensions of the categories in the 3*3 ANOVA of the sub-grid box spaces. However, the two dimensions themselves also appear to be weighted differently by participants in terms of either their salience or ease of processing with higher confidence in stimuli defined by colour. This is apparent in both the qualitative and quantitative data. Furthermore, despite the relative ambiguity of the model-based correlation analyses a model which attempted to capture different functions for confidence in each category (the scale family of models) appeared to best fit the data in both the objective and subjective boundary analyses.

In order to assess the robustness of the key results that supported these ideas, a confirmatory replication was designed and preregistered with specific and directional predictions in mind based on these results. This further served as an opportunity to check the results shown in the model-based analyses.

5.3 | Study 5

5.3.1 | Introduction

Based on the interesting patterns of results of exploratory Study 4, a confirmatory replication was conducted to understand the replicability of the patterns found and their relation to conceptualisations of decision confidence. The replication was pre-registered on AsPredicted (https://aspredicted.org/5fz7z.pdf).

As before, this study addresses the broad question of whether decision-makers utilise distance to the decision bound when reporting subjective decision confidence in a two-dimensional discrete category space with an asymmetry in the logical structure of the categories. However, based on the exploratory results of Study 4 specific and directional predictions are made of the data in Study 5. Therefore, while the methods replicate those of Study 4, the analyses act as a more directed approach to testing specific and data-driven predictions. The methods and analysis approach replicated those of Study 4.

Experimental Predictions

Basic predictions:

1) Within the □-category it is predicted that trials based on stimuli drawn from the grid box furthest from the boundary (i.e. the top right-hand corner, grid box [3,3]) will have higher self-reported confidence as compared to those trials drawn from closer to the boundary, with the lowest confidence emerging

in trials drawn from the grid box that is on the bottom left of the □-category (i.e. the bottom left-hand corner, grid box [2,2]).

2) Within the L-category stimuli drawn from the bottom left corner will have the highest confidence of the L-category trials as the dimensions are aligned in being low on both axes.

Key Predictions:

The computation of confidence will vary in the O and L-category with the latter engaging a rule-based model that weights one dimension over and above the other in different parts of the category space: e.g. once a trial is small enough on one dimension the other dimension value does not matter as it must belong to the L-category. This difference will emerge in the data in several ways:

3) Mean confidence in the lower left grid box [1,1] will be significantly higher than the adjacent grid box in the □-category (grid box [2,2]).

4) There will be an interaction between the category (L and □) and the distance from the boundary based on the arm of the L-category being examined. This will show when the confidence in trials near to and far from the boundary in each arm of the L-category are compared with confidence trials mirrored in the □-category space accounting for the participant, category, distance, and arm of the category space involved.

Stimulus-specific predictions:

Based on the data patterns found in the previous exploratory study and due to the structure of the categories across two different perceptual dimensions it was predicted that:

5) Participants will be more confident in trials of the L category that are defined by colour than size as colour is a more salient category cue.

5.3.2 | Methods

The data set, data cleaning script, pre-registration, and analysis script for this study are housed on the (OSF). The experimental scripts and relevant materials for running the data collection can be found on (GitHub).

Participant Sample

A total of 56 participants were analysed as part of this study. These participants had an average age of 28 years (range 20-52, standard deviation = 7.2). Self-identified gender splits were female (N=28), male (N=25), non-binary (N=2), and other (N=1). Handedness was reported as right (N=51), left (N=4), and ambidextrous (N=1). The average time taken by participants to complete the experiment was approximately 25 minutes (minimum 13 minutes, maximum 56 minutes, standard deviation 8.8 minutes).

Exclusion Criteria

All participants had to have completed the study fully and consented to the study prior to completion in order to be included in the final sample. Participants were collected in batches of 10 until at least 55 participants passed the pre-registered exclusion criteria. In total 6 batches of 10 participants (N=60) were collected of which N=56 met the inclusion criteria. Prior to the application of the exclusion criteria, a built-in attention check was leveraged during the study to screen for inattentiveness. Any participant who failed that attention

check was rejected from the study and a new participant was recruited in their place.

Participants who passed the attention check were subject to the pre-registered exclusion criteria. Participants were excluded if they scored an average below 55% accuracy on the test trials across all blocks. Furthermore, participants were excluded if they used confidence reports in a way that signals non-compliance, defined as using fewer than 3 unique confidence values across all trials. In this study, four participants met the first exclusion criteria by having an average accuracy below 55%. Three of those had an accuracy level between 55-52%, and one had an accuracy of just 43%. All participants were compliant in their use of the confidence reporting scale. This left a total of N=56 participants for analysis.

The local ethics committee approved the study protocol (CUREC: R68879 /RE001). Participants were provided with an information sheet and gave their consent to participate before beginning the study. The study was conducted online using the Prolific platform for recruitment and participants were compensated at a rate of £10.00 per hour for their time.

Design Overview

As in Study 4, in this study participants were once again asked to learn about the features of two species of fish, a "native" species that should be returned to the lake when caught and an "invasive" one that should be removed from the lake to protect the ecosystem. Over four blocks of training and testing, they

had to infer whether to catch or return the fish to the lake based on its size and colour. The category space and stimuli were replicated directly from Study 4 and the participant experience was identical. Once again, this experiment was designed to explicitly test the learning of categories when the observable properties of those categories are drawn from discrete distributions with an inbuilt asymmetry in the category membership criteria.

5.3.3 | Results

Basic Analyses

Regarding task performance, across all participants, the average accuracy across all blocks was approximately 78%. The minimum accuracy was approximately 57%. In terms of confidence reporting, confidence values were higher on average, within participants, for correct (M = 71.4) rather than error trials (M = 56.1), $t(55) = 10.007$, $p < 0.001$. Similarly, confidence reports in correct versus error trials showed a difference in response times with confidence reports in correct trials (M = 2153 ms) being faster than error trials (M = 2244 ms), $t(55) =$ -2.129, $p = 0.038$.

Preregistered Analyses

Grand-averaged confidence reports across the stimulus space are shown in Fig. 5.13. The first preregistered hypothesis pertained to a difference in confidence between grid box [2,2] and grid box [3,3]. It was predicted that, in line with a distance-to-bound theory of confidence, confidence would be higher in grid box [3,3] than [2,2]. A t-test indicated that this direction difference was statistically significant, $t(55) = 6.07$, $p < 0.001$.

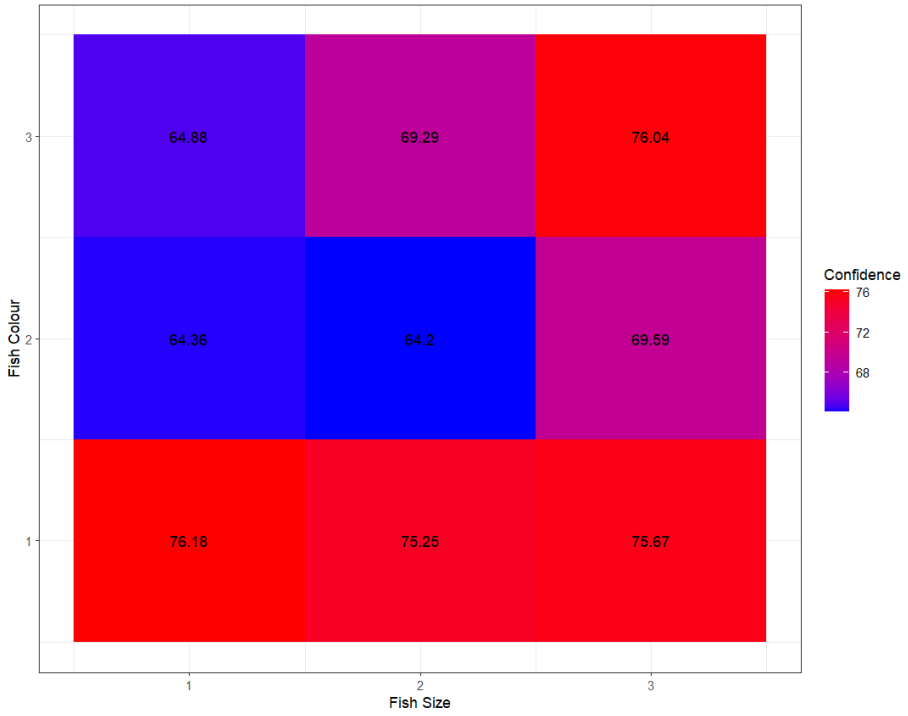


Figure 5.13: Grand mean confidence across all participants when averaged for each of the 9 grid boxes used to sample the stimulus space. The more blue the colour the lower the confidence, the more red the colour the higher the confidence. The numbers correspond to the rounded average confidence value.

The second prediction made stated that of the L-category grid boxes, grid box [1,1] would have the highest confidence in that category. This was examined using a within-participant ANOVA exploring confidence and a collapsed distance metric which compares the corner of the L-category (grid box [1,1]) to confidence 1 grid-distance-step away (a combination of data from grid box [1,2] and grid box [2,1]) and confidence 2 grid-distance-steps away (a combination of data from grid box [1,3] and grid box [3,1]). When the data was arranged in this way, as predicted, numerically, grid box [1,1] had the highest averaged confidence (76.2)

compared with the confidence 1 (69.8) or 2 (70.3) grid-distance-steps away. The ANOVA indicated that there was a main effect of distance from the corner location in the L-category ($p < 0.001$). The preregistered t-tests exploring this effect indicated that the differences were found between the different distance steps relative to grid box [1,1] ($p < 0.001$) but that the two distance steps did not have significantly different confidence reports to each other ($p = 0.57$). These results are consistent with the pre-registered predictions, however, as is clear from the Fig. 5.13 that this pattern is entirely driven by the difference in salience for the dimensions of colour versus size difference in the L-category. Note that there is essentially no difference between grid box [1,1] versus grid boxes [2,1] and [3,1].

It was also predicted that the computation of confidence will vary in the □-category and L-category with the latter engaging a rule-based model that weights one dimension over and above the other in different parts of the category space: e.g., once a trial is small enough on one dimension the other dimension value does not matter as it must belong to the L-category. Investigating this idea, the third prediction, that the mean confidence in the lower left grid box of the L-category (grid box [1,1]) would be significantly higher than the adjacent grid box in the □-category (grid box [2,2]), was tested using a t-test. Confidence reports were significantly higher in grid box [1,1] than [2,2], $t(55) = 4.79$, $p < 0.001$. This prediction was derived from the logical structure of the category space but contrasts with predictions based on simple distance-to-bound models, which would match the confidence for these two grid boxes (cf. the "pure"

models in Fig. 5.6 are symmetrical in these two grid boxes).

In order to examine prediction 4, that there will be an interaction between the category (L and □) and the distance from the boundary based on the arm of the L-category being examined, a 3-way ANOVA on confidence between distance, arm, and category was conducted, as in Study 4. As with prediction 3, above, this prediction derives both from the category structure and the findings of Study 4 and is contrary to the predictions of a distance-to-bound model of confidence. The stimulus space was again subdivided as shown in Fig. 5.12. The ANOVA indicated that the main effects of distance [$F(1, 55) = 66.62$, $p < 0.001$] and arm [$F(1, 55) = 39.39$, $p = 0.001$] were significant. Unlike in Study 4, the main effect of the category was not significant in these data [$F(1, 55) = 1.62$, $p = 0.2$]. Replicating the results of Study 4, the two-way interactions between distance and category [$F(1, 55) = 23.25$, $p < 0.001$], and arm and category [$F(1, 55) = 15.67$, $p < 0.001$] were significant. Also as in Study 4, the three-way interaction ([$F(1, 55) = 1.9$, $p = 0.17$] and the two-way interaction between distance and arm were not significant [$F(1, 55) = 1.21$, $p = 0.28$]. The main effects suggest that confidence is sensitive to both the distance between the stimuli and the category bounds, as well as the different dimensions used to define them. As before, an important finding is that confidence changes more quickly as a function of distance from bound in the L-category than in the □-category, as seen in the two-way interaction between distance and category. There also appears to be a different effect on confidence in each arm of the data depending on the category under study.

To account for the over-representation of data from the grid box [2,2] in this analysis, as above in Study 4, the 3-way ANOVA was also decomposed into two separate 2-way within-subject ANOVAs with independent variables of the category (O or L) and distance from the category bound (near or far), and the dependent variable of mean confidence report. Specifically, this sought to account for the fact that some data points contribute to more than one cell of the factorial design – data points in the lower left quadrant of [2,2] appear in the "near" distance bins of both arms of the □-category. Therefore, one of these 2-way ANOVAs was performed on each arm of the L-category and the mirrored stimulus space in the □-category. This was done to ensure that the key two-way interaction between the distance and category – i.e., the effect that suggests that distance is operating differently in each category – was not the product of double-dipping the data in the grid box [2,2]. The interaction between distance and category was significant in both the horizontal [$F(1, 55) = 10.31$, $p = 0.002$] and vertical arms [$F(1, 55) = 19.91$, $p < 0.001$] – suggesting once again that the effect was not driven purely by the double sampling that occurred in the three-way ANOVA, and that there is a clear asymmetry in distance effects either side of the category boundary.

