Investigating Individual Differences in Human Decision-Making

An Analysis of Hierarchical Bayesian Models on Bandit Problems

Patrick Laverty

THE UNIVERSITY *of* EDINBURGH

MInf Project (Part 1) Report

Master of Informatics

School of Informatics

University of Edinburgh

2019

Acknowledgements

First and foremost, I would like to say thanks to my supervisor, Chris Lucas, for his constant advice and guidance, and for selflessly donating his time to aid in debugging, as well as for providing code for the optimal Bayesian decision-making model which kick-started my entire analysis; this project would not have been completed without his help.

I would also like to say thank you Michael Lee for providing me with the data upon which my entire analysis is based, without it this project would not have been possible. I would also like to thank him for being so generous with his time and advice, and for sanity checking my code even though it was not his duty.

I also extend thanks to Pablo, Nicolas, and Vlad for allowing me to bounce my ideas off of them, and for offering their own ideas in return.

Thank you, also, to Craig Innes for providing me with an introduction to the Eddie compute cluster, without which many of the results presented in this report would not have been possible.

To Katherine and Nathaniel, thank you for sacrificing our Game of Thrones time to help me spot sentences in this report which simply did not make sense. Your watch has ended.

Finally, I would like to thank my friends for keeping me sane these past few months,

offering both guidance and distraction, and knowing which one I needed when I didn't know myself.

Abstract

Humans must regularly make decisions in unknown environments. In such situations, there is a tradeoff between repeating actions with a known outcomes, and choosing an unknown alternative in order to gain more information about the environment; this is known as the exploration-exploitation dilemma. The tradeoff has been studied through the use of bandit tasks, and strategies for acting optimally in such tasks have been developed. One set of solutions of particular interest to cognitive scientists is the use of psychologically inspired heuristic models. The parameters of such models allow for inferences to be drawn about how humans behave in sequential decision-making tasks, and allow for studies into individual differences in human performance on such tasks to be conducted. Hierarchical Bayesian models have been applied to human data gathered through completion of bandit tasks in order investigate both model and parameter differences. This project applies previously developed heuristic models to a data set consisting of decisions made on bandit tasks by 451 participants in order to test findings which claim that such models give good accounts of human decision-making. Conclusions are also drawn on the ability of those models to capture individual differences in the data set. In addition to this analysis, a new hierarchical Bayesian model, which is a mixture of previously developed cognitive models, is detailed and tested on the data set,

with results indicating that this extended mixture model is able to give a better account of human decision-making data than simpler mixture models. Finally, two mixture models are extended to allow for latent switching between models used on a per-game and per-trial basis. The posterior predictive agreement of data from these models with the data set of 451 participants is calculated and shows that a hierarchical Bayesian model with seven component models that allows for latent switching between models on a per-trial basis gives the best account of both human and optimal data than any other model considered in the analysis.

Contents

Chapter 1

Introduction

Consider the following scenario. You have decided to take a two-week vacation to a city you have never visited before, and that you do not know much about. The hotel you are staying at does not offer any self-catering facilities, and as such you have to dine out every night. You will not be satisfied, of course, with having a series of sub-par meals, so you set yourself a goal for the upcoming fortnight: you will select a series of restaurants which you hope will maximise the number of good meals you have. On the first night of your stay, you leave the hotel to find a suitable restaurant, and one question naturally arises: "Where should I eat?". You do not know anything about the restaurants in the city, so how could you possibly know which ones are good, and which are bad? You could open up a crowd sourcing app such as Yelp to help you make a decision, however, you decide that part of the fun of travelling is in the exploring. Therefore, you decide to choose a restaurant based on certain criteria: perhaps you choose the one closest to your hotel, perhaps you choose based on the average price of a meal - assuming you have such information - or perhaps you choose completely at random. After choosing a restaurant and

dining there, you now have some information you previously did not, and can make future decisions with this information in mind: for example, if the restaurant you attended served good food at a reasonable price, you might want to return there; if, however, you fell ill with food poisoning then you are unlikely to want to return. This querying of a source (the restaurants in the city) in order to gain information (whether a particular restaurant serves good or bad food) to make an informed decision (I will/will not eat at that restaurant in the future) in order to maximise a goal (to have as many good meals as possible) forms the basic structure for tasks within the field of active learning. Over the course of the next week, you decide to explore the restaurants in the city, never visiting the same restaurant twice, and noting the experiences you have at each one. At the end of the week you have a list of restaurants you have visited, as well as whether your experience at each restaurant was good or bad. With this information in mind, and given the fact that you now have one week left to reach your goal of maximising the number of good meals you have, you are presented with another question: "Should I continue to explore new restaurants, with the hope of finding one better than I have visited so far, at the risk of having a bad meal, or should I return to a restaurant I have been to before and had a good meal at?", in essence, "Should I explore, or should I exploit?".

Much work has been conducted on investigating the exploration-exploitation dilemma (Eliassen, Jørgensen, Mangel, & Giske, 2007; Aston-Jones & Cohen, 2005; Sutton & Barto, 1998), with bandit tasks offering a set of problems for conducting such analysis (Steyvers, Lee, & Wagenmakers, 2009; Shen, Wang, Jiang, & Zha, 2015). Bandit tasks have been widely used as they are a simple class of prob-

lems which allow researchers to investigate and develop theories about underlying phenomena regarding human decision-making. Optimal solutions to many bandit problems can be calculated via a variety of methods specific to the task at hand (Burtini, Loeppky, & Lawrence, 2015), and optimal decision-making data can be generated using Bellman equations in a dynamic programming solution (Kaelbling, Littman, & Moore, 1996). Psychological models have also been the subject of extensive study in sequential decision-making tasks (Steyvers et al., 2009; Zhang & Lee, 2010a), as well as other areas of cognitive science such as word learning (Frank, Goodman, & Tenenbaum, 2009) and category learning (Nosofsky, 1986). These models are of interest as they offer heuristic solutions to a variety of problems, thus reducing the computational complexity of calculating solutions. Such models are also of value for cognitive scientists as they allow conclusions to be drawn about human behaviour through analysis of psychologically relevant parameters. Not only is understanding human behaviour useful for developing new and interesting heuristics for solving cognitive tasks, but comparing human performance to optimal data via those psychological models provides a framework for analysing how close to optimal said human performance is. By analysing the values of these parameters after a model is fit on human data, we can identify where, and why, human performance falls short of optimal, providing an opportunity to teach decision-makers and help them improve their decision making strategies (Lee, Zhang, Munro, & Steyvers, 2011). A common practice in the analysis of human performance in cognitive tasks is to first aggregate data across participants before conducting an analysis (Farrell & Lewandowsky, 2018). Analysing data in the aggregate uses fewer computational resources than analysing individual data, and if the performance of subjects is the

same, aggregating can remove noise and reveal the underlying psychological phenomena (Lee & Webb, 2005). However, details on individual performance is lost, and claims made about aggregate performance do not hold when applied to the individual. Attempts have been made to model individual differences in a variety of fields ranging from studies on health (Hu, Zhang, & Wang, 2015), to economics (Pothos, Perry, Corr, Matthew, & Busemeyer, 2011), and the cognitive sciences (Steyvers et al., 2009). Accounting for individual differences allows analyses to show differences in key parameters of a model between all participants, whilst still allowing for group statistics to be calculated.

Some researchers have proposed hard-coding the number of groups before analysis based on prior information (Lee & Webb, 2005), however, others argue that it is not always possible to know the number of groups initially, and that placing a hard prior on the number of groups which can occur limits the model under consideration (Navarro, Griffiths, Steyvers, & Lee, 2006). The methods used vary from field to field, but one emerging trend is the use of hierarchical Bayesian models. In such models, the number of groups is not hard-coded prior to analysis, and instead, priors are placed on key variables within the model. Participants are assumed to belong to the same group if their parameter values have been generated from the same prior distribution. These models are often used in a Bayesian framework so that posterior distributions of participants to groups can be learned by providing a description of the model and fitting it to a set of data. These models have been used to account for within-group differences (Zeigenfuse & Lee, 2009), the notion that participants subscribe to the same general model, but differ in key parameter values. Another

way that participants can differ is by the way in which they approach a problem; this type of difference is commonly referred to as between-group differences (Navarro et al., 2006). If multiple models are available to account for how a participant solves a problem, then a mixture model can be constructed. A mixture model combines all available models, and includes a latent variable which denotes which model a particular participant is using. Models such as these have been used successfully to account for between-group differences (Zeigenfuse & Lee, 2009). Hierarchical Bayesian models also provide a framework for constructing mixture models in order to account for between-group differences, making models capable of capturing both within-group and between-group differences possible. In particular, hierarchical Bayesian models have been applied to modelling of human data on bandit tasks. Zhang and Lee (2010a) took four psychologically inspired heuristics models from the reinforcement learning literature, and built them into a hierarchical Bayesian model. These researchers used their model to analyse the performance of 451 participants who each completed a series of 300 bandit tasks. They found that there was clear evidence of individual differences - both within-group and between-group - within the data. Work has also been conducted into developing new models which describe a sequential decision-maker. One such model, dubbed the τ-switch model (Lee, Zhang, Munro, & Steyvers, 2009), performed better than any other psychological model proposed: in an analysis involving ten participants who each completed 300 bandit tasks, the τ-switch model outperformed all other psychological heuristics considered (Lee et al., 2009).

However, the τ-switch model was not included in the hierarchical Bayesian model

analysed by Zhang and Lee (2010a), nor was it applied to the larger data set gathered by Steyvers et al. (2009)[1] to test for any meaningful individual differences in parameters of the model when applied to human data. Since the data set gathered by Steyvers et al. (2009) demonstrates individual differences, it is useful to apply a new model to it to test whether or not it is able capture those differences, or perhaps, account for them in a new and meaningful way. One reason for this, as discussed by Lee et al. (2009), is that the τ-switch model was designed to be applied to two-armed bandit tasks, whereas the data used by Steyvers et al. (2009) was gathered from four-armed bandit tasks.

A goal of this project was, therefore, to extend the τ-switch model so it can be applied to N-armed bandit data, and to apply the model to this data set. The purposed of this analysis was to determine whether or not the claim still holds that it gives a better account of human performance on bandit tasks than the previously developed models, as well as to study the ability of the model to capture individual differences. Steyvers et al. (2009) also propose a Bayesian framework for analysing models by calculating Bayes Factors comparing the marginal likelihood of competing models. This could be a useful tool for analysing models and, therefore, another aim of this project was to reproduce the findings of that paper to confirm the usefulness of the proposed framework and the results pertaining to the individual differences in the data set. Without determining the reliability of the data set, any further analysis involving the data would be meaningless.

Another goal of this project was to extend the mixture model developed by Zhang and Lee (2010a) to develop a new hierarchical Bayesian model which includes the

[1]This was confirmed after reaching out to a researcher involved in both studies, Michael Lee

τ-switch model, and various other psychological heuristics which have thus far been excluded from a hierarchical Bayesian analysis. This model was then applied to the data set of 451 participants in order to determine whether it gave a better account of human performance on bandit tasks than the mixture model developed by Zhang and Lee (2010a). Of course, as was the case before extending the work of Steyvers et al. (2009), it makes sense to conduct a replication of the work conducted by Zhang and Lee (2010a) in order to test the reliability of their hierarchical Bayesian model, before additional components are added, and so, such an analysis was also conducted. Finally, two variations on the hierarchical Bayesian models discussed so far are proposed. These models have psychological underpinnings, and increase the extent to which latent states can be used in hierarchical Bayesian models to allow for a more flexible and complete account of human decision making to be accounted for. An analysis was carried out on these models in order to test their ability to account for human decision-making in bandit tasks.

In summary, the goals of this project were to:

1. Replicate the work of Steyvers et al. (2009) in order to test the proposed framework for model analysis and confirm findings relating to individual differences in data gathered during that project.

2. Extend the τ-switch model to work with an arbitrary number of arms so that it can be applied to any data set.

3. Apply the τ-switch model proposed by Zhang and Lee (2010a) to the data gathered by Steyvers et al. (2009) to test the claim that the τ-switch model gives a good account of human decision-making data, and to test the ability of the model to account for individual differences in human performance.

4. Test the findings of Zhang and Lee (2010a) relating to the hierarchical Bayesian model developed, before extending it to encompass a larger set of cognitive models.

5. Develop a new hierarchical Bayesian model which is a mixture of seven cognitive models - each analysed independently in previous bandit tasks - and to analyse parameter and model differences after applying it to human and optimal data, as well as testing its ability to describe human behaviour on bandit tasks.

6. Develop two extensions to the aforementioned hierarchical Bayesian models to allow for a more robust description of human decision-making behaviour, and to test the agreement of these models with human and optimal decision-making data.

Structure of the paper

The structure of this paper is as follows: in Chapter 2, I begin by giving an overview of the areas of research related to this paper. Chapter 3 describes the models which are subject to later analysis, as well the new models which were developed as part of this project, including: the N-arm extension of the τ-switch model; the extension to the hierarchical Bayesian model developed by Zhang and Lee (2010a); and two extensions to the aforementioned mixture models. Chapter 4 details the methods and results of the two replication studies that were conducted. In Chapter 5, the analysis of the τ-switch model on the data gathered by Steyvers et al. (2009) is described, as well as the application of the newly developed hierarchical Bayesian model, and extended mixture models, to human and optimal data. Finally, Chapter

6 summarises the findings of this project, and Chapter 7 discusses possible avenues for further research to be carried out during the second year of this project.

Chapter 2

Background and Literature Review

2.1 Exploration-Exploitation Tradeoff

The exploration-exploitation tradeoff arises regularly in our day-to-day lives: should you visit your favourite coffee shop, or try a new one; should you run along your regular route, or venture to a new area; should you stick with your current job, or go back on the market? The tradeoff also appears at a larger scale in many industries: when developing a new drug, medical researchers must decide at which point they should stop testing (exploration) and decide whether or not to bring the drug to market (exploitation). In this domain there is a lot at stake: each additional clinical trial costs time and money, and the researchers may face pressure from shareholders to develop the drug as quickly and cost-efficiently as possible, however, failure to thoroughly research the safety and efficacy of the drug may come at a detriment to the public if the drug is pushed to markets too soon.

Much work has been conducted, across many fields, into examining the nature of the exploration-exploitation tradeoff: Eliassen et al. (2007) examine the tradeoff through the lens of animal foraging; Aston-Jones and Cohen (2005) explore the effect of specific neuromodulatory mechanisms on decision making in human and nonhuman primates; and Sutton and Barto (1998) provide an extensive overview of the problems the tradeoff presents, and present strategies to manage it, in the field of reinforcement learning. Results from these studies have often provided novel insights into the problem, and inspired further research in those fields.

Recent studies have claimed that there is need for synthesis between interdisciplinary fields (Mehlhorn et al., 2015), and that seemingly disparate fields are using different terminology to explore the same underlying phenomenon (Hills et al., 2015). Hills et al. (2015) argue that a developing understanding of the neural structures deployed in various cognitive tasks has lead to the "compelling conclusion that the same cognitive and neural processes underlie much of human behavior involving cognitive search – in both external and internal environments", and in an earlier study, Hills (2006) claim that "molecular machinery that initially evolved for the control of foraging and goal-directed behavior was co-opted over evolutionary time to modulate the control of goal-directed cognition. What was once foraging in a physical space for tangible resources became, over evolutionary time, foraging in cognitive space for information related to those resources."

Whilst these researchers believe that the same mechanisms are at play in all human decision making tasks, others are more skeptical of this claim. Cohen, McClure, and Yu (2007) agree that results from different fields are revealing a set of underlying mechanisms that animals may be using to mange the exploration-exploitation

tradeoff, however, they also highlight the importance of considering other factors such as social signals, levels of abstraction, and various time-scales. Cohen et al. (2007) conclude that "[i]t is not yet clear whether neuromodulatory mechanisms serve the same function at all [levels and timescales], or whether this relies on other mechanisms that remain to be discovered". Mehlhorn et al. (2015) also support the claim that the tradeoff should be considered via a hierarchical structure encompassing different levels of abstraction, with different environmental and social factors coming into play at various levels.

In the largest such study of its kind, Todd, Hills, and Robbins (2012) gathered 44 scientists from various backgrounds to discuss findings relating to cognitive search across 4 fields: the study of animal behaviour, neurobiology, psychology, and computer science. The scientists were divided into 4 groups - one group per field - and tasked with exploring parallels in reported results, and urged to draw connections between the fields. Discussion between groups once again highlighted that the separate fields have many similarities, and the findings from one group can, and should, be utilised by others to arrive at a more comprehensive and coherent understanding of cognitive search. Todd et al. (2012) urge that "Further interdisciplinary cross-fertilization and scientific inquiry will increase our knowledge of the foundations of cognitive search, which will in turn find use in a variety of new applications." The authors highlight some promising avenues of future research such applying findings to aid in developing systems to aid decision-makers in managing the tradeoff; whilst other avenues take a purely research based approach, among which the study of "individual differences in search behavior, their genetic bases, and the possible adaptive nature of mixed strategies" is highlighted. It is upon this topic that a majority of my

research and analysis have been conducted, and remaining sections of this chapter will explore work that has been carried out in this field, including the tasks which have been used to formally investigate the exploration-exploitation tradeoff, as well as models which have been employed to formalise and better understand human behaviour in such tasks. One set of tasks readily employed to study the tradeoff are bandit problems, which are outlined next.

2.2 Multi-Armed Bandits

In bandit problems, an agent is presented with a number of options known as bandits[1], and are told to "play" the bandits with the goal of maximising their reward over some period of time or number of trials. In the classic set-up, an agent selects one bandit per time step and receives a reward according to a Bernoulli process, parameterized by some payout rate, θ, for each arm, which is unknown to the agent at the beginning of the task. The payout rate differs for each arm and it is expected that the agent will generate an estimate of each arms payout rate through repeated plays.

Many variants of this original structure exist, each motivated by the desire to capture more complex and interesting real-world phenomena, for example, bandits with an infinite number of arms have been studied in order to devise solutions the problem of controlling data routing networks (Hung, 2012). Burtini et al. (2015) give an extensive overview of the various types of bandits, and why they are worth studying, presented through the domain of clinical research. Among those described are: bandits with delayed rewards; bandits where contextual information is used;

[1]Named for their similarity to slot-machines, commonly referred to as bandits.

and non-stationary bandits with changing reward rates, exampled by clinical trials where results from a drug aren't available until a number of weeks has passed; interaction between the drug and a patient is affected by the patients demographics; and changing effectiveness of the drug on a patient over time due to evolving environmental factors, respectively.

It is important here to highlight these variants, as different experimental setups have different optimal solutions, and models applied to one task are not appropriate for study in another. I should therefore clarify at this stage that the bandits which are the subject of my study and analysis are 4-armed Bernoulli bandits with a finite horizon and a stationary reward policy, chosen both for their simplicity of analysis and availability of data.

Due to their ability to model the exploration-exploitation tradeoff in many real-world situations, bandit problems have been the subject of study in fields ranging from psychology (Steyvers et al., 2009), to reinforcement learning (Sutton & Barto, 1998), and economics (Shen et al., 2015). As a result, a wide variety of solutions to various types of bandit problems have been proposed. For the sake of brevity, I will not go into detail on specific solutions, and will instead point to the overview presented by Berry and Fristedt (1985), as well as the more recent and thorough survey conducted by Burtini et al. (2015). I would be remiss, however, to speak of the solutions to bandit problems without even a brief mention of the seminal work in calculating optimal strategies for bandit problems conducted by Gittins (1979).

In this paper, Gittins ascribes to each arm an index based on the discount factors considered, the number of times an arm has been chosen, and the successes of that arm. The index is retrieved by searching a table of published "index values" for

various discount factors. The index value represents both the expected value of choosing a certain arm based on its history, and the value of the information to be gained by choosing that arm. Gittins has shown that an optimal balance between exploration and exploitation can be attained by choosing, at each time step, the arm with the largest index value.

One limitation of the Gittins index is that it can only be used under specific assumptions, namely: bandits must use an infinite time horizon with discounted reward rates; arms which are not chosen must not change state; and the rewards of all arms must be generated independently (Gittins, 1979). Many real world problems do not abide by these assumptions, and therefore the Gittins index can not be used to act optimally, although the indices have been used as a heuristic under relaxed assumptions (Glazebrook, Owen, 1995).

Gittins indices are described by Kaelbling et al. (1996) as a formally justified technique. Kaelbling et al. (1996) describe other formally justified techniques, as well pointing to *"ad-hoc"* strategies which, although not optimal, act as reasonable and computationally tractable heuristics. The next section of this report will highlight studies which have made use of heuristic models, particularly those with roots in psychological theory.

2.3 Heuristic Cognitive Models

In the extensive survey of the bandit literature conducted by Burtini et al. (2015), the authors fail to mention psychologically inspired heuristic models. This is likely due to the fact that Burtini et al. (2015) aim to classify bandit types and their solutions based on optimality and minimum regret bounds and, as highlighted by

Kaelbling et al. (1996), heuristic strategies do not provide optimal solutions to learning problems. One of the reasons bandit problems are interesting to study is that they allow researchers to determine models and strategies which can be used to make optimal decisions in sequential decision making tasks, and help navigate the exploration-exploitation tradeoff. For cognitive scientists however, bandit problems allow conclusions to be drawn not only about optimal solutions to decision making tasks, but also about how the human mind deals with uncertainty, balancing exploration and exploitation, and the reward policies that humans are operating under. In order to investigate these questions, we need quantitative mathematical models with explicit psychological content.[2] With these models, values of psychological variables can be analysed in order to make inferences on how humans solve a task, and apply the models to both human and optimal data in order to examine the relationship between human and optimal performance. Doing so also allows researchers to determine which models give the best account of optimal data, and thus could be used as a computationally tractable heuristic - the method of generating optimal decision-making data via dynamic programming methods is computationally expensive, as will be discussed in later chapters of this report. In addition to this, we can also determine where human performance is falling short of optimal and use this information to inform and teach decision-makers.

Cognitive process models have been applied to a variety of situations including word learning (Frank et al., 2009), preference learning (Jern, Lucas, & Kemp, 2011), and categorization (Nosofsky, 1986). Cognitive process models have also been applied to bandit problems (Steyvers et al., 2009; Lee et al., 2011; Zhang

[2]commonly referred to in the literature as process models (Farrell & Lewandowsky, 2018)

& Angela, 2013). The details of the specific models under consideration is left to Chapter 3 of this report, however it is sufficient to say now that various models have given good account of both human and optimal behaviour, and analysis of the psychological variables involved have provided insights into how humans manage the exploration-exploitation tradeoff, with results from Zhang and Angela (2013) finding that humans tend to explore more than an optimal process would suggest, and Lee et al. (2011) show that in a model which allows decision makers to switch between latent exploration and exploitation states at will, that once a decision maker switches from exploration to exploitation, they never switch back. These results highlight the motivation for using non-optimal heuristics in analyses discussed in Chapter 5.

2.4 Individual Differences

"Individual differences in cognitive processes are basic, ubiquitous, and important." (Zeigenfuse & Lee, 2009).

The study of individual differences is a field in-and-of itself in the broader category of the study of psychometrics, with the study of indivudal differences in cognition dating back to 1973 (Hunt, Frost, & Lunneborg, 1973). The field is still an active area of interest, with research conducted in the fields of economics (Pothos et al., 2011), health (Hu et al., 2015), and genetics (Gottschling, Spengler, Spinath, & Spinath, 2012). It has been pointed out, however, that there has been less of a consideration of individual differences in the cognitive sciences (Zeigenfuse & Lee, 2009).

Farrell and Lewandowsky (2018) report that "[m]ost psychological experiments

report data at the group level, usually after averaging the responses from many subjects in a condition". The benefits of aggregating data include the fact that it is simpler to perform an analysis on one averaged data set than many individual ones, both in terms of human and computational resources, and the risk of overfitting on data is minimised. However, Farrell and Lewandowsky (2018) warn that "averaging may create a strikingly misleading picture of what is happening in [an] experiment". In particular, when working with human data, aggregation assumes that all individuals are the same, and ignores the complexities of human psychology and the multitude of factors, and their interactions, which come into play during decision-making. For example, in studies of individual differences in humans, increased age has been shown to lead to reduced exploration in two types of search tasks (Mata, Wilke, & Czienskowski, 2013), individuals with depression show increased exploration than non-depressive individuals (Blanco, Otto, Maddox, Beevers, & Love, 2013), and low levels of dopamine have been linked to low levels of exploration (Hills, 2006). A comprehensive summary of the range of individual differences which have been studied in relation to the exploration-exploitation trade-off is provided by Mehlhorn et al. (2015).

In the past, research has been redacted, and different conclusions have been drawn, based on whether models were fit on aggregated or individual data. In one study on response time of individuals, the power law best described aggregated performance, however when applied to the individual it gave a worse account than an exponential function (Heathcote, Brown, & Mewhort, 2000).

While it is impossible when developing cognitive process models to account for the infinite variations in factors acting upon an individual, the very basic idea that

individual variation is to be expected should be accounted for. When dealing with bandit problems we are attempting to make inferences about how humans mange the exploration-exploitation tradeoff; such inferences would be unfounded if we don't account for the fundamental fact that individuals vary in their biases towards exploration and exploitation: multiple studies have found that self-reported "maximizers" prefer to accumulate knowledge before making a decision compared to their "satisficing" counterparts (Parker, De Bruin, & Fischhoff, 2007; Schwartz et al., 2002), meaning that for certain individuals, we should expect increased levels of exploration before they converge on a single choice.

Whilst aggregating data has its disadvantages, simply fitting models to the data of each individual does not allow for meaningful analyses either. We want a way of developing models which are able to best fit individual participants, whilst also allowing inferences to be made about humans in general, and on how different individuals, and groups of individuals, compare to each other in how they solve similar problems. Frameworks for conducting such analyses are described in the next section of this report.

2.5 Hierarchical Bayesian Models

Hierarchical Bayesian Models (HBMs) have been studied as early 1972 (Lindley & Smith, 1972), however, they have been primarily of use in the field of statistics, and have not yet received widespread adoption in the cognitive sciences. In a special issue of the *Journal of Mathematical Psychology* (Lee, 2011b), all articles presented applied HBMs to previously studied tasks such as the study of confidence (Merkle, Smithson, & Verkuilen, 2011), memory (Pooley, Lee, & Shankle, 2011), and decision-

making (van Ravenzwaaij, Dutilh, & Wagenmakers, 2011). The issue highlighted the flexibility of HBMs, and the benefits they provide to researchers, including: allowing researchers to form deeper theories with richer psychological content; allowing the same set of parameters to be used to explain behaviour across different but related tasks; and allowing for fundamentally different models to be mixed and unified to better explain observed data.

This section will detail extensions to simple models of cognition which have been proposed in the past, and attempt to persuade the reader that HBMs provide a framework for implementing these extensions which leads to richer model building and more complete analyses in the cognitive sciences.

2.5.1 Hierarchical Modelling

Cognitive process models, in their simplest form, assume that data, d, is generated by some function - a model - which is parameterized by some set of parameters, θ. A graphical representation of this idea is shown in Figure 2.1.

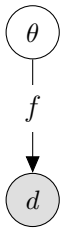


Figure 2.1: Non-hierarchical cognitive model

The exact definition of what makes a model "hierarchical" is not defined, with most researchers providing a definition by example. In order to continue with this section of the report, the definition by Lee (2011a) is adopted which states that: "we treat as hierarchical any model that is more complicated than the simplest possible

type of model shown in [Figure 2.1]."

A simple and immediate extension to this model is to assume that the parameters of the generating process, f, are themselves generated by yet another process, g, with it's own set of parameters, λ; Figure 2.2 shows this extension, and makes clear the inspiration for the name "hierarchical" model.

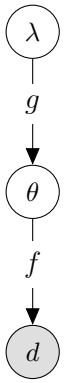


Figure 2.2: Hierarchical extension of simple cognitive model

One inspiration for the need for such models lies in the desire to accommodate individual differences. Lee (2011a) describes in his overview of HBMs that: "The non-hierarchical approach has to rely on first doing separate inference for parameters and data for each person, and then trying to say something about individual differences through post-hoc analyses. In the hierarchical approach in [Figure 2.2], the structure in individual differences is directly captured by the process [g] and its parameters [λ]."