To explore whether the processing of the two dimensions in the L-category was different, i.e., if one dimension was more salient than the other, the data relevant to prediction 5 was explored. This prediction suggested that participants would be more confident in trials of the L-category that are defined by colour than size as colour is a more salient category cue, this relates to those along the

horizontal arm in grid boxes [1,1], [1,2], and [1,3] where a stimulus is known to belong to the L-category once it is "red enough". Fig. 5.13 illustrates the spread of confidence across the stimulus space. Two things should be noted, foremost, the striking replication of the relative confidence values as compared with Fig. 5.5 from Study 4, and secondly the clear differentiation between the horizontal (colour defined) and vertical (size defined) arms of the L-category. Numerically, the trials in the L-category that can be defined by colour are more than 10 (10.84) confidence units higher than those trials defined by size, $t(55) = 5.07$, $p < 0.001$. This suggests that colour is either easier or more salient to process, a trend that is also hinted at in the category-arm interaction of the three-way ANOVA, and the resulting decomposed results in the two-way ANOVAs across the two arms.

Preregistered Exploratory Analyses

The modelling and subjective response-bound fitting from Study 4 were preregistered as secondary and exploratory analyses. No specific predictions were made about the direction or landscape of the expected results. As in Study 4, three families of models were used to simulate confidence values. The "pure" model family calculated confidence based on distance-to-bound metrics in both the L and □-categories. These models represent the classic "distance-to-bound" theories of confidence computation. The asymmetric family used distance-to-bound calculations of confidence in the □-category but in the L-category, the confidence was calculated based on the distance relative to the bound in that arm of the data

– i.e., the vertical "or" the horizontal boundary. This created an asymmetry between the category computations. Similarly, the "scale" family of models used distance computations in that □-category but scaled confidence relative to the probability space in the L-category (i.e., how small or red the fish stimulus was, rather than how distant its size and colour was from the category boundary).

Within each family of models, three approximations of "distance" were used. "Minimum distance" metrics estimated confidence based on a scaled distance from the nearest boundary value - estimating the minimum distance from either boundary in the space. "Maximum distance" looked for the furthest distance to either of the boundary values in the space, standing as the direct opposition to the minimum computations. Finally, the "summed distance" added up the values for both boundary-relevant distances - summing the distance of the y value from the horizontal boundary and the distance of the x value from the vertical boundary. These three distance metrics were simulated in each family of models creating a 3*3 grid of models to fit the behavioural data.

Once again, correlations between the simulated model values and the real data were used as a proxy for the goodness of model fit. Second-order statistics were then conducted to complete a basic model comparison procedure. The R-values of the correlations of the nine models relative to the actual confidence reports were extracted and assessed in a 3*3 ANOVA examining the type of model (pure, asymmetrical, or scaled) and the model distance (minimum, maximum, or sum). This analysis indicated that the main effects of model distance [$F(1.05, 57.97) = 12.7$, $p < 0.001$] and model type [$F(1.24, 67.99) = 7.64$, $p =$

0.004] were significant. Furthermore, the interaction between model distance and model type [$F(1.18,64.89) = 7.72$, $p = 0.005$] was also found to have significant effects on the strength of the correlations emerging from these analyses. These results replicated the direction and strength of those of Study 4, with the addition of a significant main effect on model type here.

Fig. 5.14 illustrates the landscape of correlations mapping the model simulated values to the actual confidence reports. The mapping is starkly reminiscent of the second-order statistics drawn from Study 4, displayed in Fig. 5.9. Once again the performance of the "pure" family of models is weaker than the other models that modelled the asymmetrical processing across the category boundary. This may be because the "pure" models fail to capture the confidence difference, made explicit in this study in Prediction 3, between grid box [1,1] and grid box [2,2]. Similarly, as in Study 4, the "scale" family of models appears to function well across all distance metrics. Again mirroring Study 4, the minimum distance metric is the strongest performing distance unit within each family of models, with the scaled-minimum distance model outperforming all others.

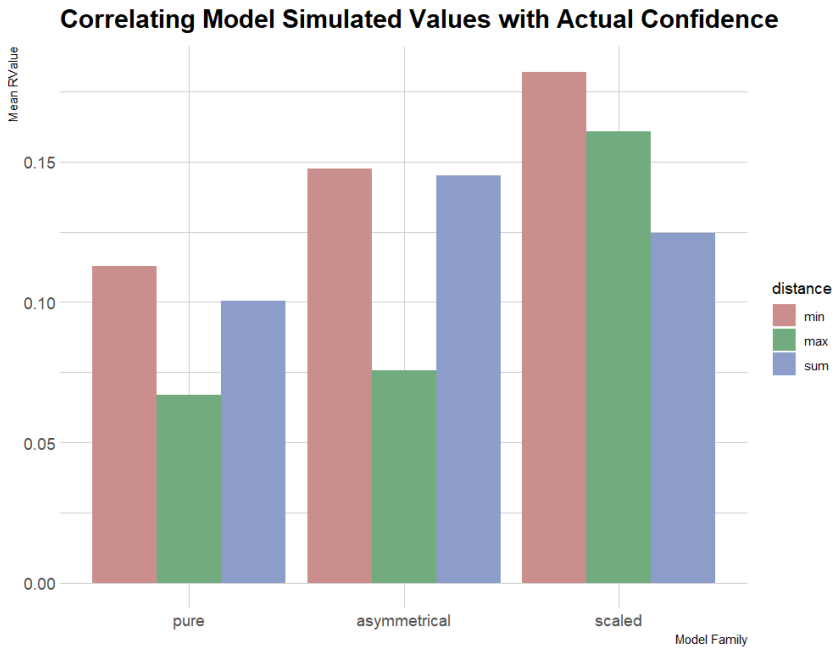


Figure 5.14: The mean R-values for each model are plotted here, as applied to the confidence data for Study 5 using the objectively defined boundaries. The bars are clustered based on the model type, while the colours indicate the approximation of distance captured by the individual model.

As in Study 4, these models were then re-fitted using the subjective response-based bounds as fit to each participant on the basis of their response patterns. The bounds were constrained such that at least 3 responses had to populate each quadrant. The patterns of responses and their relationship to the subjective bounds fit within each participant are shown in Fig. 5.15.

Another trend of note is the variability in participants' response patterns which is captured by the placement of the bounds for each participant. For example, comparing the fitted subjective bounds for PID 2 and PID 33 cap-

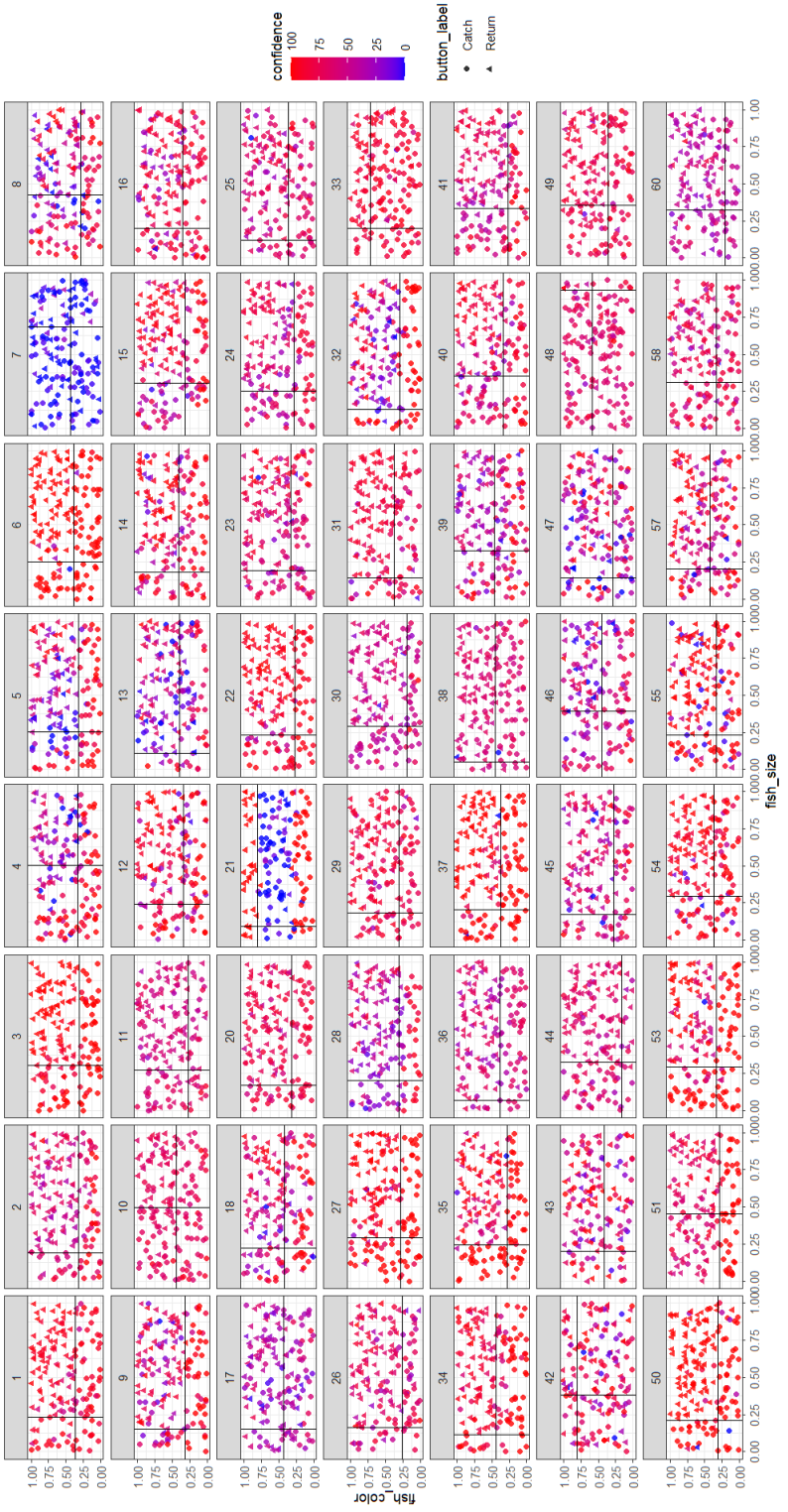


Figure 5.15: Each panel represents the response data, coloured by confidence in that response, per participant. The shape of each point represents which response the participants gave - i.e. what category they thought the trial belonged to. The fitted subjective boundaries are illustrated by the intersecting black lines on each panel. Note the variation in the category representations being used by the participants as inferred from these response-based boundaries.

tures some of this. Interestingly, looking at some participants is clear from both the bounds and the confidence gradient that some participants are integrating information from both category dimensions (e.g., PID 41) while others use only the colour dimension (e.g., PID 21). Based on these newly established subjective boundaries, I re-simulated the model-based confidence values using the subjective response boundaries. The correlations with the true confidence values were re-calculated. As in Study 4, an ANOVA was conducted on the R-values of these correlations acting as a second-order statistical model comparison procedure. The ANOVA indicated that, as with the analysis conducted on the objective category boundary, the interaction of model type and model distance had a significant effect on the strength of the correlations [$F(2.05, 112.83) = 13.03$, $p < 0.001$]. The main effects of distance [$F(1.65, 90.79) = 3.03$, $p = 0.06$] and model type [$F(1.09, 59.81) = 3.57$, $p = 0.06$] were not significant here

The average R-vales for each correlation are plotted in Fig. 5.16. Again the resemblance to the data of Study 4 is strong. While this is to be expected as the same analysis pipeline is leveraged to implement the response bounds, and the correlations based on the objective boundary were highly similar to those of Study 4, the strength of the resemblance is stark. This suggests that, on average the datasets collected in the studies are mirrored and that the patterns of variability seen are not purely noise-based but instead may reflect some specific tendencies toward specific processing styles across the category spaces in different segments of the participant pool. For example, across both studies and all versions of the analyses, in general, the family of "pure" models provide

the worst fit to the data. This is as expected given that, for example, this family of models predict a symmetry of confidence in grid boxes [1,1] and [2,2] which is contrary to the basic features of confidence in this space. Furthermore, in general, the family of "scale" models do best. Within this family, the "maximum" distance fits the most poorly. Whether "minimum" or "summed" distance fairs betters appears to depend on whether the experimenter imposed fixed and objective bounds, or the variable within participant-fitted response-based subjective bounds are assumed. While it is difficult to claim an overall best-fitting model based on these trends, particularly as none of these models begins to describe the asymmetry between the arms of the L-category that is a key feature of these data. However, the asymmetry captured in the "scale" family of models between the L and □-category seems important and is reflective of the broader trends seen across the wider analysis pool. Importantly, in contrast to the "asymmetrical" family of models which also predict different values of confidence on either side of the category bound as contrary to the distance-to-bound model, the "scale" models predict higher confidence in grid box [1,1] as compared with the arms of the L-category. While this pattern is not absolute, as it does not also demonstrate the difference between the vertical and horizontal arms of the L-category, it does highlight a key feature of the data – the high confidence of the grid box [1,1] trials – which may drive the fit of the "scale" family of models over and above the "asymmetrical" models.