The kind of model presented in Figure 2.2 not only allows for individual differences "but imposes a model structure on those differences, and allows inference about parameters – like the group mean and variance – that characterize the individual differences." (Lee, 2011a).

HBMs also have use in areas other than accounting for individual differences;

Kemp, Perfors, and Tenenbaum (2007) use the framework to capture the notion of "overhypotheses": constraints on hypotheses considered by a learner during a variety of learning tasks. The researchers developed models which acquired overhypotheses in two different tasks: word learning and categorization, and found that HBMs were able to acquire knowledge of overhypotheses. The HBMs, however, relied on the fact that the information at each level could be generated only as long as the generating process and its priors at the next level up were defined. Kemp et al. (2007) note that whilst any kind of induction is impossible without an initial baseline, the question for any learning framework is whether or not models which require no initial assumptions - besides the most fundamental - can be built. The authors are unsure whether HBMs meets this requirement, however, they believe that further research into these models is needed before any conclusions are drawn.

2.5.2 Mixture Modelling

In section 2.4, the distinction of humans into two groups of decision makers, *maximizers* and *satisficers* (Schwartz et al., 2002), was presented. We might expect that these different groups may solve a problem in fundamentally different ways, and as such, we would want to capture this phenomenon when developing a process model. There is no need, of course, to stop at only two groups: we could imagine that any number of groups may exist, and that each group solves the same problem using a different approach. In this case, a mixture model would best explain the data. According to Farrell and Lewandowsky (2018) "Mixture modeling ... [applies] to cases where we suspect that different participants might perform the task differently, either due to discrete differences in ability or due to differences in strategies

used, but where we have no external indicator of the subsets except for performance on our task." HBMs provide a ready framework for such models to be built. The model shown in Figure 2.3 contains multiple processes, f_1 through f_n, each with its own set of parameters, θ_i. A latent variable, z, along with a mixture process, h, is added to the model to determine how the different processes are used to generate the data: this may be a discrete process which allows for only one model to be used at a time, or it may combine results from each process, weighting them according to the latent mixture parameter - the specifics of the model will change from task to task, however these different options are highlighted to exemplify the flexibility of such models.

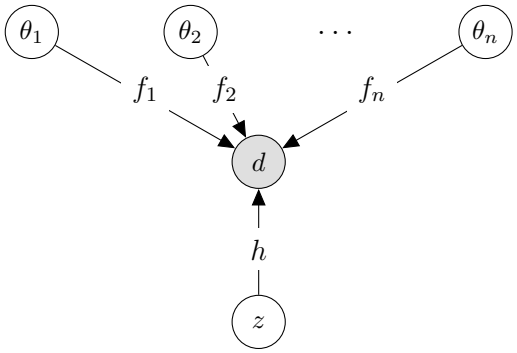


Figure 2.3: Mixture model allowing processes to combine to produce the observed data

Finally, the most ambitious type of HBM is presented in Figure 2.4.

This model assumes that a mixture of processes can be used to explain the observed data, with each process' parameters differing, but stemming from the same underlying distribution. Of course we don't have to stop at just one level of hierarchical abstraction: we can - as some models detailed in Chapter 3 of this report do - encode in our model the notion that each θ_i is generated according to its own process, g_i, parameterized by some values λ_i. We can then place a prior on these

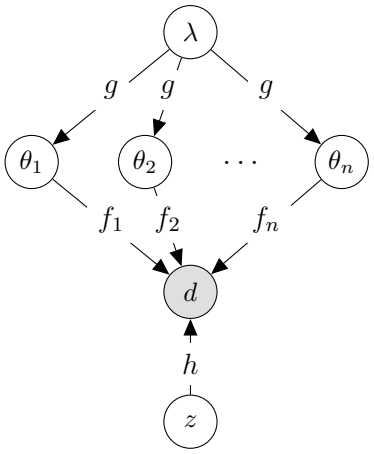


Figure 2.4: Hierarchical extension of a mixture model

values of λ_i, assuming perhaps that they are generated by some underlying process, m, parameterized by yet another value, ϕ; or we can go further up the hierarchy, assuming individual generating processes for each λ_i, and so on. This process can be repeated ad infinitum, and again highlights the flexibility of HBMs: once the structure of the original cognitive model has been defined, extra processes and hierarchical extensions can be added, the model can be fit to data, analyses generated, and further modifications to the model made. Of course, each additional feature adds complexity to the model and will lead to increased computational time and resources per analysis, as well as increasing the likelihood of model overfitting, highlighting the need for formal justification of methods used during model design and building.

The models described thus far have one fundamental limitation: they place a prior assumption on the number of groups present in the data. This limitation has been highlighted by Navarro et al. (2006), who adopt a Dirichlet process prior to model group assignment. By doing so, the researchers assume that there are an infinite number of possible groups to which participants can be assigned, but in any given data-set, only some finite number of groups will actually be observed. This

removes the need for a hard-coded prior on the number of groups present, and instead allows the number of groups to expand in order to better explain the data. One motivating reason for adding hierarchical elements to the mixture model is that it allows for increased capacity for capturing individual differences. In their conclusion, Navarro et al. (2006) report that a natural extension to their "infinite-groups" model is to allow for variation within a group. The idea proposed here is motivated by the intuitive notion that people may differ broadly in the strategies used to solve a problem - referred to as "between-group" differences - and that people who belong to the same group may also differ from each other, albeit to a lesser extent than to how they differ from individuals in other groups - referred to as "within-group" differences. Zeigenfuse and Lee (2009) develop upon this idea by constructing a non-parametric Bayesian model which allows for both between-group differences - referred to in the paper as "discrete" differences - as well as well allowing for variation within the group - dubbed "continuous" differences. They apply their model - known as the "discrete and continuous individual differences (DCID)" model - to a set bandit problems, and find that allowing for both discrete and continuous differences, as opposed to only discrete differences as is the case with the model developed by Navarro et al. (2006), better fits the observed data.

Research has also been conducted into applying HBMs to bandit problems, and has lead to inferences into how people differ in the fundamental problem solving approaches on such tasks, as well as the more subtle ways in which individuals using the same decision-processes differ from one another (Zhang & Lee, 2010b, 2010a; Lee & Newell, 2011). Lee and Newell (2011) argue that "Taking existing successful

models of cognition and embedding them within a hierarchical Bayesian framework opens a vista of potential extensions and improvements to current modeling...". In the next chapter, existing models which have been applied to bandit problems are described, and details are presented as to how they can be expanded via the hierarchical Bayesian framework.

Chapter 3

Models

In this section I will give a brief outline of the models from the bandit literature which I have implemented and evaluated during my project. For each model, I will provide a simple description of the model, a graphical model - where appropriate - and a function describing the likelihood of observed data under the model.

Some models described, such as the ϵ-greedy model and its variants, are common in the reinforcement literature and will be familiar to the reader, whilst other models have been introduced by researchers who have studied bandit problems, namely the τ-switch model, and one Bayesian mixture model developed by Zhang and Lee (2010b). Finally, I will outline a new hierarchical mixture model which has been developed during the course of this project.

Where any substantial changes to a single model have been made, both the original and the modified version are presented. For example, the original τ-switch model (Lee et al., 2009) is designed for two-armed bandit tasks. The data[1] used in my later analysis was gathered by presenting a group of people with four-armed

[1]This is the same data gathered by Steyvers et al. (2009) and was generously provided by one of the original authors, Michael Lee

bandit problems. As a result, I have had to make significant changes to the model so that it can be applied to four arms. The changes made are sufficient for the τ-switch model to be applied to an arbitrary number of arms.

A note on graphical models and terminology

Graphical models have seen recent use in representing probabilistic generative models in the cognitive sciences (Shiffrin, Lee, Kim, & Wagenmakers, 2008) with Zhang and Lee (2010a) reporting that "[t]he practical advantage of graphical models is that sophisticated and relatively general-purpose Markov Chain Monte Carlo (MCMC) algorithms exist that can sample from the full joint posterior distribution of the parameters conditional on the observed data". Multiple introductions to these models are available (L Griffiths, Kemp, & B Tenenbaum, 2008; Lee, 2008). In this report, I adopt the formalism presented by Lee (2008) in which "*... nodes represent variables of interest, and the graph structure is used to indicate dependencies between the variables, with children depending on their parents. The conventions of representing continuous variables with circular nodes and discrete variables with square nodes and of representing unobserved variables without shading and observed variables with shading are used. Stochastic and deterministic unobserved variables are distinguished by using single and double borders, respectively. Plate notation, enclosing with square boundaries subsets of the graph that have independent replications in the model, is also used.*"

Other notation in the graphical models, and likelihood functions, used in this report includes:

- D to represent the set of decisions made by a participant, with a decision in

game g and trial k denoted by D_{gk}

- R denoting the rewards received during the task
- S and F denoting the number of successes and failures thus far for a particular arm for the current game

3.1 Guessing

This model is the simplest one considered: given N arms, the model chooses between them uniformly at random. The purpose of this model is to act as a baseline for comparison of other models. The likelihood function for this model is as follows:

$$p(D_k^g = i | M_{guessing}) = \frac{1}{N} \tag{3.1}$$

3.2 Win-Stay Lose-Shift

The Win-Stay Lose-Shift (WSLS) model is a classic model in the reinforcement learning literature (Sutton & Barto, 1998). In its deterministic form, the model assumes that people stick with a choice while they continue to receive a reward from it, and when the arm no longer pays off, they switch to another. The stochastic version of the model assumes that the participant sticks with a "winning" arm with (high) probability γ, and on any given trial will switch from a winning arm with probability $1 - \gamma$. This parameter, γ, has been described as an "accuracy of execution parameter (Steyvers et al., 2009), and is useful in describing noisy decision-making data. Many of the models described in this chapter have some

form of accuracy of execution parameter to account for suboptimal decisions in human data. The likelihood function for this model is:

$$p(D_k^g = i|R, \gamma, M_{WSLS}) = \begin{cases} \frac{1}{N} & \text{if } k = 1 \\ \gamma & \text{if } k > 1,\ D_{k-1}^g = i \text{ and } R_{k-1}^g = 1 \\ \frac{1-\gamma}{N-1} & \text{if } k > 1,\ D_{k-1}^g \neq i \text{ and } R_{k-1}^g = 1 \\ 1-\gamma & \text{if } k > 1,\ D_{k-1}^g = i \text{ and } R_{k-1}^g = 0 \\ \frac{\gamma}{N-1} & \text{if } k > 1,\ D_{k-1}^g \neq i \text{ and } R_{k-1}^g = 0 \end{cases} \tag{3.2}$$

Figure 3.1 shows the graphical model for the WSLS model, providing the first example of how a simple cognitive model can be extended via the hierarchical Bayesian framework described in section 2.5 of this report.

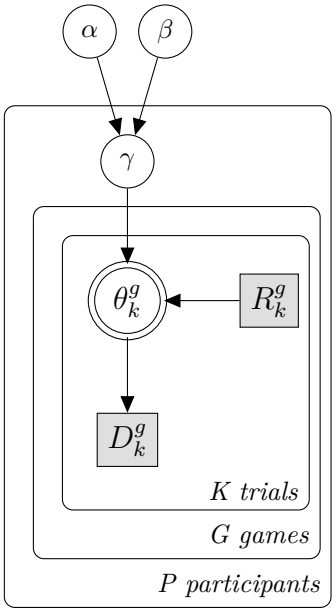


Figure 3.1: Bayesian graphical model for the WSLS model

Figure 3.1 shows how we can allow for individual variation in the "sticking" probability, , used by each participant, whilst assuming commonality in all participants

by assuming this parameter comes from some underlying process, described here by a Beta distribution, parameterized by α and β. The use of a Beta distribution for modelling human assumptions about the environment is standard practice when working with bandit problems (Steyvers et al., 2009). One reason for using a Beta distribution is that it allows for an intuitive explanation of how we believe people might think about the environment: the two parameters used, α and β, can be thought of as a count of prior successes and prior failures, respectively. With these parameters at hand, we can construct psychological assumptions about the level of optimism a player has about the environment as $\frac{\alpha}{\alpha+\beta}$, and the level of certainty they have in their optimism as $\alpha + \beta$. We should expect that people will behave very differently based on these prior assumptions on the environments: where levels of optimism are high, we might expect higher levels of exploration as participants believe there are many high paying arms. Conversely, in situations were optimism is lower, or participants are less certain in their assumptions, they may explore less often and choose to stick with whichever arm they have relatively more information on. Once a model has been fit to the data, we can analyse the posterior estimates of α and β to make inferences about participants assumptions about the environment, as will be seen in later analyses.

3.3 Extended Win-Stay Lose-Shift

The extended Win-Stay Lose-Shift (e-WSLS) is very similar to the original WSLS model described above, except that the probability of staying with a winning arm and the probability of switching from a losing arm are no longer parameterized by the same probability γ: the probability of staying with a winning arm is γ_w, while

the probability of switching from a losing arm is now γ_l. The likelihood function, and graphical model, for this model are shown below:

$$p(D_k^g = i|R, \gamma_w, \gamma_l, M_{e-WSLS}) = \begin{cases} \frac{1}{N} & \text{if } k = 1 \\ \gamma_w & \text{if } k > 1,\ D_{k-1}^g = i \text{ and } R_{k-1}^g = 1 \\ \frac{1-\gamma_w}{N-1} & \text{if } k > 1,\ D_{k-1}^g \neq i \text{ and } R_{k-1}^g = 1 \\ 1 - \gamma_l & \text{if } k > 1,\ D_{k-1}^g = i \text{ and } R_{k-1}^g = 0 \\ \frac{\gamma_l}{N-1} & \text{if } k > 1,\ D_{k-1}^g \neq i \text{ and } R_{k-1}^g = 0 \end{cases} \tag{3.3}$$

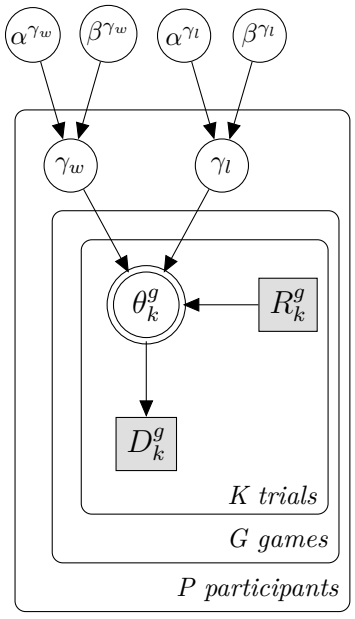


Figure 3.2: Bayesian graphical model for e-WSLS model

3.4 ϵ-Greedy

The ϵ-greedy model features prominently in many domains[2] and is a classic in the reinforcement learning literature. On any given trial, the model chooses an arm at random according to probability ϵ, otherwise it chooses the "best" arm. In order to choose the "best" arm in this context, the model maintains a record of the proportion of successes and failures of each arm, and chooses an arm based on it's ratio of successes to failures. The likelihood function for the ϵ-greedy model is as follows:

$$p(D_k^g = i|S, F, \epsilon, M_{\epsilon-greedy}) = \begin{cases} \frac{1}{N} & \text{if } k = 1 \\ \frac{\epsilon}{N_{max}} & \text{if } k > 1 \text{ and } i \in \arg\max_j \frac{S_j+1}{S_j+F_j+2} \\ \frac{1-\epsilon}{N-N_{max}} & \text{otherwise} \end{cases} \quad (3.4)$$

Figure 3.3 shows the graphical model for the ϵ-Greedy model.

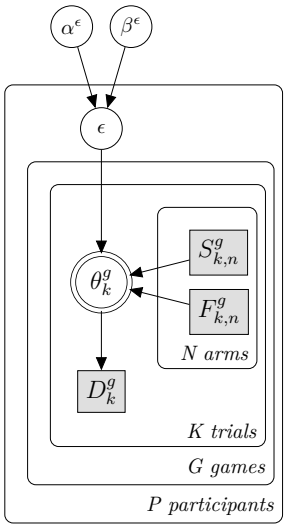


Figure 3.3: Bayesian graphical model for the ϵ-Greedy model

[2]In the paper by Steyvers et al. (2009), the ϵ-greedy model is used under the name Success Ratio.

3.5 ϵ-Decreasing

The ϵ-decreasing model is a simple variant of the ϵ-greedy model presented above. The key difference between the two models is that the value of ϵ remains fixed in the ϵ-greedy model, whilst it decreases as time progresses in the ϵ-decreasing variant. The psychological content of this model is suggestive of the fact that people tend to explore less as time goes on and they gather more information. The value of ϵ can decrease in many ways: some researchers have proposed decreasing it according to a power law, others by an exponential law, whilst others suggest a simple linear decrease. I will follow the method suggested by Zhang and Lee (2010b) and decrease ϵ linearly with time; one of the aims of this paper is to reproduce findings reported of Zhang and Lee (2010b), and so ensuring my models are consistent with the ones in that paper is vital to this analysis. The likelihood function for this model is:

$$p(D_k^g = i|S, F, \epsilon, M_{\epsilon-decreasing}) = \begin{cases} \frac{1}{N} & \text{if } k = 1 \\ \frac{\epsilon/k}{N_{max}} & \text{if } k > 1 \text{ and } i \in \arg\max_j \frac{S_j+1}{S_j+F_j+1} \\ \frac{1-(\epsilon/k)}{N-N_{max}} & \text{otherwise} \end{cases} \tag{3.5}$$

and the graphical model is shown in Figure 3.4:

3.6 π-First

The π-first is usually referred to as the ϵ-first model in the reinforcement literature (Sutton & Barto, 1998), however, Lee et al. (2009) introduce the name π-first in

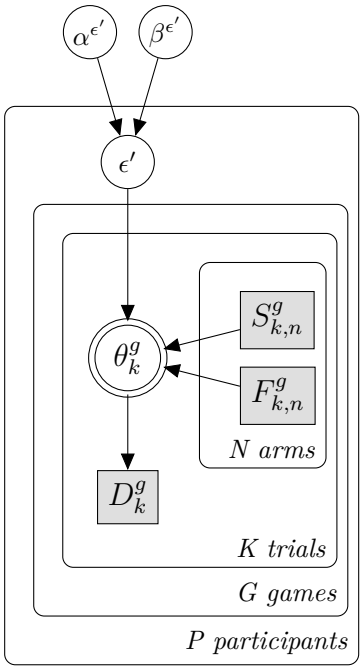


Figure 3.4: Bayesian graphical model for the ϵ-Decreasing model

order to highlight the difference between it, and the two other variants on the ϵ model described thus far. I will adopt the name used in that paper for sake of consistency.

The model describes decision-making as taking place in two distinct phases: an exploratory phase in which all decisions are made at random, and exploitative phase where the best arm is greedily chosen. The time spent in the latent "explore" phase is controlled by the parameter π, which dictates after which trial number the participant switches to the "exploit" phase. The researchers are not clear in whether or not the information gathering phase stops at the end of the exploratory phase, or whether results during exploitation are also put to use. Whilst it may seem trivial and obvious that information encountered should be used to update the model regardless of the latent phase the participant is in, I would like to formally state here that this was the assumption I made, so as to resolve this slight ambiguity.

Lee et al. (2009) veer away from the purely deterministic version of the model

and, similarly to their implementation of the WSLS model, include an accuracy of execution parameter, γ, to allow for suboptimal decisions to be made. Lee et al. (2009) do not give a formal definition of the likelihood function for this model, and so I present one here, constructed one from the description of the model given in their paper:

$$p(D_k^g = i|S, F, \pi, \gamma, M_{\pi-first}) = \begin{cases} \frac{1}{N} & \text{if } k \leq \pi \\ \frac{\gamma}{N_{max}} & \text{if } k > \pi \text{ and } i \in \arg\max_j \frac{S_j+1}{S_j+F_j+1} \\ \frac{1-\gamma}{N-N_{max}} & \text{otherwise} \end{cases} \quad (3.6)$$

Figure 3.5 shows the graphical model for the π-first model.

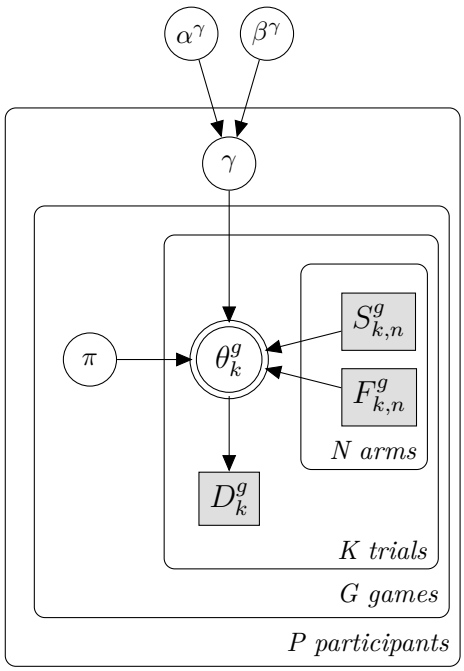


Figure 3.5: Bayesian graphical model for for the π-first model

3.7 Latent State and τ-Switch Models (2-arms)

3.7.1 Latent State Model

The Latent State Model is first proposed by Zhang, Lee, and Munro (2009), with a full description of the model provided by Lee et al. (2009) and again later by Lee et al. (2011). Lee et al. (2009) develop a full latent state model which extends the π-first model to allow for repeated switching between the two latent states, rather than simply switch once from exploration to exploitation as outlined in the π-first model. The other difference between the π-first and their new latent state model is the mechanism by which the model makes "explore" and "exploit" decisions. In the π-first model, exploring consists of choosing an arm at random, whilst exploiting means greedily choosing the arm with the highest proportion of successes encountered so far. The new latent state model proposed distinguishes between 3 different situations, and acts according to both the latent state - explore or exploit - and the situation the model finds itself in. The 3 situations are described in the original paper as follows:

"In the same situation, both alternatives have the same number of observed successes and failures. ... For the same situation, both alternatives have an equal probability of being chosen."

"In the better–worse situation, one alternative has more successes and fewer failures than the other alternative (or more successes and equal failures, or equal successes and fewer failures). In this situation, one alternative is clearly better than the other. ... For the better–worse situation, the better alternative has a high probability, given by a parameter γ, of being chosen. The probability the worse

alternative is chosen is $1 - \gamma$*."*

"In the explore–exploit situation, one alternative has been chosen more often, and has more successes but also more failures than the other alternative. In this situation, neither alternative is clearly better, and the decision-maker faces the explore–exploit dilemma. Choosing the better-understood alternative corresponds to exploiting, while choosing the less well-understood alternative corresponds to exploring. ... In this situation, our model assumes the exploration alternative will be chosen with the high probability γ *if the decision-maker is in a latent 'explore' state, but the exploitation alternative will be chosen with probability* γ *if the decision-maker is in the latent exploit state. In this way, the latent state for a trial controls how the exploration versus exploitation dilemma is solved at that stage of the problem."* (Lee et al., 2011)

Lee et al. (2011) applied this latent model to a set of data collected from 10 participants who completed a verity of bandit problems which ranged in the number of trials completed, and the latent reward states of the environment. The researchers found high agreement between the model with both human and optimal data.

3.7.2 τ-Switch

Upon applying the full latent state model to both human and optimal data, Zhang et al. (2009) observed that once the model switched from the latent explore to exploit state, it never switched back. In a follow-up paper, Lee et al. (2009) simplified the model to be more similar to the original π-first model, removing the latent state parameter for each trial and instead incorporating a single switch-point parameter, τ, giving rise to the τ-switch model. This model is now more similar to the π-first model in that a single parameter encodes the switching point, however, it is still

fundamentally different from the π-first model due to it's modelling of 3 different situations which are used to control the models decision making.

One pitfall of the original model is that it is defined only for two-armed bandits; this fact was confirmed after reaching out to one of the authors[3] of the original paper. An aim of this project was therefore to extend this model so that it could be applied to bandits with an arbitrary number of arms. The motivation for this extension was due to the results presented in a follow-up paper by Lee et al. (2009): the researchers found that the τ-switch model gave the best account of both human and optimal data out a variety of heuristic models considered, including all of the models previously defined thus far. This result, however, was obtained by applying the models to a set of data collected over only 10 participants. In order to apply the Latent State and τ-switch model to a larger data set, the first study of its kind to do so, the model first had to be extended to work with N-armed bandits, with a formal definition of the model presented in the next section.

3.8 Latent State and τ-Switch (N-arms)

To define the Latent State and τ-switch model for N arms it is helpful to define two sets: S and F, where S consists of all of the arms who have number of successes equal to the maximum number of successes of all arms, and F consists of all of the arms with number of failures equal to the minimum number of failures. Each of these sets is calculated at the beginning of each trial based on previous successes and failures. These sets can then be used to determine behaviour of the model in

[3]Michael Lee

each of the 3 situations outlined in the original paper.

Same situation

The same situation was the easiest to modify: rather than considering only whether two arms have the same number of successes and failures, the model is in the same situation if two *or more* arms appear in both S and F, i.e. two or more arms have both the maximum number of successes, and the minimum number of failures, therefore all alternatives have equal probability of being chosen.

Better-Worse situation

We encounter the better-worse situation if only one arm is in both S and F, meaning only one arm has had both more successes **and** fewer failures than all other arms, so it is clearly the best choice. This arm is therefore chosen with probability γ, or one of the worse alternatives are chosen with probability $\frac{1-\gamma}{N-1}$.

Explore-Exploit situation

The explore-exploit situation arises when $S \cap F = \emptyset$, and both S and F contain at least one element. This means that at least one arm has had more successes than all other arms, but also more failures, meaning more information is available for these arms and so choosing one of these arms would constitute and exploit decision. Alternatively, choosing an arm from those in F - the ones we have less information on - would constitute an explore decision. Again, an accuracy of execution parameter, γ is included in the model to allow for suboptimal decisions. Note that this situation is different from the better-worse situation since that the arm with the greatest number of success no longer as also has the minimum number of failures.

The likelihood function for the τ-switch model for N-armed bandits is presented as follows, with the latent state model following a similar structure with τ replaced with a latent switching parameter z:

$$p(D_k^g = i|S, F, \tau, \gamma, M_{\tau-switch}) = \begin{cases} \frac{1}{|S_k \cap F_k|} & \text{if } i \in S_k \cap F_k, \text{ and } |S_k \cap F_k| > 1 \\ \gamma & \text{if } i \in S_k \cap F_k, \text{ and } |S_k \cap F_k| = 1 \\ \frac{1-\gamma}{N-1} & \text{if } i \notin S_k \cap F_k, \text{ and } |S_k \cap F_k| = 1 \\ \frac{\gamma}{|F_k|} & \text{if } k \leq \tau,\ i \in F_k, \text{ and } |S_k \cap F_k| = 0 \\ \frac{1-\gamma}{|F_k|} & \text{if } k > \tau,\ i \in F_k, \text{ and } |S_k \cap F_k| = 0 \\ \frac{1-\gamma}{|S_k|} & \text{if } k \leq \tau,\ i \in S_k, \text{ and } |S_k \cap F_k| = 0 \\ \frac{\gamma}{|S_k|} & \text{if } k > \tau,\ i \in S_k, \text{ and } |S_k \cap F_k| = 0 \end{cases} \tag{3.7}$$

Figure 3.6 shows the graphical model for the τ-switch model.