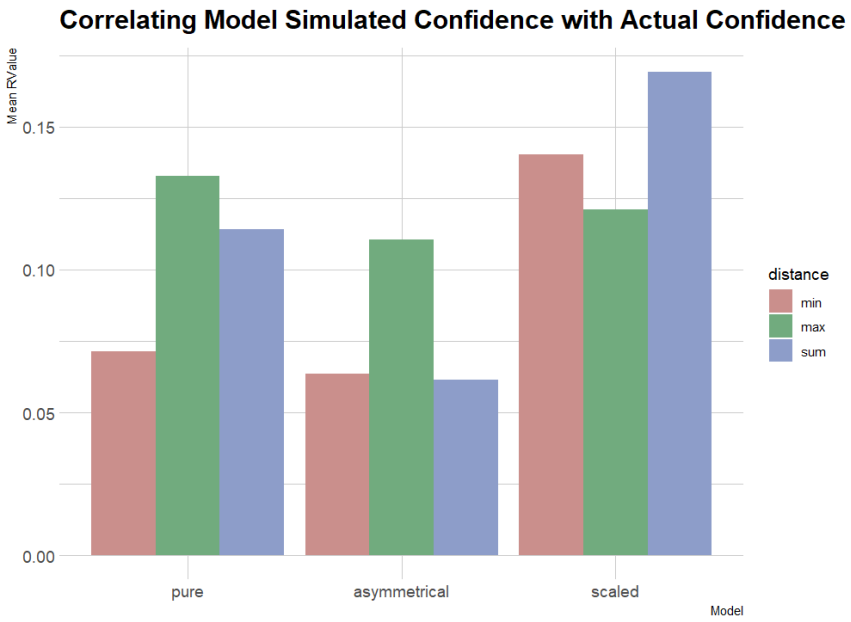


Figure 5.16: The mean R-values for each model, as applied to the confidence data for Study 5 using subjectively defined boundaries. The bars are clustered based on the model type, while the colours indicate the approximation of distance captured by the individual model.

Another line of preregistered secondary analyses includes investigating the reaction times of participants across the stimulus space and how that compares to the patterns of confidence. This is relevant as the work of Nosofsky and colleagues (e.g. Little *et al* (2011), Fific *et al* (2010)) leans on reaction times as an outcome variable to differentiate the logical models of categorisation being used by their participants. It is expected that these results will correlate strongly – as the values of confidence trend with reaction time on average (Pleskac & Busemeyer, 2010; Vickers & Packer, 1982). However, it is unclear if the larger scale of possible response times will allow for greater or lesser differentiation

between the categories as compared with the metacognitive confidence reports.

Looking at the landscape of response times, as measured by the time (in milliseconds) between stimulus onset and the response recorded by button press, in Fig. 5.17 a few comparisons can be made to the confidence values of Fig. 5.13. Foremost, as expected the overarching pattern is similar with the fastest reaction times occurring in trials drawn from the horizontal arm of the L-category. Similarly, the slowest trials matched the lowest confidence trials in grid box [2,2] in the □-category. However, there are some differentiations in Fig. 5.17 which do not appear in the confidence values, such as the intermediate response time of the vertical arm of the L-category. These grid boxes had the lowest confidence values in Fig. 5.13, but not the highest reaction times, which were instead found in the grid box [2,2]. This suggests that perhaps confidence was sensitive to the uncertainties that emerge from the processing of the size values in this arm of the L-category in a way that cannot be captured by reaction time alone.

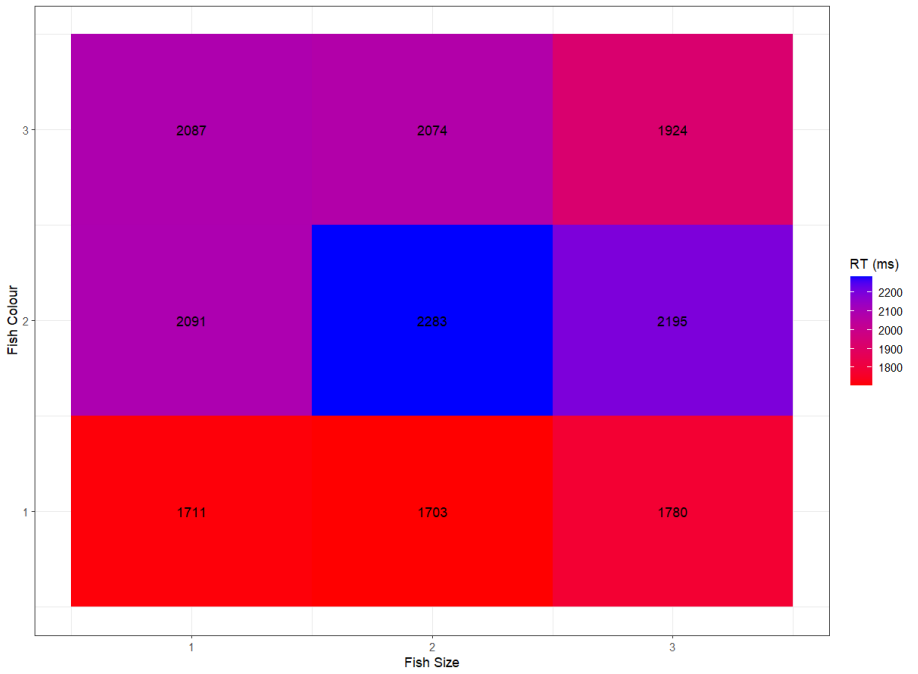


Figure 5.17: The mean response times (in milliseconds) for each grid box of the stimulus space are plotted here. Red colours indicate fast times (which correlated with higher response times, and blue hues denote slower response times.

Statistically speaking, predictions 1 and 3, which compared confidence values for grid boxes [2,2] and [3,3] and grid boxes [1,1] and [2,2] respectively were replicated in the reaction time data ($t <$ -5.1 and $p < 0.001$ in both cases). Interestingly, however, while the confidence values between grid [1,1] and [3,3] were not significantly different, the reaction times were, $t(55) = -2.95$, $p = 0.004$). While the confidence values were high in both grid box [1,1] and grid box [3,3], reaction times were faster for category judgements made in grid box [1,1] than in grid box [3,3]. This difference in reaction time aligns with the serial application of rules as described by (Little et al., 2011), however, the equitable confidence

values indicate a unique feature of confidence in these logically defined categories - that even if the trial-wise processing time is long the confidence report on a trial may still be high once processing is complete – because the "AND" rule of grid box [3,3] have been satisfied, for example.

Exploratory Analyses

As with Study 4, a final set of exploratory analyses was conducted on the qualitative text response item that participants completed post-experiment. These responses were prompted by the following prompt: *"Please describe what strategy you were using to reach your decision about whether or not to catch a fish."*.

Of the 56 participants who were included in the analyses, N = 54 responded to this prompt, of whom N = 51 gave valid responses that mentioned a strategy related to at least one of the category dimensions. N = 16 responses mentioned using both dimensions, with some level of integration for both categories, of whom N = 14 mentioned colour prior to size, e.g., *"The colors and sizes of the fish told me they were more likely to be invasive or native." (Participant 39.* When only 1 dimension was mentioned it was colour rather than size in 4/6 responses, e.g., *"I tried to group the fish in terms of closeness to red or blue, in respect to trying to memorize the trend in invasive fish colours that lean more to the red colour and the native fish colours ranging more on the blue/ purple colour." (Participant 46).* When only one category was mentioned it was overwhelmingly the L-category responses (the invasive species) that were the focus of the text response (8/9 responses) e.g., *"I focused on the things*

that made the fish that I need to catch unique, their size, or specific color." (Participant 16).

In terms of differences in expressing the processing of the two categories, a number of responses described the L-category in terms of serial or separate one-dimensional judgements on the two dimensions (N = 11). E.g. *"I noted the red ones as well as all the small ones were invasive" (Participant 50).* Within these responses, the majority (7/11) then defined the □-category on the basis of one dimension only - size e.g., *"I found that the small round fish and the bright red ones were always invasive and the large ones were always native" (Participant 22).* This suggests a deconstruction of the task and a purposeful focus away from the colour dimension in the □-category where the blue tones were non-identifying unless the size was above the boundary.

Collectively these data are suggestive of both the salience of the colour dimension for many participants and the differential processing that occurred across participants. This variable processing was seen both in terms of the single-dimension classification versus integration of the two dimensions as well as the differential approaches used by participants across the categories between the L and □ categories.

5.3.4 | Discussion

Study 5 acted as a direct replication of Study 4 with confirmatory predictions and specific analyses detailed. This was a necessary addition to the investigation as it was not clear if the nuanced analyses, which yielded interesting results, were actually differentiating between signal and noise in the data. In Study 5, the results indicated that the patterns and trends of confidence within and between the categories in the experimental stimulus space were robust across both samples.

The predictions regarding the differences in the behaviour of confidence, implying a difference in the functioning of distance metrics, across the category boundary were upheld. For example, the differences in confidence between both grid boxes [1,1] and [2,2] and grid boxes [2,2] and [3,3] demonstrated the different processing of confidence relative to distance in the L-category relative to the □-category. This asymmetry in confidence processing relative to distance was also evidenced in the interaction between distance and category in both arms of the data.

This tendency toward asymmetry across the bound was also evidenced in the replicated results of the model-based analyses in which the "scale" models, which computed confidence differently in each category, were the best fit to the data in both the objective and subjective boundary analyses. The predicted asymmetry of confidence, as related to the differential processing of the two dimensions in the L-category, was also upheld.

5.4 | General Discussion

The experiments in this chapter investigated confidence in discrete, multi-dimensional category spaces. The use of this type of category structure allowed for specific investigations of distance-to-bound models of confidence to be carried out. While these stimulus spaces – with multiple dimensions, discrete bounds, and an asymmetry of decision rule application – have been used to explore logical rule models of categorisation (Fific et al., 2010; Little et al., 2011), the expected behaviour of metacognitive confidence is unclear relative to the classic theories of confidence as distance-to-bound (Kepecs et al., 2008; Komura et al., 2013; Vickers & Packer, 1982).

By eliciting confidence in a two-dimensional space with a discrete bound that exists at the intersection of those dimensions, several research questions arose. In particular, how confidence should operate around a discrete category boundary, as opposed to a graded perceptual overlap, was of interest. Indeed, whether confidence should be expected to reflect the discontinuity was unclear. Furthermore, the definition of "distance", and therefore the behaviour of confidence reports based on classic distance-based models, was less apparent in categories defined based on multiple dimensions. Finally, the inbuilt asymmetry of categories used, with one lending itself to an "or" rule-based classification system (red or small), while the other was defined based on the intersection of the two dimensions (blue and big), was of interest. It was unclear if confidence reports would reflect these differences in processing in a way that would suggest

a different computation of confidence on either side of the discrete bound.

The results across the exploratory and confirmatory versions of the task were replicated robustly across the two experiments. The specific tests of differences in confidence processing within and across the boundary were upheld in both datasets. Importantly, these results suggested that the operation of confidence in each category was functioning differently.

Foremost, the difference in confidence values in trials based on stimuli drawn from the grid box furthest from the boundary (i.e., the top right-hand corner, grid box [3,3]) had higher self-reported confidence as compared to those trials drawn from closer to the boundary, with the lowest confidence emerging in trials drawn from the grid box on the bottom left of the □-category (i.e. the bottom left-hand corner, grid box [2,2]). This gradient of confidence in the □-category, which replicated across both qualitative visualisations (see Fig. 5.5 and Fig. 5.13) as well as in quantitative analyses.

By comparison, confidence in the L-category appeared to be less graded. Indeed although, on average, the results suggested a descending confidence gradient moving towards the grid box [1,1], in Study 5, this averaging obscured the clear qualitative differences visible in both studies between the level of confidence in the two arms of the L-category. Analytically speaking, analyses indicated that distance, as defined as being near or far from the bound, was behaving differently in the L versus the □-category on both arms of the boundary based on the significant distance-category interactions. A key finding was that confidence changed more quickly as a function of distance from bound in the L-category

than in the □-category, as shown by these interactions between distance and category. Furthermore, the different dimensions (as indicated by the effect on the "arm") seemed to have different effects on confidence in each category. These differences across the category boundary were complemented by an investigation of the differences in confidence in grid boxes [1,1] and [2,2], and further grid boxes [2,2] and [3,3]. These analyses suggested that confidence in trials in the □-category appeared to align with a distance-to-bound description of confidence, while computations in the L-category did not. This raises questions about how confidence is computed in relation to the logical category structure and rule-based decision-making strategies that are employed by participants in the L-category.