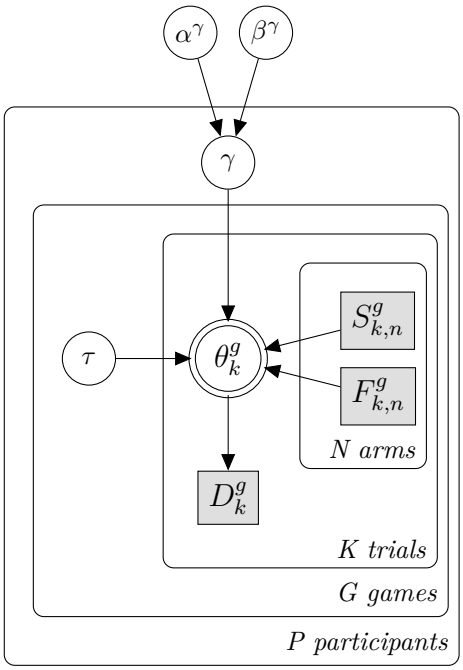


Figure 3.6: Bayesian graphical model for the τ-switch model

3.9 Optimal

For a fixed environment, we can calculate the optimal decision process according to the dynamic programming method outlined by Kaelbling et al. (1996). The expense of calculating the optimal decision policy for a bandit problem is exponential, however, for a small enough number of trials and arms, the computation is feasible. The idea behind the method, as is the case with all dynamic programming methods, is that we choose a base case, in this case the last trial, and assume that on the final trial the arm with the greatest expected reward is chosen. From here, we choose the arm on the second-last trial which gives the greatest expected reward over the final two trials, assuming that the final decision is made optimally. We iterate backwards through the sequence of trials in this manner, and in the end we will have an optimal sequence of decisions for the entire problem.

In order to calculate the optimal policy, we must compute a mapping between belief states and actions (Kaelbling et al., 1996). A belief state is a representation of the information available to the agent: Kaelbling et al. (1996) define this as a vector containing the number of times an arm has been pulled, n and the rewards, w received so that a belief state takes the form $\{n_1, w_1, ..., n_k, w_k\}$; Steyvers et al. (2009) use a slightly different notation where a belief state for a given game, g, and given trial, k, is represented as the number of successes and failures for each arm, denoted s_i and f_i respectively for arm i so that the belief state is of the form: $\{s_1, f_1, ..., s_N, f_N\}$. I will adopt the latter notation for the remainder of this report for consistency, as the work of Steyvers et al. (2009) is the subject of the replication study described later.

The expected additional reward to be gained by acting optimally for the remainder of trials, given that we are currently at trial k, is denoted $V_k^*(s_1, f_1, ..., s_N, f_N)$, where the expected additional reward for the final trial $V_K^*(s_1, f_1, ..., s_N, f_N) = 0$. The value of V_k^* for any other trial can be calculated recursively according to the following equation, with the probability of success and failure on each trial defined for Beta(α, β) environments as defined by Steyvers et al. (2009):

$$V_k^*(s_1, f_1, ..., s_N, f_N) = \max_i E\left[\begin{array}{l}\text{Future reward if the agent chooses arm i}\\ \\ \text{and then acts optimally for the remainder of pulls}\end{array}\right]$$

$$= \max_i \left[\begin{array}{l}\frac{s_i+\alpha}{s_i+f_i+\alpha+\beta}V_{k+1}^*(s_1, f_1, ..., s_{i+1}, f_i, s_N, f_N)+\\ \frac{f_i+\beta}{s_i+f_i+\alpha+\beta}V_{k+1}^*(s_1, f_1, ..., s_i, f_{i+1}, s_N, f_N)\end{array}\right]$$

Steyvers et al. (2009) present the likelihood for data under the optimal model as follows, where α and β the participants assumptions about the environment, and w acts as an accuracy of execution parameter:

$$p(D_k^g = i|S, F, \alpha, \beta, w, M_{opt}) = \begin{cases}\frac{w}{N_{max}} & \text{if the } i^{th} \text{ alternative maximises total expected reward}\\ \\ \frac{1-w}{N-N_{max}} & \text{otherwise}\end{cases} \tag{3.8}$$

3.10 Hierarchical Mixture Model

Zhang and Lee (2010a) develop a hierarchical mixture of four decision-making models - WSLS, e-WSLS, ϵ-greedy, and ϵ-decreasing - in order to accommodate for differences not only the parameters used for a given model, but for differences in

the model used by each participant, indicated by the parameter z_p which indicates which model the p^{th} participant is thought to be using. The assignment parameter, z_p, has prior $z_p \sim Categorical(\phi)$ with a uniform Dirichlet prior placed on ϕ so no model one is preferred. In analysis of model differences, the posterior expectation of ϕ is analysed to infer the proportion of participants using each model.

Figure 3.7 shows the graphical model for this mixture model (with Beta, rather than Gaussian priors).

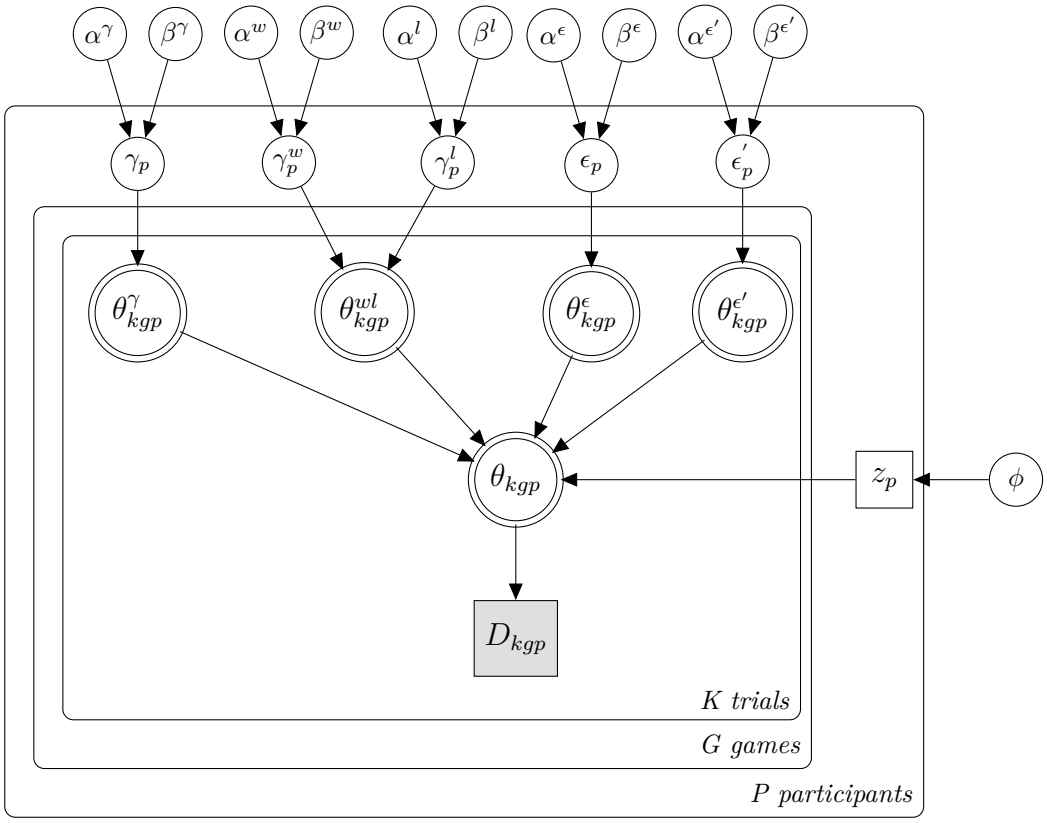


Figure 3.7: Bayesian graphical model for mixture model proposed in (Zhang & Lee, 2010a)

3.11 Extended Hierarchical Mixture Model

The extended mixture model proposed here is inspired by the findings of (Lee et al., 2009) who detail that latent state models give a good account of human decision making data, with the τ-switch model giving the best account of human data of any of the models considered in said paper. Mixture models are constructed in the cognitive sciences in order to allow for more robust accounts of human performance on various cognitive tasks to be accounted for. Since the cognitive models proposed by Lee et al. (2009) are believed to give good descriptions of human decision-making, I believe it is worthwhile to add these models as components in the mixture model developed by Zhang and Lee (2010a). The π-first model is also included in this model so as to further broaden the range of models considered

Figure 3.8 shows the graphical model for the proposed mixture model, with latent switch parameters for the τ-switch and π-first models excluded for sake of readability.

3.12 Extensions to Mixture Models

Throughout our lives, we are subject to many different schools of thought, and taught multiple ways to tackle a problem. Current work in modelling individual differences in the way that different people solve a problem stops at the level of assuming that people differ only from each other, and that each person subscribes to exactly one decision-making policy. I believe that developing latent models which allow a single person to engage with multiple problem solving strategies is an important consideration to make, and allows for a more meaningful and complete account

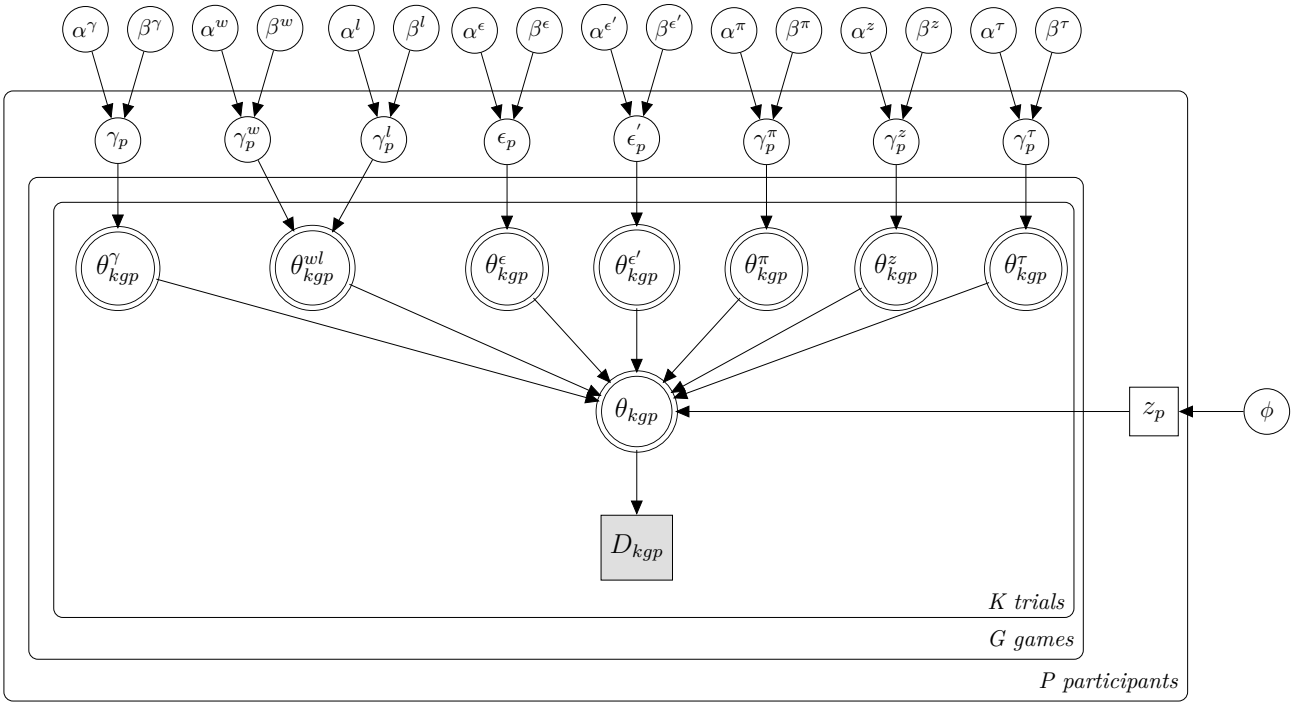


Figure 3.8: Bayesian graphical model for the newly proposed Extended Mixture Model

of human behaviour. I believe that two natural extensions to the mixture models presented in sections 3.10 and 3.11 is to allow participants to switch to a different decision-making policy on either each new game, or each new trial encountered. The idea for allowing game-by-game basis is motivated by the idea that people may have different strategies in mind, and change which policy they follow based on the problem context: a builder know that both a hammer and screwdriver are useful tools, and knows which one to use based on the job that needs to be done. The motivation for the per-trial changes stems from the understanding that people are likely to adapt their problem solving strategies throughout the course of a single problem based on whether or not they think it is working. For example, a programmer may begin to solve using simple linear programming techniques, however when they become aware of time constraints imposed by this approach, they may switch

to a parallel processing solution.

For sake of brevity, I have not included graphical models for the newly proposed models as they are very similar to the graphical models presented in Figures 3.7 and 3.8, with the only difference being that the latent allocation parameter, z is moved inside of the *game* plate and *trial* plate for the per-game and per-trial extensions, respectively.

Chapter 4

Replication

4.1 Model Identifiability and Individual Differences

Motivation

Steyvers et al. (2009) study individual differences in terms of how people balance exploration and exploitation when solving bandit problems. They provide a formal characterization of the optimal Bayesian decision process for bandit problems and compare performance of this model on data gathered from 451 participants to three heuristic models. The performance of models are compared based on Bayesian model selection methods. The authors find evidence of individual differences in the data set, in terms of This data set has been the subject of much study in the field of human performance on Bandit problems: Zhang and Yu (2013) use it in evaluating a new model which approximates the Bayes-optimal policy; and Zhang and Lee (2010a) use it when studying individual differences in how different

individuals employ completely different decision making policies by applying mixture model to the data set and analysing how individuals are assigned to different models. This paper presents two findings relevant to my project. First, the authors claim that they have found clear individual differences between participants in this data set. Secondly, they provide a clear framework for evaluating models in a Bayesian fashion using Bayes Factors (Kass & Raftery, 1995), a standard method in Bayesian model selection. Bayes factors "compare the average likelihood of a participant's decisions being generated under one model rather than the other." The authors use the guessing model as the low-end benchmark when comparing the marginal densities of each model. An example of the ratios calculated is as follows:

$$\begin{aligned} BF_{opt} =& \frac{p(D|M_{wsls})}{p(D|M_{guess})} \\ &= \frac{\int p(D|\lambda, M_{wsls})p(\lambda)d\lambda}{p(D|M_{guess})} \\ &= \frac{\int [\prod_{g=1}^{G}\prod_{k=1}^{K} p(D_k^g|R_{k-1}^g, \lambda)]p(\lambda)d\lambda}{\prod_{g=1}^{G}\prod_{k=1}^{K} p(D_k^g|M_{guess})} \end{aligned} \tag{4.1}$$

This project involves evaluating a newly proposed model on human data to evaluate human performance on bandit tasks and determining whether that model is sufficiently able to capture individual differences; Steyvers et al. (2009) detail both a data set and a framework for such analysis. Before beginning to test new models on the given data set, I believe it is important to first confirm the findings of the original paper, namely that there are individual differences to be found among these 451 participants, and the framework for Bayesian model comparison is appropriate and sufficient to draw conclusions on the performance of different models.

Method

The first part of this replication involves testing model recovery ability as detailed in Section 4 of (Steyvers et al., 2009), in order to test the usefulness of the Bayes Factor model selection framework, as well as general identifiability of the models considered. The models implemented and tested are the Guessing, WSLS, ϵ-greedy, and optimal models. In this paper, the authors used the name "Success Ratio (SR)" model to denote the ϵ-greedy model. For sake of consistency when referring to figures and results from the original paper, I will adopt that name for the remainder of this section. As mentioned above, the authors use Bayes Factors to compare models. In order to calculate the appropriate Bayes Factors, the authors compute the marginal density of each model using a grid based method. For example, in order to calculate the marginal density for the WSLS model (likelihood function is equation 3.2), the authors use a brute-force methods on a grid of 40 evenly-spaced points over the domain $(0, 1)$ for the parameter λ[1] in order to approximate the integral in equation 4.1.

Steyvers et al. (2009) generated 4 artificial decision-making data sets, with each of the data sets corresponding to one of the 4 models: guessing, WSLS, SR, and optimal. In generating the artificial data sets, the authors set the accuracy of execution parameters for each model - λ, δ, and w for the WSLS, SR, and optimal models respectively - to 1, corresponding to the ideal execution of the decision strategy.[2]. With the artificial data sets generated, the authors performed 2 analyses: first, they calculate the log BF measures to find which of the models was best supported for

[1] similar grids were used for δ and w for the Success Ratio and optimal models, respectively

[2] Steyvers et al. (2009) tested parameter recovery was tested with different values and found recovery to be excellent. Such an extended analysis was beyond the scope of this replication

each data set; second, they found the maximum a posteriori (MAP) estimates of the parameters of the best supported model.

In order to replicate the results, a series of classes in Scala[3] were written to generate data according to an optimal policy - with code for the optimal model provided by my supervisor, Chris Lucas - and to calculate the marginal density of observed data under a given model. For each model, two classes were implemented: one which contains methods to generate an artificial data set of decisions made by following that model, and another which calculates the likelihood of observed data under a given model. Common functionality across models was extracted to parent classes which allowed for common interface to provide parameters such as number of trials and participants. One issue that I ran into during this analysis was the computational intensity of the optimal model. As described in Chapter 3, the optimal model is essentially a lookup table of histories to future expected rewards. This meant that in order to calculate the likelihood of a single participants data under the model, I had to maintain a history of their decisions and rewards, and use this to look-up their future expected reward. Whilst in theory these look-ups should be done in near-constant time, in practice the time taken for each lookup was non-negligible. This, compounded by both the large size of the table to account for all possible decision and reward combinations, and the number of look-ups to be performed for each participant meant that the time taken to calculate the a single participants data took over 4 seconds on an 8-core machine. Calculating the marginal density of the optimal model involved estimation the value of an integral similar to that in equation 4.1. For the optimal model, there were 3 parameters to be

[3]https://www.scala-lang.org/

marginalised over: w, α, and β, with w; ranging over 40 points on the domain $(0, 1)$, and each of α and β taking on values between 0.2 and 5.0 in increments of 0.3. For each participant, therefore, calculating the log BF involved calculating the likelihood function $11,560$ times. In practice, it took on average 17 hours[4] to calculate the Bayes Factor for 1 out of 451 participants decisions. I attempted to optimize this by parallelizing the code on a single machine, however, this still resulted in an average of 8 hours per participant. With 451 participants per data set, and 5 data sets in total, calculating the results for the optimal model in this manner was not feasible. To overcome this bottleneck, I was pointed towards the Eddie compute cluster by my supervisor and given guidance for interacting with the cluster by Craig Innes. I modified the Scala code I had written in order to accept a variable number of participants. I wrote scripts to initialise an Eddie node, and split processing into 4 batches, with just over 110 participants per batch. Each Eddie node had 32 cores and as such the paralleization code I had written could be utilised to a greater effect. By parallelizing within each node, as well as splitting participants across multiple nodes, I was able to generate results for the optimal model for all data sets, and complete the first stage of the replication.

The next stage of the replication involved confirming the findings in Section 5 of paper by Steyvers et al. (2009). In their original report, the authors applied the models to human decision making data in order to make inferences about which model best described their decision-making, as well as calculating MAP estimates of the accuracy of execution parameters conditioned on the model with the largest support as dictated by the Bayes Factors calculations. When applying the models

[4]clock time

to human data, the authors assume uniform priors over the accuracy of execution parameters. The authors note that this may not be a reasonable prior assumption, and instead expect that it is likely that a participant will accurately follow of model with high probability. They test this by repeating the analysis with 8 different prior assumptions, each increasing the prior probability that a participant will accurately follow a model, and report how this affects distribution of participants to models. The authors find that there are only minor differences in final distributions when different priors are used. Due to the computational intensity involved in calculating Bayes Factors for the optimal model, and since the authors conclude that different prior assumptions do not have a great effect on final distributions, this replication does not repeat this analysis.

Scala code to calculate Bayes Factors was reused here, and the Eddie cluster used again to compute the marginal density for the optimal model - with only minor changes in analyzing the real decision-making data. The minor changes involved were in loading the data set: the participants completed the same set of bandit tasks, but in a different order. Along with the original data, the block order in which the participants completed the tasks was also provided allowing me to reorganise the decision-making data to allow for 1:1 comparisons between participants and models on any given bandit task.

The results from both of these stages of the replication, reported alongside the original results are described in the next section of this report.

Results

Steyvers et al. (2009) report that the inference methods were able to recover the underlying decision models well, with the correct[5] model always having the most evidence for data generated under WSLS, SR, and optimal models, and 98% correct for Guessing model data. The authors conclude that the results mean they can have confidence in applying the log BF measure to participants in the real behavioural data. Table 4.1 shows the model recovery performance reported by Steyvers et al. (2009), as well as the results obtained from my analysis.

Recovery Model	Guessing	WSLS	Success Ratio	optimal
Guessing	98 (97)	0 (0)	0 (0)	2 (3)
WSLS	0 (0)	100 (100)	0 (0)	0 (0)
Success Ratio	0 (0)	0 (0)	100 (100)	0 (0)
optimal	0 (0)	0 (0)	0 (5)	100 (95)

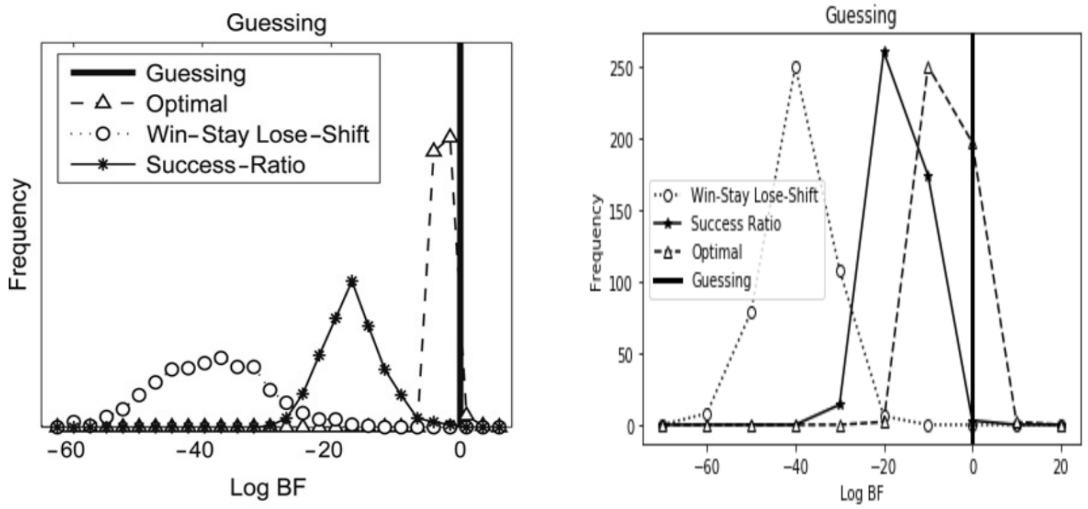
Table 4.1: Model recovery performance (replicated results shown in brackets)

It is clear from the results in Table 4.1 that the original model recovery was almost perfect, and results of this replication are of a very similar nature, with only slight differences in results involving the optimal model.

Figures 4.1 - 4.4 show the original and replicated log BF measures for each model under the artificial . It can be seen that there are very minor differences in the shapes of individual distributions, for example, the peak of the WSLS model in Figure 4.3b is slightly higher in the replicated than in the original results (Figure 4.3a). Since the original frequencies were not reported, I am unfortunately unable to report the

[5]correct as being defined as the model which generated the artificial data set

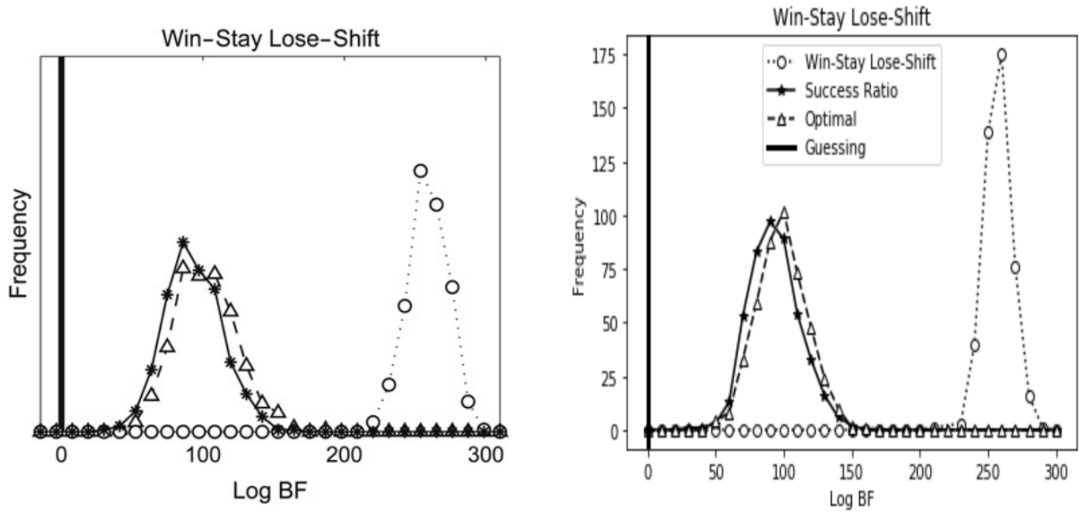
actual differences in values in log BF measures. That being said, the general shape and overall rank of each model per artificial data set is extremely similar to the original result, indicating a successful replication.



(a) Original

(b) Replicated

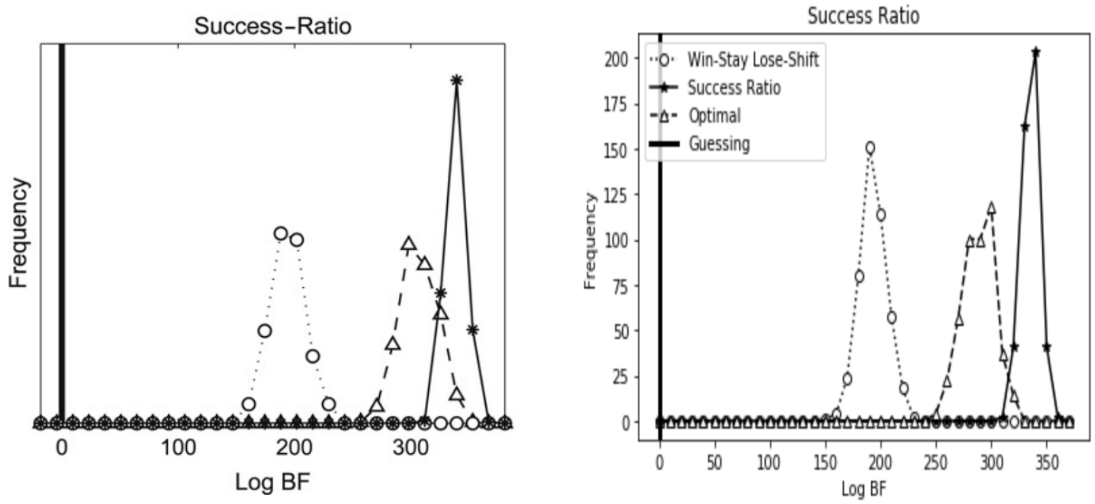
Figure 4.1: log BF measures for models applied to the artificial Guessing data set



(a) Original

(b) Replicated

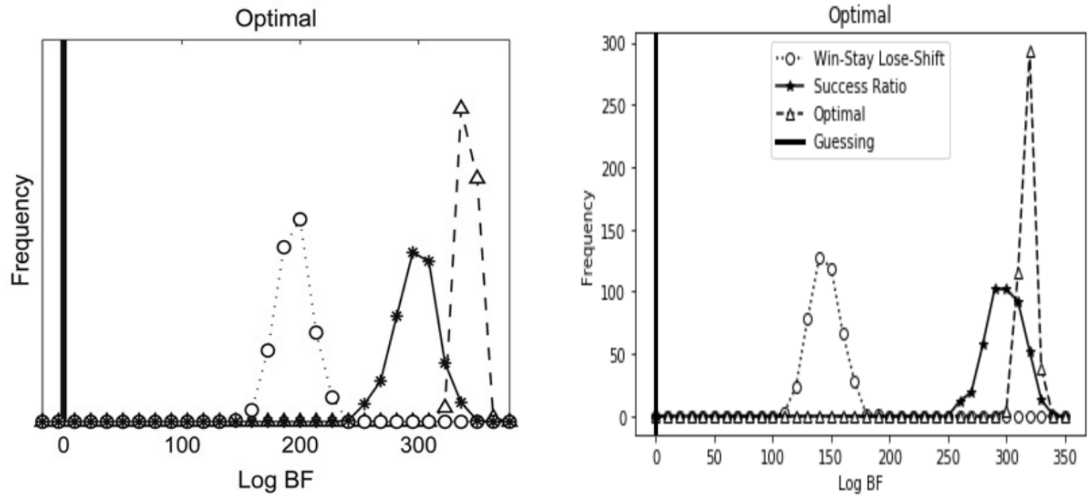
Figure 4.2: log BF measures for models applied to the artificial WSLS data set