Modelling distance in this space allowed for consideration of how to capture the idea of distance-to-bound in a multi-dimensional and geometric category space, as well as how to describe confidence based on the asymmetries evidenced across the bound. By testing different models of distance, which captured different units of distance as well as different levels of symmetry across the boundaries, a few key findings were upheld. When fit relative to the objective category boundary, the "pure" model family, where confidence was based purely on distance in both categories, performed relatively poorly. This may be due to the inability of this model to account for the pattern of confidence difference in grid box [1,1] and [2,2]. However, in both data sets the scaled-sum model showed the strongest correlations to the data on average. This model has a few key features to highlight. The scaled-sum model captures an asymmetry

between the categories with distance being used in the □-category and both dimensions being relevant to that conceptualisation of distance. This means it describes confidence in the □-category as a distance-based metric that integrates both category boundaries in its computation. Furthermore, the "scaled" confidence computation in the L-category seems to capture the high level of confidence in the grid box [1,1] by looking at the probability of the stimulus features relating to category membership. However, even this "best-fitting" model failed to capture key features of the empirical data including the asymmetry of confidence across the two dimensions in the L-category.

The qualitative data analysis in both studies complements these findings. Several key threads are worth reiterating. Foremost, the salience of the colour-based dimension was apparent in the descriptions used. Furthermore, an overarching focus on the L-category was seen with participants prioritising the classification of these fish and other stimuli being "othered" relative to those rules. This may be because of the ease of rule learning in this category, or the increased relative risk associated with this stimulus category which incurred a cost if mis-classified. It would be necessary to repeat the study design with a flipped reward matrix such that the risky category sat in the □-space in order to understand the origin of the focus on the L-category more clearly.

These results hark back to the discussion of rule-based learning based on the ability to articulate a simple rule (Livesey & McLaren, 2009) discussed in previous chapters. The ease of this articulation process was greater in the L-category and was dominated by the colour dimension with statements of "All

red fish were invasive" being highly prevalent across both studies. This type of processing was less developed in the □-category where a simple rule seemed to be less intuitive as, if a simple rule were to be employed based on size, the colour dimension had to be discarded on those trials. Although some participants appeared to achieve this kind of expertise, as inferred from their single-dimension rule-based descriptions of the □-category, they were not in the majority.

The evidence presented in this chapter suggests that the predictions of simple distance-to-bound models of confidence fail to capture the apparent sensitivity of confidence to complex task features. For example, the rule-based nature of many categorisation tasks (Bourne, 1970) create models of processing which make specific and directional predictions about reaction times (Fific et al., 2010; Little et al., 2011) which can be extrapolated to confidence. The nuances of these models, as necessitated by the richness of the category structures, test the distance-to-bound model of confidence in ways that perceptual discrimination tasks, by definition, cannot. While the exploration of these types of structured and rule-based category spaces highlights the paucity of the predictions of distance-to-bound models relative to the empirical data in multi-dimensional categories with discrete bounds, it does not necessarily specify the origin of the confidence reports themselves. If not based on distance on a trial-by-trial basis - where precisely are metacognitive confidence reports emerging from in these spaces?

Particularly in spaces where the integration or serial processing of multiple dimensions is required, and where the salience of those dimensions is not neutral,

the origin of confidence is not apparent. By combining the observations of the confidence and reaction time data across the L and □-category one pattern stands out that may help to begin to unpick this problem: Whereas there is a clear asymmetry between the dimensions in the L-category, this is not seen in the □-category. This makes sense in terms of the fixed-serial application of logical rules (Little et al., 2011), if people checked colour first and then size the asymmetry of the L-category would be apparent. The problem is then – how to combine the two dimensions when categorising trials belonging to the □-category?

In terms of the latter, it can be noted that both confidence values and reaction times are very similar for grid boxes [2,3] and [3,2] across both data sets. The stark gradient of the □-category, from grid box [3,3] to [2,2] may be best described by a summing of the distance across the two boundaries (as seen in the "sum" distance models in this chapter) – or as per the symmetrical geometry of the space taking the average distance to the boundary across the two dimensions to be integrated. Yet, even if this hybrid distance model functions to effectively approximate confidence in the □-category, a distance model of confidence would still fail to accurately capture the asymmetries of the L-category which cannot be accounted for without an additional dimension-based weighting function.

Taken together, the results of this chapter suggest that the participants were sensitive to the asymmetry in single versus dual dimension classification in the different categories and that this was expressed in their confidence reports. This asymmetry of fit is aligned with theories of logical models of categorisa-

tion which have previously been explored using response times in this type of stimulus space, (Little et al., 2011). The alignment of the patterns of confidence across the category bounds with the families of models that described the space in an asymmetrical manner indicates that an approach in which the conceptualisation of confidence in the two categories was distinct would better describe the behavioural data. This suggests that confidence is not described by distance alone, or at least does not use distance in the same way on either side of the category bound, and instead may arise from context-based computations which account for the task structure and the relevance of different information to the decision or category classification at hand.

CHAPTER 6

Discussion & Synthesis

6.1 | Introduction

Across the chapters of this thesis, decision-making and decision confidence have been investigated in multiple behavioural tasks with different parameters, decision boundaries, and distribution properties. Through these experimental designs and their accompanying analyses, the theoretical and methodological intersections between the decision confidence and categorisation literatures have been explored. What began as an investigation into task demands and approximations of uncertainty evolved into a series of experiments that sought to leverage the distinctive features of categorisation tasks to better conceptualise metacognitive decision confidence and its utility in understanding cognitive representations. In this way, this thesis attempts to begin mapping out the usefulness of metacognitive confidence in categorisation tasks in order to understand what sort of information is being attended to in what types of contexts, and how that translates to confidence computations themselves. By allowing the intersection of these theories of categorisation and decision confidence to become central in the discussions, the sensitivity of confidence to the structure and properties of category distributions and its utility as a lens to understand cognitive representations of categories becomes clear.

Taken together, the three experimental chapters detailed in this thesis begin to illustrate the richness and diversity of the information processing and decision space representation that occurs as part of the flexible cognitive processing in human decision-making. The manipulations used in my experiments

attempted to emulate some of that diversity. For example, by stepping away from perceptual discrimination tasks which often impose a hard boundary in a uniform distribution across the feature space, Chapters 3 and 4 explored the sensitivity of confidence to the graded membership often expressed in natural categories by creating overlapping distributions to define category membership (Rosch, 1973; Rosch & Mervis, 1975). By varying the distribution properties in Chapter 4 and creating categories with different variances and separable means, I sought to understand the relationship between confidence and these more ecologically valid descriptions of perceptual feature overlap which can be seen in everyday examples of category judgements. From deciding if an umbrella is needed based on the colour of the clouds, to a medical professional grappling with complex presentations of symptoms, these more varied category spaces are ubiquitous. In Chapter 5, the use of multi-dimensional category definitions highlighted, as described by confidence, the different kinds of classification that can occur across a category boundary. Whether it is sorting cups and bowls, or tables from chairs, these kinds of spaces highlight the variety of strategies that can be used to make sense of the diverse families of categories in the world.

Conjointly, these experiments help to illustrate how category confidence can be understood in terms of its behaviour and derivation in a variety of category spaces as well as conversely highlighting the methodological utility of category tasks in interrogating models of decision confidence. Therefore, this thesis stands as a small contribution to this branch of scientific exploration although the task designs, category spaces, and analysis pathways leveraged remain far

from exhaustive.

Over the course of this thesis project, the direction of the work began to shift away from more computational research, by necessity, towards well-defined exploratory behaviour tasks. The project began with secondary data analysis of a more classic perceptual decision-making task which, while not purely perceptual because even with perfect perception the task cannot be done perfectly, represented a more standard approach to probing confidence. The work then pivoted and the task manipulations and structures evolved until discrete multi-dimensional categories became the focus of the final experimental chapter. The shift was driven by a need for greater control of the experimental tasks in order to assess the patterns of confidence more explicitly. However, this shift also leveraged interesting and rich category spaces to address the modelling of confidence in a diverse set of tasks.

Through this work, it has become apparent that metacognitive confidence can act as a valuable lens for understanding the task representations used by participants when learning about and navigating through novel or complex stimulus spaces. The sensitivity of confidence reports to the dimensionality, the variable spread, and the boundedness of categories and category spaces suggest that it may be meaningfully used to test theoretical models of categorisation and decision-making. Perhaps more clearly, these kinds of spaces allow for new insight and understanding into the nature and basis of confidence. These ideas and the conclusions that can be drawn from the findings of this thesis will be detailed in the subsequent sections starting with a summary of the results, an

integrative summary of the themes, and concluding with a discussion of future directions and conclusions that emerge across the span of the work discussed here.

6.2 | Overview of Results

The first empirical chapter, Chapter 3, aimed to explore whether decision-makers' confidence reports aligned with a Bayesian model of confidence in terms of using the relevant distributional properties of the categories appropriately. This exploration took the form of a model comparison exercise to assess whether a Bayesian or rule-based heuristic model of confidence best described the confidence reports generated under different task demands. This work emerged from studies suggesting that human decision-makers would adopt different approaches to learning and navigating tasks depending on task difficulty (Wills & Mackintosh, 1998) and the ease with which a verbal rule can be generated to define the task space (Livesey & McLaren, 2009). Broadly speaking, therefore, this chapter began to explore the balance between the objective accuracy of ideal observer models and the subjective utility of heuristic models that describe task space well enough to allow for effective navigation of the stimuli.

The results of this chapter were neither unambiguous nor in line with the original hypotheses. Whereas I had predicted that the best-fitting model would change relative to the difficulty within individual participants, the data suggested that participants tended to persist with a single cognitive representation of the task based on the session difficulty during their initial exposure to the stimuli. Indeed, participants' behaviour was best captured by a mix of different but relatively stable strategies across individuals. Although a full Bayesian model, for which confidence reflects a normative calculation of the

P(*category*|*stimulus*), was found to be the overall best-fitting model at the group level, it did not tend to provide the best fit for many individual participants. By contrast, a rule-based heuristic model of confidence, for which the decision rule was based on a discrete boundary and confidence was judged relative to that bound based on a noisy stimulus percept, was a good fit for more participants.

It may be possible to extend the approach used in Chapter 3 to provide clearer evidence for Bayesian representations. For example, as discussed in Chapter 3, the results may be disambiguated by reducing the stimulus noise or by creating a between-participants design. However, in this thesis, I opted for a different approach that involved looking for stronger evidence by varying the categorical properties more directly. In Chapter 4 this corresponded to manipulating the variability of the distributions, changing the distribution properties directly. By manipulating the underlying structures of the categories, it became possible to look for effects on people's confidence reports more easily without having to assume task-demand-based differences in distribution representation.

Therefore, moving forward from these initial modelling attempts, the behavioural studies in Chapter 4 built on the idea that directly manipulating the distribution properties would produce clearer results as compared with inferring patterns from a standardised task. The success of this approach, as compared with the ambiguity of Chapter 3, was apparent. Behaviourally the patterns of confidence showed a sensitivity to the category distribution variances as would be expected of more sophisticated models of confidence. Confidence was greater

when judging stimuli as belonging to a category with a wider, as compared with a narrower, distribution in stimuli space. This evidence is consistent with the idea of Bayesian encoding of category spaces and distribution representations (Beierholm, 2013). Indeed, when variations of Bayesian models of confidence were assessed on these data, a model that captured the asymmetry of the distributions fit better than one that simplified the task space by equalising the distribution variances, even within a Bayesian framework. However, this is not to suggest that the true underlying mechanism of these confidence patterns is Bayesian. Indeed, a more sophisticated distance-to-bound model could capture this asymmetry of slope with an arguably lower computational load than the comparative Bayesian models. It is unclear how those two model families could be best disambiguated in these data, or whether they would result in meaningfully different conclusions being drawn. What is clear, however, is the sensitivity of confidence as a self-reported variable to the distribution approximations of the participants.