(a) Original

(b) Replicated

Figure 4.3: log BF measures for models applied to the artificial SR data set



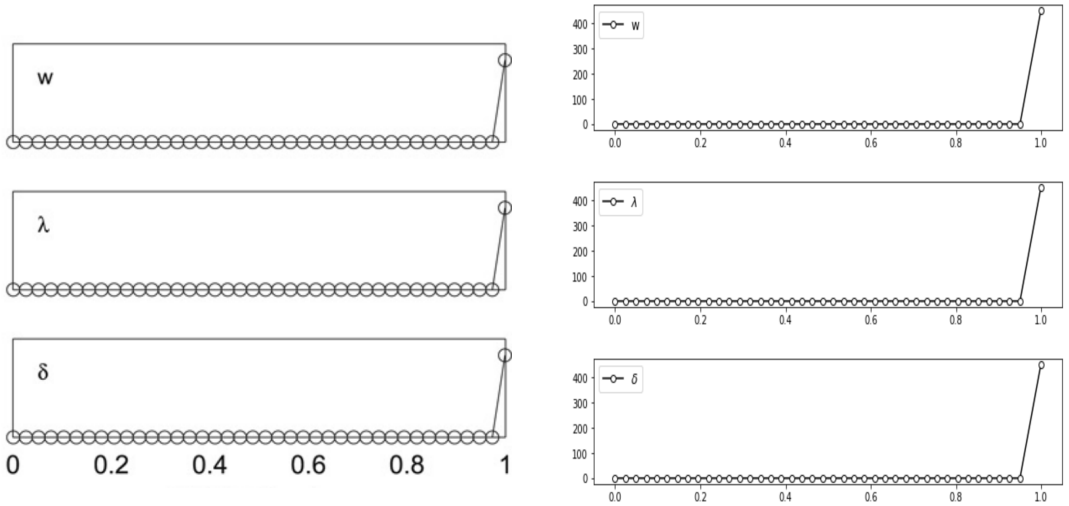
(a) Original

(b) Replicated

Figure 4.4: log BF measures for models applied to the artificial optimal data set

Next, the authors report the distribution of MAP estimates of the accuracy of execution parameters - conditional on the model being the one recovered by the log BF measures - as well as α and β values for the optimal model. These results are shown in Figures 4.5a and 4.6a, where it is clear that parameter recovery is perfect. The distribution of MAP estimates from my replication are shown in Figures 4.5b

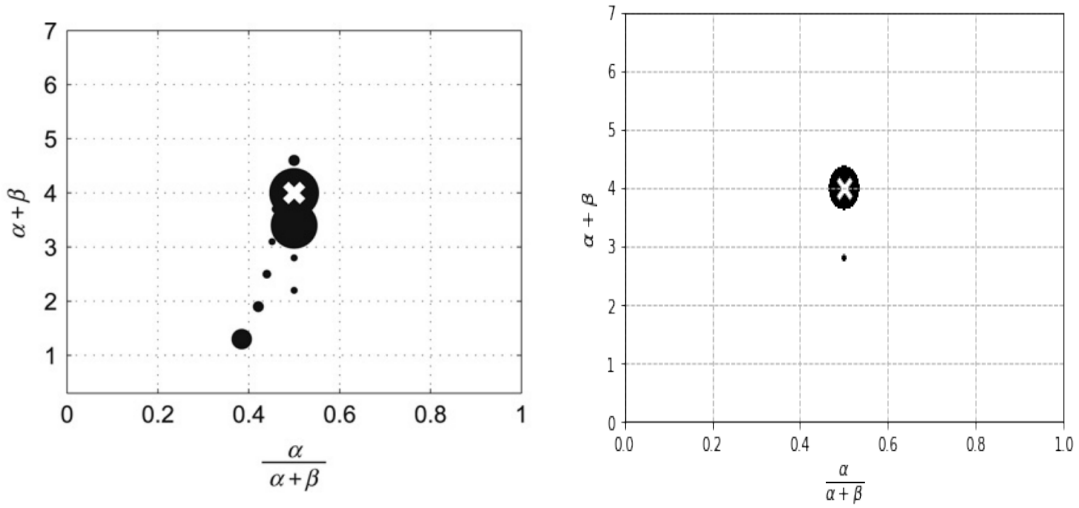
and 4.6b, which again shows perfect parameter recovery, and near perfect agreement with the results of Steyvers et al. (2009), albeit some minor differences in α and β values.



(a) Original

(b) Replicated

Figure 4.5: MAP estimates of accuracy of execution parameters



(a) Original

(b) Replicated

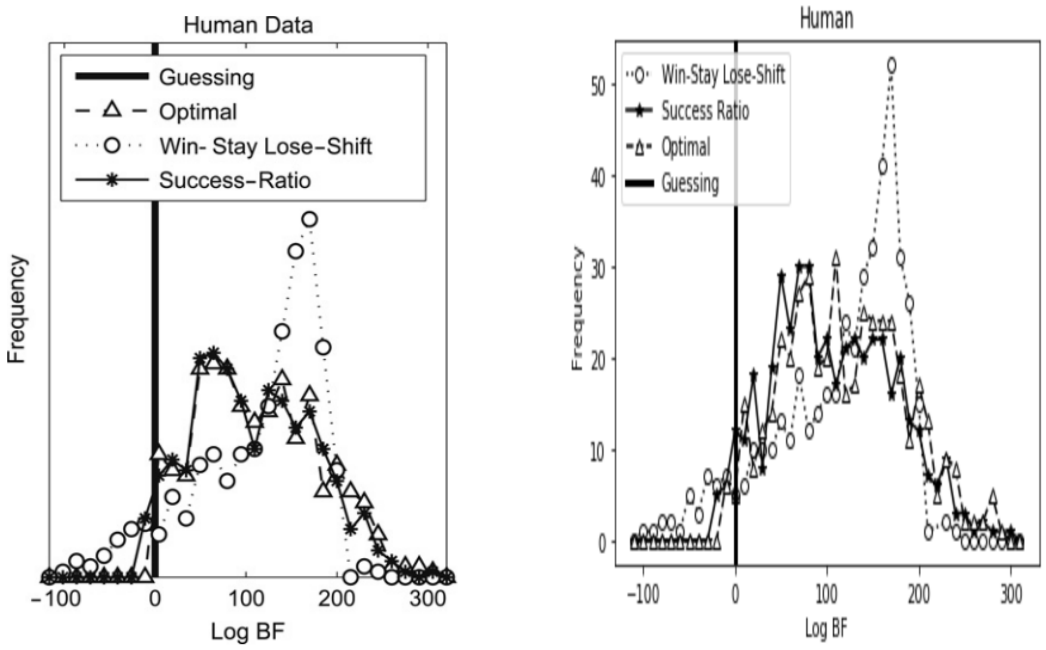
Figure 4.6: MAP estimates of α and β

Finally, Section 5 of the paper by Steyvers et al. (2009) reports the distribution of

the best model for the given participants, as well as the MAP estimate of the relevant parameters for each model, conditioned on it having the best support according to Bayes Factors. The division of participants according to the best supported model in both the original paper and the results found in this analysis is shown in Figure 4.8. It can be seen that the original results were replicated to a high degree of accuracy, with only minor differences in the final divisions. I suspect that these results may be due to minor differences in implementation of the likelihood functions of the models considered, particularly the optimal model. These differences are more evident in Figure 4.7. There is slightly more noise in the results of this analysis. The original values of the log BF measures were not reported, and as such, exactly how different the replicated results are from the original results can not be reported. However, a comparison of the distributions in Figure 4.7 shows that the overall shape and rank of each distribution is almost perfectly replicated.

I was also able to reproduce MAP estimates of the accuracy of execution parameters, as well as α and β values. Unfortunately, the true values of the estimates were not provided in the original paper, so the exact degree to which parameters were recovered is unknown, a comparison of Figures 4.9 and 4.10 shows that recovery seems quite accurate.

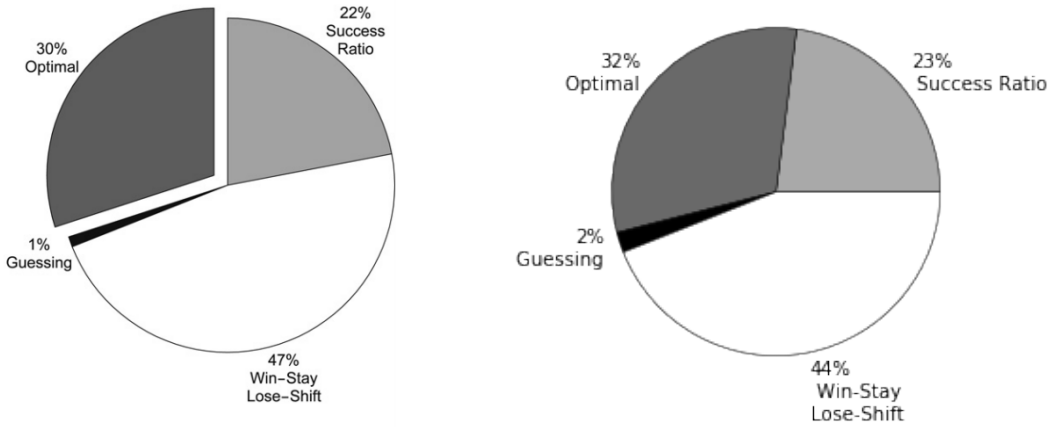
In summary, the original findings within a very small margin of error, with minor discrepancies in values, likely due to differences in implementations of the models analysed. The key findings of the original paper were confirmed: Bayes factors were shown to be a useful and worthwhile tool for comparing models in cognitive science as evidenced by their near perfect model identifiability and parameter recovery; individual differences of the same relative degree as originally reported for this data



(a) Original

(b) Replicated

Figure 4.7: log BF measures for models applied to the real decision-making data

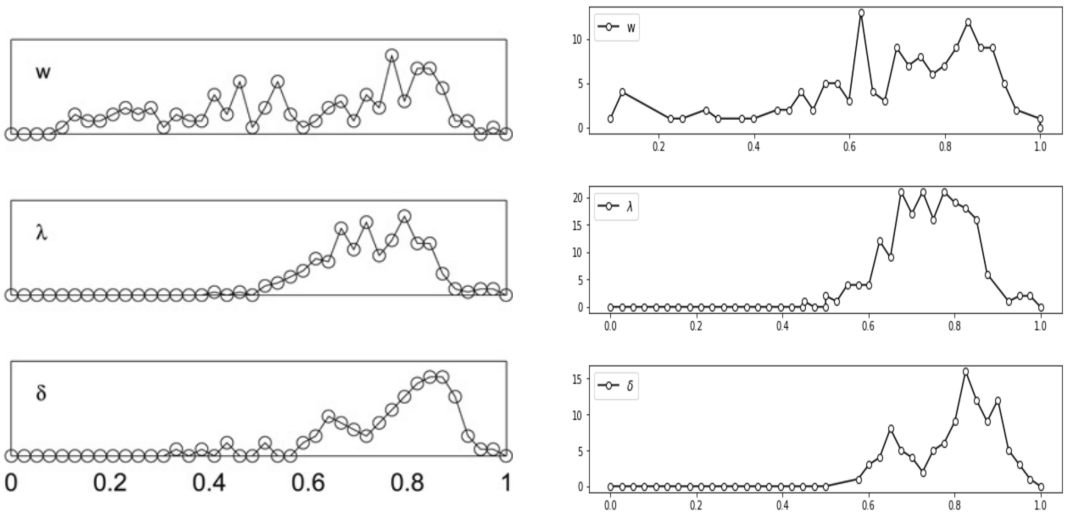


(a) Original

(b) Replicated

Figure 4.8: Division of participants according to model with largest support

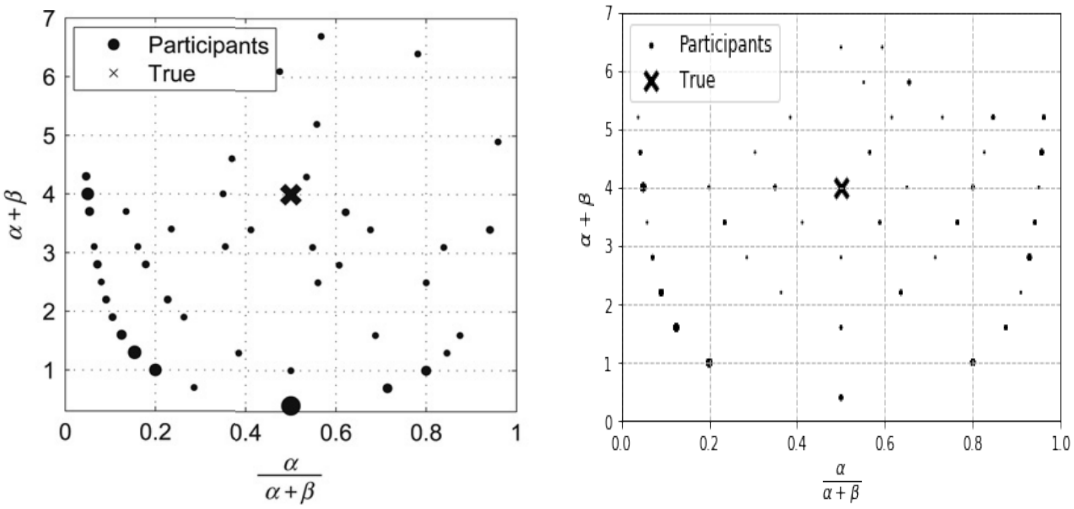
set were identified.