Chapter 5 explored these ideas in a complex category space where the behaviour of decision confidence was difficult to predict because of the multidimensional nature of the category structure. From the discrete category boundary and the asymmetry of the salience of the dimensions across the category space, predictions based on classical distance-to-bound models of confidence struggled to capture the behavioural and confidence data (Kepecs et al., 2008; Komura et al., 2013; Vickers & Packer, 1982). Instead, the data suggest that confidence would be better described by distinct models in each category in the

stimulus space. There were asymmetries in the strength of confidence relative to distance to bound in each category and the data appeared to be best fit by a mixed model which reflected this imbalance. The asymmetrical sensitivity of participants to the different category dimensions was magnified by the asymmetry in single versus dual dimension classification in the different categories, which emerged as AND/OR style use of the dimensions across the bound. This asymmetry of fit was in line with theories of logical models of categorisation which had previously been explored using response times in this type of stimulus space, (Little et al., 2011). The alignment of the patterns of confidence across the category bounds with the different families of models that could describe the space was complemented by the modelling attempts made. Overall, the family of models which did not capture any type of asymmetry across the bound did not correlate well with the data, by comparison, models which built in the asymmetry of computations of confidence across the boundary appeared a better fit to the data. While these models still failed to capture the nuance of the space fully, they did appear to indicate that an approach in which the conceptualisation of the two categories was distinct would better describe the behavioural data.

6.3 | Themes

Across these chapters, several key threads can be drawn out and the relationship between the tasks and the corresponding results lends itself to an integrative discussion. Foremost, a clear link across the three empirical chapters, and their results, is the tension that exists in the literature – and in the current data – between Bayesian and heuristic models of confidence. Based on the first set of model comparisons, in Chapter 3, several assertions could be made. Foremost, individualised and variable representations of tasks and decision spaces were both possible and probable. For example, this was evidenced by the variation in best-fitting models across participants, which appeared to depend on the initial task-based experience and subsequent persistence of the task representation. Furthermore, as seen in Chapter 4, a Bayesian model of confidence – with its corresponding sensitivity to distribution variance – was a reasonable fit to the data in both instances. Although in Chapter 3 the Bayesian model was not the best-fitting model for many participants, it was numerically the best fit overall as compared with either dynamic model types which changed based on task demands or a rule-based heuristic model – despite the theoretically grounded predictions that were made (Wills & Mackintosh, 1998). Similarly, the Bayesian model that captured the asymmetry of variance in Chapter 4 was a better fit to the data than a simplified symmetrical Bayesian model. While a comparison cannot be made directly between the two datasets, or the model comparisons therein, the conclusion is similar. A more complex model

that encompasses an accurate description of the distribution properties at hand served as a better approximation of the metacognitive confidence of the task participants. However, whether this suggests these participants were acting as ideal Bayesian observers in these tasks is much less clear. Once again the utility of models is apparent - as is the potential that they are inaccurate (Box & Draper, 1987).

Indeed, based on these data it is not possible to ascertain if patterns of decision confidence in these tasks are best described by ideal observer models or not. It is entirely plausible that another mechanism creates patterns of output that appear to adhere to the Bayesian principles established in these analyses (Geisler & Ringach, 2009). As described above, a more involved distance-to-bound model could capture this asymmetry of slope with an arguably lower computational load than the comparative Bayesian models. This complexity, the attempts to disambiguate between what are often deemed to be "optimal" decision processes (Bowers & Davis, 2012) and heuristic alternatives, is not easily solved (Adler & Ma, 2018). However, the discrete category space leveraged in Chapter 5 lends strength to the notion that optimality is context-based. The asymmetry of the required approaches in the discrete multi-dimensional task suggests that it is not always necessary to leverage all information and integrate it into a classic ideal observer model in order to achieve "optimal" behaviour pattern, or metacognitive accuracy. Indeed, in the category space utilised in Chapter 5, there is a clear argument that in the L-category space the observer should ideally only attend to one of the two dimensions as the decision should be

clear-cut once that is assessed relative to the boundaries in the space. I would argue that, in fact, it would be inaccurate to assume that Bayesian modus operandi would produce more optimal results in terms of the relative time and cognitive capacity required to achieve a roughly parallel performance based on a single-dimension classification.

Regardless of whether Bayesian models are in fact correct, what is apparent is that metacognitive confidence does not simply depend on distance to bound in category tasks. A simple rule-based model, which approximates a distance-rule, is not the best-fitting model in Chapter 3 and it is contradicted by the patterns observed in the data in Chapter 4 and Chapter 5. There is a similarity in the deviation from the distance models in the results of Chapter 4 and Chapter 5 in terms of the predictability of the variation in confidence above and beyond distance-to-bound effects. In Chapter 4, confidence differed systematically as a function of variance, and it varied as a function of the intrinsic and logical AND/OR structure of the category space on either side of the discrete boundary in Chapter 5. This work may therefore serve as a counterpoint to existing literature, based largely on one-dimensional perceptual discrimination tasks, which suggest that formal Bayesian models of confidence poorly reflect the actuality of human metacognition but that variations of distance to the bound may better describe the data (Adler & Ma, 2018; Xue et al., 2023). Instead, my results suggest that, unlike the assumptions of simple distance-to-bound representations, confidence is instead a rich signal which captures information about the category and stimulus space representations held by the participants.

A second key consideration that emerges across these chapters is the utility of decision confidence as an outcome variable in these spaces. As discussed in Chapter 3, confidence is a rich and continuous signal as compared with the response data which is a simple binary. Confidence is also meaningful at the trial level whereas response accuracy is inherently an aggregate measure. The similarity of fit despite the inclusion of responses in the extended models of Chapter 3 suggests that confidence can alone be a helpful window into the mechanisms responsible for behaviours. Furthermore, the sensitivity of metacognitive confidence as a self-reported insight into the cognitive representations of task spaces appears to match behavioural variables such as decision reaction time often leveraged by categorisation tasks (Little et al., 2011). This was seen in the patterns of confidence and reaction time that were explored across the stimulus space in Chapter 5. In some places, confidence seemed to differentiate between aspects of the space that reaction time did not and vice versa. For example, although reaction times were slower for classifying stimuli which seemed to use two dimensions (e.g. grid box [3,3]) than the comparator space in the L-category (e.g. grid box [1,1]) confidence was consistent and high in both of these spaces – once the processing was complete the answer was clear for these stimuli. The conscious level of access to metacognitive confidence lends additional credence to the findings as does the use of confidence to feed forward into future learning and behaviour (Persaud et al., 2007; Rhodes, 2019). Collectively these ideas suggest that intervention at the level of confidence holds promise for adjusting away from the overconfidence bias seen in many domains

(Atir et al., 2015; Jackson & Kleitman, 2014) and that confidence is useful as a converging measure with reaction time and response accuracy to understand human decision-making and category judgements.

Another key thread in this thesis is the categories themselves and the act of categorisation as relevant to decision confidence. Indeed, categories are attractive both in terms of their methodological utility and in terms of what we can learn about cognitive representations through sampling of their spaces both within and across category boundaries. They are also fundamental to these discussions due to their psychological importance and their ubiquity. Moving away from the noisy perceptual uncertainty elicited in the original data in Chapter 3, the focus on the uncertainty at the level of classification or categorisation, as described by the overlap of the distributions, in Chapter 4 allowed for an explicit test of the distance-to-bound model and the sensitivity of confidence to distributional properties such as variance. This evolved further when the discrete boundary in 2-D space in Chapter 5 pushed the idea of "distance" further and opened up the ideas that decision-making, and confidence, could be operating in different ways in adjacent conceptual spaces. The design of categorisation tasks and the diversity of category structures lends itself to explicit testing of the assumptions and theoretical underpinnings of models of confidence.

6.4 | Future Directions

This thesis and its findings contribute to an already rich literature on models of decision confidence. However, by bridging into categorisation task spaces and by learning from distinct models and spaces in the category learning literature, the findings of this thesis map out a fruitful intersection of the two fields. The scope for future work based on this thesis is clear in the multitude of different kinds of category structures and the rich space of category definitions. Indeed, future directions from which this work could grow are manifold.

Speaking first to the overlapping asymmetry of the tasks used in Chapter 4, tasks based on distributions with a more extreme spread could test the limits of the fidelity of participants' confidence judgements to the task distributions. By having more extreme differences in the distributions it would become possible to test whether people only have a general estimate of the distribution variance or a quantitative computation. This could be seen based on fitting the variance of the distributions as a free parameter, or in the relative fidelity of the confidence reports across different differences of variance. This would enable a clearer test of the tipping point that might exist between a "true" representation of a distribution as compared with a "good enough" model of the space that achieves a reasonable level of task and metacognitive accuracy. Indeed, while the results of this thesis align with the idea that confidence is sensitive to complex categorical features, it is unclear how complex the underlying cognitive representations held by the participants are. This speaks directly to the afore-

mentioned tension and the related questions, between rational versus heuristic models of decision-making and decision confidence (Adler & Ma, 2018; Xue et al., 2023). For example, trialling stimuli distributions that have a clear skew in opposing directions could test the extent to which decision-makers attempt to represent the steepness of the slope and long-tailed overlap to the distributions when a simplified heuristic model could suffice to represent the vast majority of trials (see panel 1 and 2 of Fig. 6.1). Similarly, it would be informative to explore how confidence might be modelled, as a proxy for the distribution representations held by the participants, in a stimulus space that resembles a signal detection task. For example, consider the case where a uniform distribution spreads across the entire space in the case of one category, with the other category existing as a narrow peak in the centre of that distribution (see panel 3 of Fig. 6.1). Whether participants would represent the relative uncertainty in that narrow peak zone, which should exist on the basis of the uniform distribution overlapping absolutely with that space, is unclear. Instead, they could create two discrete boundaries that delineate the narrow peak from the rest of the space and confidence could either be absolute on either side of that boundary in line with a rule-based decision rule or could be derived relative to that bound.

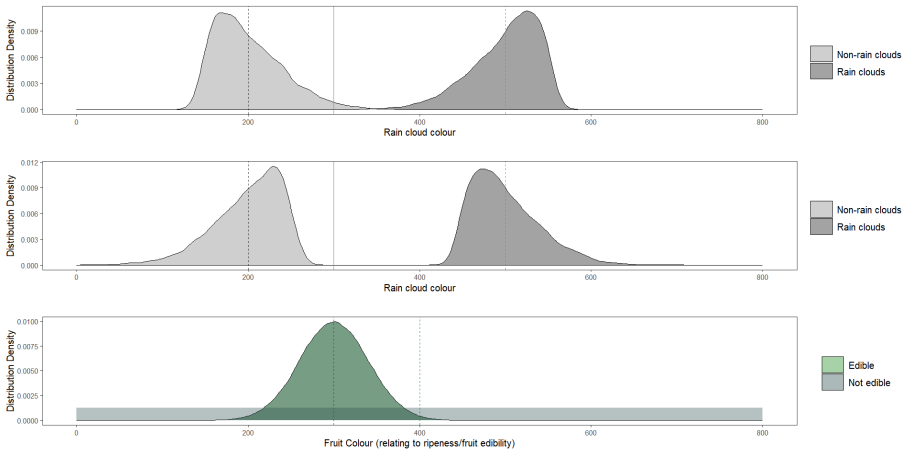


Figure 6.1: Panel 1 and 2 illustrate the kind of space where overlapping skewed category distributions might spread across the stimulus space, in alternate ways. How confidence would reflect the tails of these distributions is not evident. Panel 3 denotes two category distributions, one uniform spanning the whole space and one narrow normal distribution. Whether and how metacognitive confidence would account for the overlap of these distributions is unclear.

The scope of the possible work based on these findings extends when thinking about category learning and the sensitivity of confidence to distribution properties over time. Consider, for example, the impact of different levels of training on a task on the category distribution representation and how that might translate into metacognitive confidence reports. It is very plausible to predict that under-training would lend itself to low confidence reports across the space as there is a layer of task-based uncertainty that biases confidence downward away from the emerging realities of the category distributions. By comparison, achieving an expert level in the task, as a result of extensive training, could result in cognitive heuristic shortcuts that mean task accuracy diminishes slightly but cognitive load is reduced. While this can be detected in changes in accuracy,

the continuous nature of confidence reports, particularly as measured in this thesis allows for a more nuanced examination of the ways in which participants are representing the context they are navigating over time. This idea probes the possibility that the debate between Bayesian and heuristic confidence reporting does not lie solely in the adjudication between the two models overall but may instead depend on the degree of expertise or experience in a given environment or task.

Within multi-dimensional categories, the possibilities for further research are even more apparent. Beyond simply increasing the number of dimensions involved, which would in turn increase the complexity of weighting and integration across dimensions, the category boundary placement becomes more interesting in terms of how it relates to the expected behaviour of confidence and what it suggests about how individuals are navigating the space. Consider, for example, a multi-dimensional category that is nested within another in n-dimensional space, for example, a species of bird within a wider genus – such as a rockhopper penguin relative to other penguins. The category boundary becomes a sort of ring or circle (or sphere) within which all exemplars are rockhoppers and outside of which all exemplars are other penguins (see Collins & Quillian (1969), and Quillian (1967) for discussion of these kinds of hierarchies in semantic concepts, and Rogers & McClelland (2004) for discussion of how these relationships emerge in conncectionist networks). In this case, the distance-to-bound model of confidence could operate with some success but it is probable that a prototype-based category model would better approximate

the behaviour of confidence. In essence, it could be predicted that confidence be framed in terms of distance to the prototypical rockhopper penguin rather than the boundary between the species and the wider genus on whatever set of dimensions are deemed relevant. This might stimulate the re-imagining of confidence as a flexible indicator of the cognitive representation of particular concepts or context which may not, and arguably should not, necessitate a universal computational approach.

Even with just two dimensions being used to define category membership, it would be informative to explore how the dimensions are weighted, and whether there are individual differences in perception and ranking of those dimensions. For example, colour was found to be more salient in Chapter 5 relative to size. Leaning into this differentiation, directly modelling how the individual subjective category boundary might be pushed or pulled relative to the objective bound could shed light on the origin of the individual differences in model fits or confidence usage. The relevance of understanding confidence in these kinds of spaces becomes apparent when again looking at the world of medical diagnoses. In the case of our dermatologist classifying skin lesions perhaps here too, like in Studies 4 and 5, colour is a more salient cue for malignancy than size. In the case of a medical practitioner trained in the UK, for example, how then might confidence be expressed in cases where patients present with lesions or growths on non-white skin where discolouration is less apparent? Overconfidence in the benign nature of a lesion due to over-reliance on the salient colour cue could be fatal. Therefore, understanding how the behaviour of confidence changes

relative to a ranking of preferred or more salient category properties would allow for more targeted interventions to be developed to improve the metacognitive calibration of decisions in the medical field and beyond.

6.5 | Conclusions

This thesis explored two adjacent fields of cognition, decision-making and categorisation, bridging the two by leveraging models and self-reports of metacognitive decision confidence. By examining patterns and model comparisons of decision confidence, I investigated the way that a specific type of decision, a categorisation, is navigated by decision-makers. The potential for increasing understanding of both decision confidence and concept representation by integrating the ideas and methods from these two fields of literature is clear. This thesis acts as an exploration into this integration and while yielding interesting and distinctive results it serves more clearly to stimulate discussion on the potential for this rich experimental space.

The sensitivity of confidence to complex features of tasks in category spaces, and their representations that give rise to decisions, was demonstrated throughout. This was seen most clearly in the ability to read out properties of these representations in the confidence reports of individuals. The scope of these categorical spaces and the potential for further research are evident in the many variations of category dimensionality and boundedness that could be leveraged. The exploration of these spaces lends itself to the idea of the contextualisation of cognitive approaches to specific task or problem structures. The variety of approaches that can be made by the observer in terms of distribution representation and task space navigation are apparent in a more complex categorical stimulus space. Indeed, across the multi-dimensional task space used, it seems

that not all confidence reports were created equal and some reflected single versus dual-dimension classification strategies.

Therefore, based on the work in this thesis, I believe that by using the lens of decision confidence to examine the "glue that holds our mental world together." (Murphy, 2002, p. 1), concepts and categorisation, we can shed light on the way in which individuals navigate and explore their worlds, process evidence and information, and ultimately make contextualised decisions and category judgements.

Bibliography

Acerbi, L., Dokka, K., Angelaki, D. E. & Ma, W. J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLOS Computational Biology*, *14*(7), e1006110. https://doi.org/10.1371/journal.pcbi.1006110

Acerbi, L. & Ma, W. J. (2017). Practical bayesian optimization for model fitting with bayesian adaptive direct search. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1834–1844.

Ackerman, R. & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Adler & Ma. (2017). *Limitations of proposed signatures of bayesian confidence* (preprint). Neuroscience. https://doi.org/10.1101/218222

Adler & Ma. (2018). Comparing bayesian and non-bayesian accounts of human confidence reports. *PLOS Computational Biology*, *14*(11), e1006572. https://doi.org/10.1371/journal.pcbi.1006572

Adler & Pointon. (1993). *Vision: Coding and efficiency*. Cambridge University Press.

Ais, J., Zylberberg, A., Barttfeld, P. & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386. https://doi.org/10.1016/j.cognition.2015.10.006

Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. (2015). Doubly bayesian analysis of confidence in perceptual decision-making. *PLOS Computational Biology*, *11*(10), e1004519. https://doi.org/10.1371/journal.pcbi.1004519

Alais, D. & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Andrew, S. C., Perry, C. J., Barron, A. B., Berthon, K., Peralta, V. & Cheng, K. (2014). Peak shift in honey bee olfactory learning. *Animal Cognition*, *17*(5), 1177–1186. https://doi.org/10.1007/s10071-014-0750-3

Arnold, D. H., Petrie, K., Murray, C. & Johnston, A. (2019). Suboptimal human multisensory cue combination. *Scientific Reports*, *9*(1), 5155. https://doi.org/10.1038/s41598-018-37888-7

Atir, S., Rosenzweig, E. & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, *26*(8), 1295–1303. https://doi.org/10.1177/0956797615588195

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G. & Frith, C. D. (2010). Optimally interacting minds. *Science (New York, N.Y.)*, *329*(5995), 1081. https://doi.org/10.1126/science.1185718

Balakrishnan, J. D. & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 615. https://doi.org/10.1037/0096-1523.22.3.615

Bang, D. & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082–6087. https://doi.org/10.1073/pnas.1800795115

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y. F., Roepstorff, A., Rees, G., Frith, C. D. & Bahrami, B. (2014). Does interaction matter? testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition: An International Journal*, *26*, 13–23. https://doi.org/10.1016/j.concog.2014.02.002

Bao, S. (2015). Perceptual learning in the developing auditory cortex. *European Journal of Neuroscience*, *41*(5), 718–724. https://doi.org/10.1111/ejn.12826

Baranski, J. V. & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 929. https://doi.org/10.1037/0096-1523.24.3.929

Baranski, J. V. & Petrusic, W. M. (2001). Testing architectures of the decision–confidence relation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *55*(3), 195. https://doi.org/10.1037/h0087366

Barnes, J. (1991). *The complete works of aristotle the revised oxford translation* (Vol. 1). Princeton University Press.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629–654. https://doi.org/10.1037/0278-7393.11.1-4.629

Barthelmé, S. & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(48), 20834–20839. https://doi.org/10.1073/pnas.1007704107

Battaglia, P. W., Jacobs, R. A. & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *JOSA A*, *20*(7), 1391–1397. https://doi.org/10.1364/JOSAA.20.001391

Beierholm, U. R. (2013). Bayesian approaches in computational neuroscience: Overview. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of computational neuroscience* (pp. 1–2). Springer New York. https://doi.org/10.1007/978-1-4614-7320-6_778-2

Beierholm, U. R. (2020). Bayesian approaches in computational neuroscience: Overview. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of computational neuroscience* (pp. 1–2). Springer. https://doi.org/10.1007/978-1-4614-7320-6_778-3

Ben-David, I., Graham, J. R. & Harvey, C. R. (2013). Managerial miscalibration. *The Quarterly Journal of Economics*, *128*(4), 1547–1584. Retrieved August 21, 2023, from https://www.jstor.org/stable/26372532

Bonaccio, S. & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bontempo, R. N., Bottom, W. P. & Weber, E. U. (1997). Cross-cultural differences in risk perception: A model-based approach. *Risk Analysis*, *17*(4), 479–488. https://doi.org/10.1111/j.1539-6924.1997.tb00888.x

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review*, *77*(6), 546–556. https://doi.org/10.1037/h0030000

Bowers, J. S. & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389. https://doi.org/10.1037/a0026450

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

Box, G. E. P. & Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons.

Brenner, L. A., Koehler, D. J., Liberman, V. & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*(3), 212–219. https://doi.org/10.1006/obhd.1996.0021

Butler, J. S., Smith, S. T., Campos, J. L. & Bülthoff, H. H. (2010). Bayesian integration of visual and vestibular signals for heading. *Journal of Vision*, *10*(11), 23. https://doi.org/10.1167/10.11.23

Cabantous, L., Hilton, D., Kunreuther, H. & Michel-Kerjan, E. (2011). Is imprecise knowledge better than conflicting expertise? evidence from insurers' decisions in the united states. *Journal of Risk and Uncertainty*, *42*, 211–232. https://doi.org/10.1007/s11166-011-9117-1

Carlebach, N. & Yeung, N. (2020). Subjective confidence acts as an internal cost-benefit factor when choosing between tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(7), 729.

Carlebach, N. & Yeung, N. (2023). Flexible use of confidence to guide advice requests. *Cognition*, *230*, 105264. https://doi.org/10.1016/j.cognition.2022.105264

Charles, L. & Yeung, N. (2019). Dynamic sources of evidence supporting confidence judgments and error detection. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(1), 39–52. https://doi.org/10.1037/xhp0000583

Cheng, K., Spetch, M. L. & Johnston, M. (1997). Spatial peak shift and generalization in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *23*(4), 469–481. https://doi.org/10.1037/0097-7403.23.4.469

Clayton, D. A., Eguchi, M. M., Kerr, K. F., Miyoshi, K., Brunyé, T. T., Drew, T., Weaver, D. L. & Elmore, J. G. (2022). Are pathologists self-aware of their diagnostic accuracy? metacognition and the diagnostic process in pathology. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*. https://doi.org/10.1177/0272989X221126528

Collins, A. M. & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Davidson, J., Deuser, R. & Sternberg, R. (1994). The role of metacognition in problem solving. https://doi.org/10.7551/mitpress/4561.003.0012

De Martino, B., Kumaran, D., Seymour, B. & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, *313*(5787), 684–687. https://doi.org/10.1126/science.1128356

de Gardelle, V., Corre, F. L. & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *PLOS ONE*, *11*(1), e0147901. https://doi.org/10.1371/journal.pone.0147901

de Gardelle, V. & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science*, *25*(6), 1286–1288. https://doi.org/10.1177/0956797614528956

Desender, K., Boldt, A. & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, *29*(5), 761–778. https://doi.org/10.1177/0956797617744771

Desender, K., Murphy, P., Boldt, A., Verguts, T. & Yeung, N. (2019). A postdecisional neural marker of confidence predicts information-seeking in decision-making. *Journal of Neuroscience*, *39*(17), 3309–3319. https://doi.org/10.1523/JNEUROSCI.2620-18.2019

Dienes, Z. (2007). Subjective measures of unconscious knowledge. In R. Banerjee & B. K. Chakrabarti (Eds.), *Progress in brain research* (pp. 49–269). Elsevier. https://doi.org/10.1016/S0079-6123(07)68005-4

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. https://doi.org/10.1016/j.concog.2009.09.009

Dunlosky, J. & Metcalfe, J. (2008). *Metacognition*. Sage Publications, Inc.

Dunlosky, J. & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*(4), 374–380. https://doi.org/10.3758/BF03210921

Edwards, W. (1955). The theory of decision making. *Psychological Bulletin*, *51*(4), 380. https://doi.org/10.1037/h0053870

Ernst, M. O. & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Farrell, S. & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Fific, M., Little, D. R. & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review*, *117*, 309–348. https://doi.org/10.1037/a0018526

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906. https://doi.org/10.1037/0003-066X.34.10.906

Fleming, S. M. & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M. & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. Retrieved October 14, 2022, from https://www.frontiersin.org/articles/10.3389/fnhum.2014.00443

Folke, T., Jacobsen, C., Fleming, S. M. & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), 1–8. https://doi.org/10.1038/s41562-016-0002

Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., Fine, P. L., Miller, T. M. & Elstein, A. S. (2005). Do physicians know when their diagnoses are correct? implications for decision support and error reduction. *Journal of General Internal Medicine*, *20*(4), 334–339. https://doi.org/10.1111/j.1525-1497.2005.30145.x

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. https://doi.org/10.3758/BF03196546

Geisler, W. S. & Ringach, D. (2009). Natural systems analysis. *Visual Neuroscience*, *26*(1), 1–3. https://doi.org/10.1017/S0952523808081005

Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1992). Probabilistic mental models: A brunswikian theory of confidence. *Psychological Review*, *98*(4), 506. https://doi.org/10.1037/0033-295X.98.4.506

Gold, J. I. & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Wiley.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Guggenmos, M. (2022). Reverse engineering of metacognition (V. Wyart, M. J. Frank & S. Fleming, Eds.). *eLife*, *11*, e75420. https://doi.org/10.7554/eLife.75420

Hampton, J. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *18*, 441–461.

Hangya, B., Sanders, J. I. & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation*, *28*(9), 1840–1858. https://doi.org/10.1162/NECO_a_00864

Hasan, E., Eichbaum, Q., Seegmiller, A. C., Stratton, C. & Trueblood, J. S. (2023). Harnessing the wisdom of the confident crowd in medical image decision-making. *Decision*. https://doi.org/10.1037/dec0000210

Helmholtz, H. v. (1867). *Handbuch der physiologischen Optik*. Voss.

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*, 186–201. https://doi.org/10.1037/h0074579

Hershberger, W. (1971). Attached-shadow orientation perceived as depth by chickens reared in an environment illuminated from below. *Journal of*

Comparative and Physiological Psychology, *73*(3), 407. https://doi.org/10.1037/h0030223

Howard, I. P., Bergström, S. S. & Ohmi, M. (1990). Shape from shading in different frames of reference. *Perception*, *19*(4), 523–530. https://doi.org/10.1068/p190523

Hubel, D. H. & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591. Retrieved February 23, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363130/

Hull, C. L. (1920). Quantitative aspects of evolution of concepts: An experimental study. *Psychological Monographs*, *28*(1), i. https://doi.org/10.1037/h0093130

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Iyer, E. S., Weinberg, A. & Bagot, R. C. (2022). Ambiguity and conflict: Dissecting uncertainty in decision-making. *Behavioral Neuroscience*, *136*(1), 1–12. https://doi.org/10.1037/bne0000489

Jackson, S. A. & Kleitman, S. (2014). Individual differences in decision-making and confidence: Capturing decision tendencies in a fictitious medical test. *Metacognition and Learning*, *9*(1), 25–49. https://doi.org/10.1007/s11409-013-9110-y

Jazayeri, M. & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026. https://doi.org/10.1038/nn.2590

Jordano, M. L. & Touron, D. R. (2018). How often are thoughts metacognitive? findings from research on self-regulated learning, think-aloud protocols, and mind-wandering. *Psychonomic Bulletin & Review*, *25*(4), 1269–1286. https://doi.org/10.3758/s13423-018-1490-1

Juslin, P. & Olsson, H. (1997). Thurstonian and brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*(2), 344. https://doi.org/10.1037/0033-295X.104.2.344

Kahneman, D. & Tversky, A. (1974). Subjective probability: A judgment of representativeness. In C.-A. S. Staël Von Holstein (Ed.), *The concept of probability in psychological experiments* (pp. 25–48). Springer Netherlands. https://doi.org/10.1007/978-94-010-2288-0_3

Kepecs, A. & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231. https://doi.org/10.1038/nature07200

Kersten, D., Mamassian, P. & Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Kiani, R. & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., IJzerman, H., Nilsonne, G., Vanpaemel, W. & Frank, M. C. (2018). A practical guide for transparency in psychological science (M. Nuijten & S. Vazire, Eds.). *Collabra: Psychology*, *4*(1), 20. https://doi.org/10.1525/collabra.158

Knill, D. C. & Richards, W. (1996). *Perception as bayesian inference*. Cambridge University Press.

Komori, M. (2016). Effects of working memory capacity on metacognitive monitoring: A study of group differences using a listening span test. *Frontiers in Psychology*, *7*. Retrieved June 29, 2023, from https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00285

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T. & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, *16*(6), 749–755. https://doi.org/10.1038/nn.3393

Körding, K. P. & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247. https://doi.org/10.1038/nature02169

Körding, K. P. & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326. https://doi.org/10.1016/j.tics.2006.05.003

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107. https://doi.org/10.3758/BF03212211

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S. & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21. https://doi.org/10.1111/j.1467-7687.2006.00468.x

Lak, A., Nomoto, K., Keramati, M., Sakagami, M. & Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology*, *27*(6), 821–832. https://doi.org/10.1016/j.cub.2017.02.026

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, *124*(2), 161–180. https://doi.org/10.1037/0096-3445.124.2.161

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 695–711. https://doi.org/10.1037/0278-7393.24.3.695

Lichtenstein, S. & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183. https://doi.org/10.1016/0030-5073(77)90001-0

Little, D. R., Nosofsky, R. M. & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1–27. https://doi.org/10.1037/a0021330

Livesey, E. J. & McLaren, I. P. L. (2009). Discrimination and generalization along a simple dimension: Peak shift and rule-governed responding. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(4), 554–565. https://doi.org/10.1037/a0015524

Lockhead, G. R. (1972). Processing dimensional stimuli: A note. *Psychological Review*, *79*(5), 410. https://doi.org/10.1037/h0033129

Loftus, E. F. (1996). *Eyewitness testimony*. Harvard University Press.

Mamassian, P. & Goutcher, R. (2001). Prior knowledge on the illumination position. *Cognition*, *81*(1), B1–B9. https://doi.org/10.1016/S0010-0277(01)00116-0

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. The MIT Press. Re-

trieved November 15, 2020, from https://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262514620.001.0001/upso-9780262514620

Massoni, S. (2014). Emotion as a boost to metacognition: How worry enhances the quality of confidence. *Consciousness and Cognition*, *29*, 189–198. https://doi.org/10.1016/j.concog.2014.08.006

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines*, *1*(2), 185–196. https://doi.org/10.1007/BF00361036

McCloskey, M. & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*, 462–472. https://doi.org/10.3758/BF03197480

McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 771–792. https://doi.org/10.1037/0278-7393.24.3.771

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mervis, C. B., Catlin, J. & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society*, *7*(3), 283–284. https://doi.org/10.3758/BF03337190

Metcalfe, J. & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Metcalfe, J. & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing.* https://doi.org/10.7551/mitpress/4561.001.0001

Meyers, J., Lytle, S., Palladino, D., Devenpeck, G. & Green, M. (1990). Think-aloud protocol analysis: An investigation of reading comprehension strategies in fourth-and fifth-grade students. *Journal of Psychoeducational Assessment*, *8*(2), 112–127. https://doi.org/10.1177/073428299000800201

Milton, F., Longmore, C. A. & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(3), 676–692. https://doi.org/10.1037/0096-1523.34.3.676

Movellan, J. R. & Nelson, J. D. (2001). Probabilistic functionalism: A unifying paradigm for the cognitive sciences. *Behavioral and Brain Sciences*, *24*(4), 690–692. https://doi.org/10.1017/S0140525X0154008X

Murphy, G. (2002). *The big book of concepts.* A Bradford Book. Retrieved December 16, 2021, from https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=78121&site=ehost-live&authtype=ip,uid

Navajas, J., Bahrami, B. & Latham, P. E. (2016). Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences*, *11*, 55–60. https://doi.org/10.1016/j.cobeha.2016.05.005

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E. & Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature Human Behaviour*, *1*(11), 810–818. https://doi.org/10.1038/s41562-017-0215-1

Nelson. (1996). Consciousness and metacognition. *American Psychologist*, *51*(2), 102. https://doi.org/10.1037/0003-066X.51.2.102

Nelson & Narens. (1990). Metamemory: A theoretical framework and new findings. *Psychology of learning and motivation* (pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5

Nosofsky. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114. https://doi.org/10.1037/0278-7393.10.1.104

Nosofsky. (1992). Exemplar-based approach to relating categorization, identification, and recognition. *Multidimensional models of perception and cognition* (pp. 363–393). Lawrence Erlbaum Associates, Inc.

Olawole-Scott, H. & Yon, D. (2023). Expectations about precision bias metacognition and awareness. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001371

O'Reilly, J. X., Jbabdi, S. & Behrens, T. E. J. (2012). How can a bayesian approach inform neuroscience? *European Journal of Neuroscience*, *35*(7), 1169–1179. https://doi.org/10.1111/j.1460-9568.2012.08010.x

Overgaard, M. & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1287–1296. https://doi.org/10.1098/rstb.2011.0425

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L. & Roberts, R. D. (2002). The role of individual differences in the

accuracy of confidence judgments. *The Journal of General Psychology*, *129*(3), 257–299. https://doi.org/10.1080/00221300209602099

Payne, S. J. & Duggan, G. B. (2011). Giving up problem solving. *Memory & Cognition*, *39*(5), 902–913. https://doi.org/10.3758/s13421-010-0068-6

Peirce, C. S. & Jastrow, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences*, *3*, 73–83.

Persaud, N., McLeod, P. & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*(2), 257–261. https://doi.org/10.1038/nn1840

Piray, P., Dezfouli, A., Heskes, T., Frank, M. J. & Daw, N. D. (2019). Hierarchical bayesian inference for concurrent model fitting and comparison for group studies. *PLOS Computational Biology*, *15*(6), e1007043. https://doi.org/10.1371/journal.pcbi.1007043

Platt, M. L. & Huettel, S. A. (2008). Risky business: The neuroeconomics of decision making under uncertainty. *Nature neuroscience*, *11*(4), 398–403. https://doi.org/10.1038/nn2062

Pleskac, T. J. & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. https://doi.org/10.1037/a0019737

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363. https://doi.org/10.1037/h0025953

Posner, M. I. & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*, 304–308. https://doi.org/10.1037/h0028558

Pouget, A., Drugowitsch, J. & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Pushkarskaya, H., Smithson, M., Joseph, J. E., Corbly, C. & Levy, I. (2015). Neural correlates of decision-making under ambiguity and conflict. *Frontiers in Behavioral Neuroscience*, *9*. Retrieved February 10, 2023, from https://www.frontiersin.org/articles/10.3389/fnbeh.2015.00325

Qamar, A. T., Cotton, R. J., George, R. G., Beck, J. M., Prezhdo, E., Laudano, A., Tolias, A. S. & Ma, W. J. (2013). Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(50), 20332–20337. https://doi.org/10.1073/pnas.1219756110

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, *12*(5), 410–430. https://doi.org/10.1002/bs.3830120511

Rahnev, D. (2023). Measuring metacognition: A comprehensive assessment of current methods. https://doi.org/10.31234/osf.io/waz9h

Ratcliff, R. & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333

Resulaj, A., Kiani, R., Wolpert, D. M. & Shadlen, M. N. (2009). Changing your mind: A computational mechanism of vacillation. *Nature*, *461*(7261), 263–266. https://doi.org/10.1038/nature08275

Rhodes, M. G. (2019). Metacognition. *Teaching of Psychology*, *46*(2), 168–175. https://doi.org/10.1177/0098628319834381

Rogers, T. T. & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press.

Rosch. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and acquisition of language* (pp. 111–144). Academic Press. https://doi.org/10.1016/B978-0-12-505850-6.50010-4

Rosch & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rouault, M., Dayan, P. & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, *10*(1), 1141. https://doi.org/10.1038/s41467-019-09075-3

Rouault, M., Seow, T., Gillan, C. M. & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological Psychiatry*, *84*(6), 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017

Sanborn, A. N., Griffiths, T. L. & Shiffrin, R. M. (2010). Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, *60*(2), 63–106. https://doi.org/10.1016/j.cogpsych.2009.07.001

Sanders, J. I., Hangya, B. & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron*, *90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

Schellings, G. L. M., van Hout-Wolters, B. H. A. M., Veenman, M. V. J. & Meijer, J. (2013). Assessing metacognitive activities: The in-depth comparison of a task-specific questionnaire with think-aloud protocols. *European Journal of Psychology of Education*, *28*(3), 963–990. https://doi.org/10.1007/s10212-012-0149-y

Shadlen, M. N. & Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Current Opinion in Neurobiology*, *4*(4), 569–579. https://doi.org/10.1016/0959-4388(94)90059-0

Shekhar, M. & Rahnev, D. (2021). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12–23. https://doi.org/10.1016/j.tics.2020.10.007

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*(1), 54–87. https://doi.org/10.1016/0022-2496(64)90017-3

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows

presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simon, H. A. (1990). Bounded rationality. In J. Eatwell, M. Milgate & P. Newman (Eds.), *Utility and probability* (pp. 15–18). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-20568-4_5

Smith & Medin, D. L. (1981). *Categories and concepts*. Harvard University Press.

Smith & Ratcliff. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Smith, Shoben & Rips. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*, 214–241. https://doi.org/10.1037/h0036351

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Strand, J. F. (2023). Error tight: Exercises for lab groups to prevent research mistakes. *Psychological Methods*. https://doi.org/10.1037/met0000547

Summerfield, C. & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *Frontiers in Neuroscience*, *6*, 70. https://doi.org/10.3389/fnins.2012.00070

Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Tunney, R. J. (2005). Sources of confidence judgments in implicit cognition. *Psychonomic Bulletin & Review*, *12*(2), 367–373. https://doi.org/10.3758/BF03196386

Tunney, R. J. & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition*, *31*, 1060–1071. https://doi.org/10.3758/BF03196127

Tversky & Kahneman. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky & Kahneman. (1986). Rational choice and the framing of decisions. *The Journal of Business*, *59*(4), S251–S278. Retrieved October 8, 2022, from https://www.jstor.org/stable/2352759

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N. & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence (T. E. Behrens, Ed.). *eLife*, *5*, e12192. https://doi.org/10.7554/eLife.12192

van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N. & Wolpert, D. M. (2016). Confidence is the bridge between multi-stage decisions. *Current Biology*, *26*(23), 3157–3168. https://doi.org/10.1016/j.cub.2016.10.021

van Opheusden, B., Acerbi, L. & Ma, W. J. (2020). Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *arXiv:2001.03985 [cs, q-bio, stat]*. Retrieved November 15, 2020, from http://arxiv.org/abs/2001.03985

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*(1), 37–58. https://doi.org/10.1080/00140137008931117

Vickers, D. & Packer, J. (1982). Effects of alternating set for speech or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, *50*(2), 179–197. https://doi.org/10.1016/0001-6918(82)90006-3

Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(1), 103–112. https://doi.org/10.1037/0096-1523.9.1.103

Weber, E. U., Blais, A.-R. & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290. https://doi.org/10.1002/bdm.414

Wills, Inkster, A. B. & Milton, F. (2015). Combination or differentiation? two theories of processing order in classification. *Cognitive Psychology*, *80*, 1–33. https://doi.org/10.1016/j.cogpsych.2015.04.002

Wills & Mackintosh, N. J. (1998). Peak shift on an artificial dimension, 32.

Wilson, R. A. & Keil, F. (1999). *The MIT encyclopedia of the cognitive sciences*. Retrieved October 7, 2022, from https://ezproxy-prd.bodleian.ox.ac.uk:11190/ehost/ebookviewer/ebook/bmxlYmtfXzg2MDU4X19BTg2?sid=78ed1984-8f79-415c-af68-edb4c0b6de66@redis&vid=0&format=EB&lpid=lp_ix&rid=0

Wittgenstein, L. (1953). *Philosophical investigations*. John Wiley & Sons.

Xue, K., Shekhar, M. & Rahnev, D. (2023, September 4). Challenging the bayesian confidence hypothesis. https://doi.org/10.31234/osf.io/mf5zp

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Yates, F. A. (2011). *The art of memory*. Random House.

Yee, E. & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, *23*(4), 1015–1027. https://doi.org/10.3758/s13423-015-0948-7

Yeung, N. & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Yeung, N. & Summerfield, C. (2014). Shared mechanisms for confidence judgements and error detection in human decision making. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 147–167). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-45190-4_7

Yu, S., Pleskac, T. J. & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology: General*, *144* (2), 489–510. https://doi.org/10.1037/xge0000062

APPENDIX A

Appendices

A.1 | Study 2: Confidence Simulations

In order to maximise the power of the experimental design, simulations of experimental trials were run to simulate confidence reports based on likelihoods at different values of distribution means and variances. These simulations aimed to find the values for the narrow and wide distributions that would provide maximally different objective confidence reports moving away from the mid-line of the category distributions.

A.1.1 | Simulation Constraints

The simulation values were constrained by the size of the experimental screen, which spanned 800 pixels. Distributions that would largely draw trials outside this range would not be appropriate for this study as the required re-sampling rate would skew the distributions away from their actuality. Therefore, distributions whose locations up to 2 standard deviations from the mean fell within the screen dimensions were selected as candidates for comparison. This meant that the variances and the mean locations were both constrained accordingly.

A.1.2 | Defining the Category Boundary

The category mid-line for each category pair was defined mathematically by finding the unique root of both distributions between the two distribution means. In other words, the mid-line was located by searching that space for the point at which the distributions cross over, where the likelihood of each distribution

occurring is exactly 0.5. These mid-lines were then used to compare confidence differences on either side of the cross-over point in order to find maximal confidence differences.

A.1.3 | Simulation Method

Initially, the values used in the pilot study were used to establish a baseline of confidence differences. This created a wide distribution with a mean of 500 pixels and a variance of 100 pixels and a narrow distribution with a mean of 200 pixels and a variance of 70 pixels. These values were simulated over N = 10,000 trials each, with locations drawn randomly from each distribution. Samples which fell outside the 0-800 range were dropped without replacement. As the values of the distributions had been constrained by the screen size, the number of trials that were dropped without replacement was not substantial. The mathematical cross-over point was defined at 332 pixels - this deviated from the previously defined mid-line of 350 in the pilot experiment where the mid-line was defined in geometric space without the true likelihoods accounted for.

In 5 pixel bins across the full 800 pixel space (160 bins in total) the objective likelihood of each location was defined. The simulation then proceeded by allocating the distribution for each bin in line with the most likely distribution for that bin and setting the objective confidence accordingly. The bin which contained the mid-line in that iteration was then identified and tagged. Finally, the differences between confidence reports for each distribution at each bin step

away from the mid-line bin were calculated in the central location values of primary interest between the means and 100 pixels on either side of the means (i.e. the 100-600 pixel space in this instance). This indicated that the original experimental distributions had a mean difference of just 2.7(on a scale from 0-100) across this space with the majority of confidence bin differences falling between 0-5 see Fig. A.1.

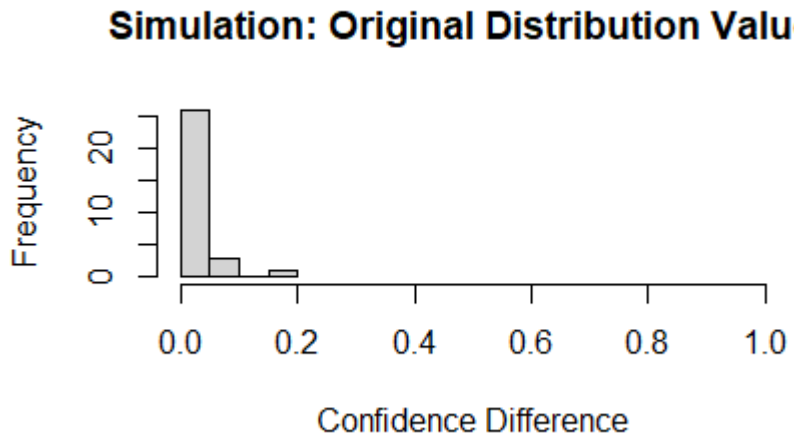


Figure A.1: Histogram showing the range of differences in confidence simulated using the original experimental values for the wide and narrow distributions. Confidence difference is reported on the x-axis and is scaled between 0-1. These values translate to confidence differences on a 0-100 scale such that a difference of 0.05 reported here equates to a difference of 5 on the 0-100 scale. Note that the majority of confidence differences lie below the 0.05 point for the original values utilised in the pilot study.

Given the assumed metacognitive noise and allowing for errors and lapses both in decision-making and confidence reports, these values are very low. As we are looking to find differences, if they are there to be found, between confidence reports for a narrow versus wide distribution - it is important to create these two

distributions using values that would illicit a noticeable difference in confidence reports if participants were utilising this source of information appropriately.

Therefore, the simulations reported here moved beyond the original pilot values and iterated through different possible standard deviation pairs for the wide and narrow distributions to detect the pair of variances that elicited the maximum average objective confidence difference across the central location bin. Once again, the variances were limited by the necessity to keep the majority of the locations within the 800pixel screen. Therefore, simulated narrow distribution standard deviations ranged from 10 to 80 in distance steps of 10, and simulated wide distribution standard deviations ranged from 80-150 in distance steps of 10. The simulations assessed each possible pair of variances and their associated objective confidence differences.

A.1.4 | Simulation Results

The simulations indicated that the pair of distributions with the narrow distribution at a standard deviation of 80, and the wide distribution with a standard deviation of 140 was the optimal set of values to maximise confidence difference in the specified area of interest. Concurrent evaluations using the median confidence difference gave the same results. At these values, the average confidence difference in the location bins stepping away from the mid-line rose to 8.7 with most values lying between 5-10 and some even reaching above 20, see Fig. A.2.

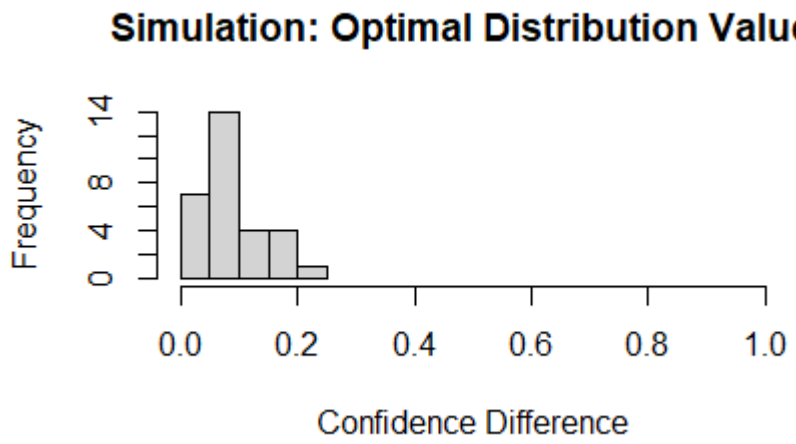


Figure A.2: Histogram showing the range of differences in confidence simulated using the optimal experimental values for the wide and narrow distributions. These values translate to confidence differences on a 0-100 scale such that a difference of 0.05 reported here equates to a difference of 5 on the 0-100 scale. Note that the majority of confidence differences now lie above the 0.05 point for the value pair chosen for the upcoming studies, a marked change from the pilot study simulations.

With these standard deviation values allocated to the wide and narrow distributions respectively, the means and mid-lines were reevaluated. By centring the mid-line on 350 pixels and using these values for the distribution standard deviations the simulation-based calculations indicated that the mean for the narrow distribution should be 222 and the wide 522, (offset by +22pixels). When the wide distribution was flipped to sit on the left of the screen, to give more space for the wide distribution, the mid-line was re-centred on 450 pixels. This required means of 278 for wide and 578 for narrow respectively, (offset by +100-22pixels). These values were selected as the objectively optimal values to maximise the study design power to detect differences in confidence reports if

they were there to be seen. These values were used in the experimental design described below to create distributions with wide and narrow spreads within the constraints of the screen size.

A.2 | Study 2: Response Analysis

The response patterns of participants were also explored. The responses were sensitive both to the distribution spreads and the associated distribution outcome - either bombs or coins. When the outcome is a gain, of coins, and requires a response of "Retrieve" to achieve a correct response the pattern of responses is different than for the same distribution when paired with a loss outcome. This is well described in Fig. A.3 which indicates a shift in response pattern for blocks that required a "Retrieve" response for the wide distribution, as compared with blocks requiring a "Zap" response for the wide distribution trials. In the wide "Zap" blocks, where incorrectly retrieving a wide trial would result in a loss of coins, the participants were relatively conservative in their responses. They were more likely than expected to respond as though the middle bin trials (location bin 6 in Fig. A.3) were drawn from the wide distribution. This is a sensible strategy as incorrectly zapping a coin package resulted in no score change as compared with accidentally retrieving a bomb package which would result in a loss of 3 coins. By comparison, when the wide response was a "Retrieve" action for coin collection, an incorrect retrieval would result in coin loss. Here, trials in the middle bin were more likely to be zapped than retrieved - again

showing a sensitivity to the reward matrix that is layered over the distribution spreads themselves. In essence, the reward matrix, while balanced overall and across blocks and individuals, may have distorted the subjective perception of the distribution spreads pulling the narrow distribution wider to create a more conservative approach to those middle bin trials, see Fig. A.3 for details.

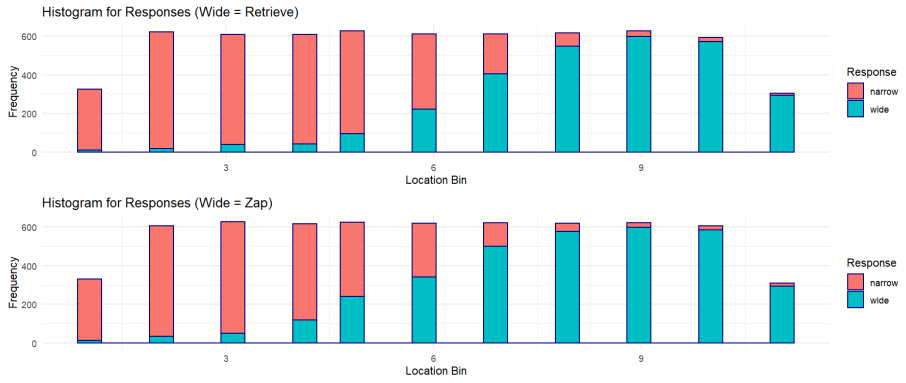


Figure A.3: Response-based analysis of responses to the wide and narrow distribution split by the associated response outcome. The x-axis displays the Location Bin and indicates the total number of responses given in each bin overall. The objectively defined decision boundary lies in bin 6 with the responses transformed such that all trials where participants responded as though it was the narrow distribution lie on the left of bin 6, and all trials where participants responded as though it was the wide distribution lie on the right of bin 6. The top panel indicates the data drawn from blocks where the wide distribution was paired with a positive outcome of coins, requiring a "Retrieve" response. The lower panel plots data drawn from blocks where the wide distribution was associated with a negative outcome, requiring a "Zap" response.

"I know a person small -

She keeps ten million serving-men,

Who get no rest at all!

She keeps 'em abroad on her own affairs,

From the second she opens her eyes -

One million Hows, two million Wheres,

And seven million Whys."

RUDYARD KIPLING