(a) Original

(b) Replicated

Figure 4.9: MAP estimates of accuracy of execution parameters



(a) Original

(b) Replicated

Figure 4.10: MAP estimates of α and β

4.2 Hierarchical Models

Motivation

Zhang and Lee (2010b) use the bandit problem in order to study the "wisdom of the crowds" i.e. the idea that the aggregate performance of a group of people on

a task is better than the performance of any given individual. In order to investigate this effect, the authors implemented four decision-making models: the ϵ-greedy, ϵ-decreasing, WSLS, and e-WSLS models. The models were implemented hierarchically so as to allow for individual differences in the key parameters of each model.

One problem identified during a reading this paper was that a Gaussian prior is placed on the psychological parameters in each model, and the value of the hyperparameters in each prior is not stated. In my attempts to overcome this lack of description, I placed an uninformative prior on these values: 0 for μ, and 1000 for the σ, standard practice in Bayesian prior setting. Doing so meant that negative probabilities were possible which, of course, they should not be. I considered taking the absolute value of the probability generated, however, I did not want to interfere with the values generated at such a direct level. In a paper by the same authors in the same year as Zhang and Lee (2010a), Zhang and Lee (2010b) again used Bayesian graphical models in their efforts to identify optimal parameters for experimental design of bandit problems which maximize the ability of the researchers to distinguish between models. In this paper, the authors analyzed performance of three decision-making models on bandit problems: WSLS, e-WSLS, and ϵ-greedy. The graphical models used in this analysis are exactly those described in Figures 3.1, 3.2, and 3.3, with Beta priors placed on the psychological parameters, as is normally done in bandit tasks. The authors placed a $Poisson(10)$ prior on all α and β values. Since the use of Beta priors is standard in bandit problems, and due to the missing details in the paper by Zhang and Lee (2010a) [6] I adapted the models by Zhang and Lee (2010a) to use Beta priors, as is done in the later study by the same

[6] I reached out to one of the authors, Michael Lee, to clarify the parameters used, however, the code was no longer available

authors. A further motivating factor for the unification of these two papers is that there is a large overlap between the models analysed: all of the models considered by Zhang and Lee (2010b) were also used by Zhang and Lee (2010a). In addition to this, the researchers also applied their models to the same data set, the same one used in the previous replication study by Steyvers et al. (2009), and the one used in my analysis. As a result, the parameter values estimated in both papers for each model could be compared: this comparison showed that there were discrepancies in estimated parameter values between the two papers, even though the same models and data were used - the results are shown in Table 4.2, alongside the results found during this replication. A possible reason for the discrepancies in results could be due to the different priors placed on the estimated parameters. Another possible explanation for the discrepancies in the results is the difference in the likelihood functions detailed for the models considered in each paper. The likelihood functions detailed by Zhang and Lee (2010a) are exactly those in Chapter 3 of this report, and are defined for four-armed bandit tasks, whereas the likelihood functions detailed by Zhang and Lee (2010b) are defined for two-armed bandit problems, and were applied to data from a four-armed bandit problem. Whilst it is, of course, very likely that the authors adapted their likelihood functions to account for the difference in number of arms, the precise likelihood functions are not available, and thus may differ from those detailed in Zhang and Lee (2010a).

One motivation for this analysis was to unify the findings of these two papers by using the likelihood functions defined by Zhang and Lee (2010a) - as they are correctly defined for four-armed bandit problems - and placing a Beta prior over

the psychological parameters of each model, as done by Zhang and Lee (2010b), as this is the case in all other papers dealing with cognitive models on bandit tasks (Steyvers et al., 2009; Zhang & Yu, 2013; Zhang & Angela, 2013). Finally, using a Beta prior overcomes the problem of negative probabilities presented by the use Gaussian priors.

Another motivating factor for reproducing the methods used in these papers, particularly those reported by Zhang and Lee (2010a), is that the a successful replication of their results will provide more credibility for the results of my later analysis, due to the similarities in the mixture model developed by Zhang and Lee (2010a) (3.7 in Chapter 3), and the model developed during this project (3.8).

Method

All four models, as well as the hierarchical mixture model, implemented by Zhang and Lee (2010a), and the three models by Zhang and Lee (2010b) were implemented according to their Bayesian graphical model - the models are those shown in Figures 3.1, 3.2, 3.3, 3.4 and 3.7. The authors reported in both papers that the models were implemented in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). In order to implement the graphical models, I used JAGS (Plummer et al., 2003) rather than WinBUGS, as more support is available for JAGS on MacOS. Since JAGS offers no functionality for processing of Markov chain Monte Carlo (MCMC) samples, postprocessing was carried out in R[7] using the rjags[8] package. Graphical models were implemented according to the formalism outlined in (Lee & Wagenmakers, 2014). During initialisation of the JAGS models, I encountered out of memory

[7]https://www.r-project.org/

[8]https://cran.r-project.org/web/packages/rjags/index.html

errors in R due to the large graph size of the JAGS models. In order to overcome these issues, I wrote a series of scripts to fit the models and process the results on the Eddie cluster, initialising a new node for each JAGS model.

Results

Table 4.2 shows the original mean and standard deviations (shown in brackets) reported by Zhang and Lee (2010a), and the results obtained during this replication. The table also shows estimates of the same parameters reported in (Zhang & Lee, 2010b) - where parameter estimates were not given in this paper, a $-$ was added. Figure 4.11[9] shows a plot of the parameters from Zhang and Lee (2010a) against the replicated results, where a dotted line indicates values from Zhang and Lee (2010a) and a dashed line indicates the replicated results - the results from the ϵ-decreasing model are not included in the plot, as the low standard deviation resulted in a very large density around the mean, larger than all other densities by a factor of 10, thus making it impossible to clearly distinguish the other results.

As shown can be seen from Table 4.2 and Figure 4.11, the original results from Zhang and Lee (2010a) for γ and γ^l were almost perfectly replicated. There was also a large overlap in the value of ϵ, as well as γ^w. The value of ϵ' was not well replicated. It is clear from Table 4.2 that the two papers disagree in their estimates of ϵ, and that the replicated value is much closer to the value reported by Zhang and Lee (2010b) than by Zhang and Lee (2010a). Again, this discrepancy is likely due to the different priors placed on the psychological parameters. I believe these different priors may also explain the difference in posterior estimates of ϵ', however,

[9]Colors are used in this plot to distinguish between different parameters

Parameter	(Zhang & Lee, 2010a)	(Zhang & Lee, 2010b)	Replicated results
γ	0.71 (.10)	0.7 (-)	0.705 (0.0966)
γ^w	0.99 (0.27)	$\approx$ 0.99 (-)	0.806 (0.194)
γ^l	0.59 (0.25)	$\approx$ 0.6 (-)	0.567 (0.216)
ϵ	0.24 (0.10)	0.3 (-)	0.314 (0.128)
ϵ'	0.61 (0.11)	- (-)	0.989 (0.0135)

Table 4.2: Comparison of means (and standard deviations) of the group distributions for each parameter in the four mixture models (- denotes the value wasn't reported)

as this model was not implemented by Zhang and Lee (2010b) and prior values were not provided for the models developed by Zhang and Lee (2010a), this cannot be confirmed. One further piece of evidence that differences in priors were the reason for the differences in posterior estimates are the results of the posterior estimates of the hyperparamters for the Beta prior placed on the models detailed by Zhang and Lee (2010b) and the estimates in this replication: these results are presented in Table 4.3, with the distributions plotted in Figure 4.12[10]. The ϵ-decreasing model is omitted from this analysis since it isn't subject to analysis of Zhang and Lee (2010b).

It is clear from Figure 4.12 that the choice of prior distribution used has a large effect on the posterior densities, with the replicated results being almost identical to the results report by Zhang and Lee (2010b).

Finally, Table 4.6 summarises the results from applying the hierarchical mixture model to the human decision-data, and reports the number of people inferred to be

[10]Colors are again used to distinguish between different parameters

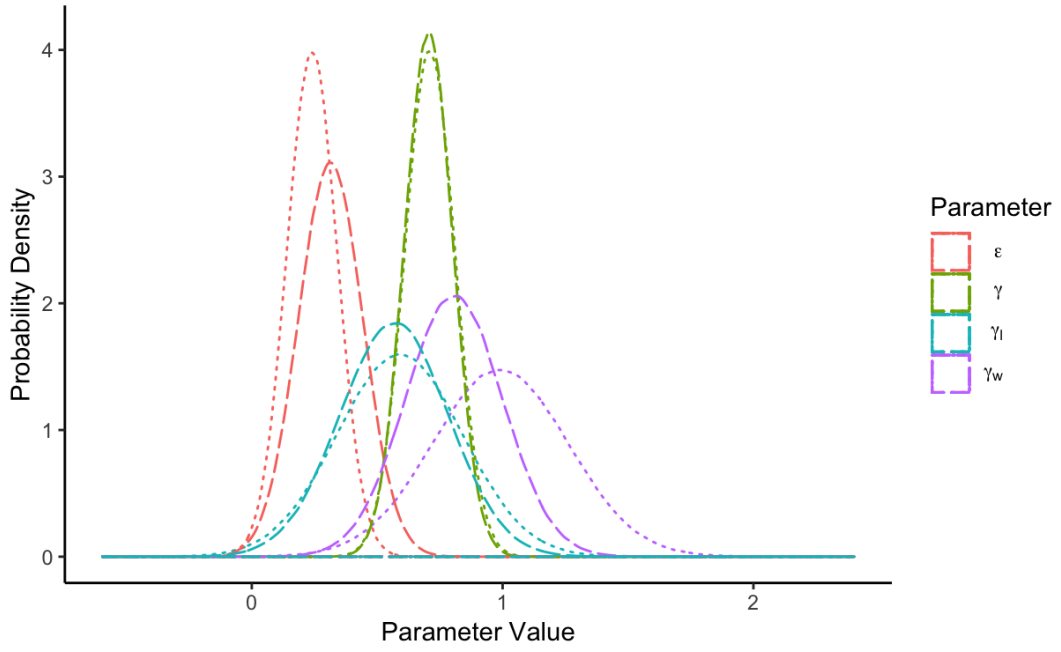


Figure 4.11: Plot of posterior estimates of psychological parameters in this analysis versus those reported by (Zhang & Lee, 2010a)

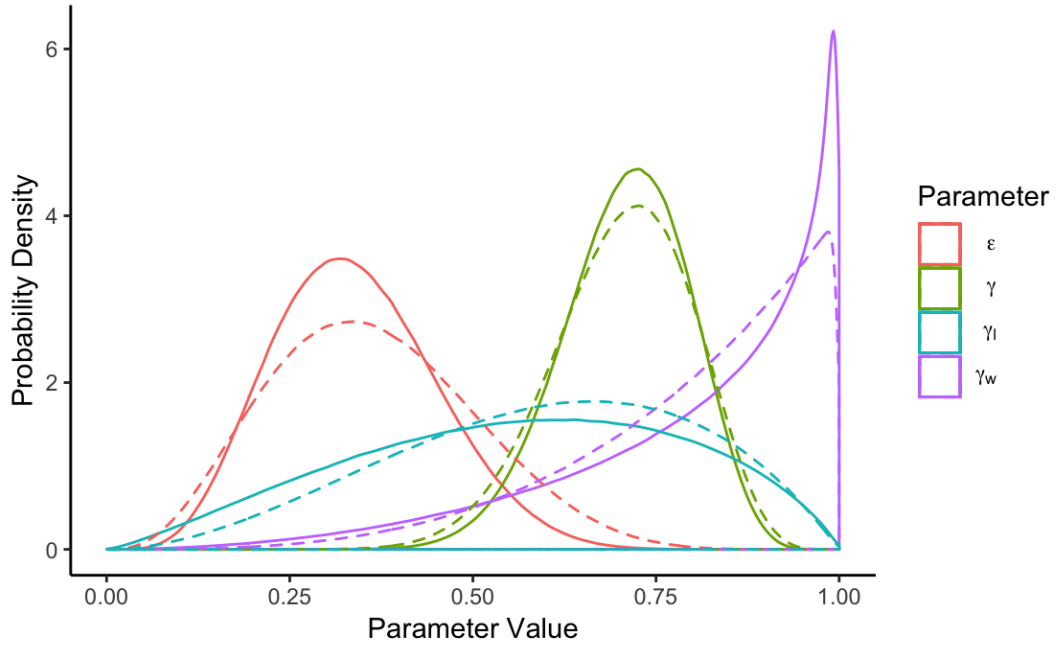


Figure 4.12: Plot of posterior estimates of psychological parameters based on hyperparameters in this analysis versus those reported by (Zhang & Lee, 2010b)

	(Zhang & Lee, 2010b)		Replicated results	
Parameter	α	β	α	β
γ	18.9	0.7.8	15.44	6.499
γ^w	2.9	0.7	4	1
γ^l	2.3	1.8	2.977	2
ϵ	5.8	11.3	4	7

Table 4.3: Comparison of means of the hyperparameters for the priors placed on psychological parameters

using each model in both the original paper (Zhang & Lee, 2010a) and this analysis.

Model	(Zhang & Lee, 2010a)	Replicated results
WSLS	0%	2.72%
Extended WSLS	75%	68.23%
ϵ-greedy	22%	27.7%
ϵ-decreasing	3%	1.32%

Table 4.4: Comparison of proporiton of participants using each model

There are slight differences in the distribution of participants to models in the replicated results when compared to the original, however, the proportion of participants to each model is of the same relative degree, and it is again likely that the differences in values are due to the different priors used in each model.

In order to test this theory, the priors reported by Zhang and Lee (2010a) were hard-code into the hierarchical mixture model, and the model was applied to the human decision-making data. The results of this experiment are shown in Table

4.5. Hard-coding the priors had a drastic effect on distribution of participants to each model, and results are not at all similar to the original results reported by (Zhang & Lee, 2010a). To understand this, it is worth examining the posterior estimates of each parameter in the replicated results after fitting on human data in Table 4.6; estimates of the parameters are very different when a mixture model is applied to the data. I therefore believe that the prior values reported by Zhang and Lee (2010a) would be different for the mixture model than for the individual hierarchical models. Unfortunately, these values are not reported in the original paper, so again this theory cannot be confirmed. Of course, these prior values are likely omitted since the part of the paper analysing the hierarchical model was interested in investigating model differences, not parameter differences.

Model	(Zhang & Lee, 2010a)
WSLS	43.65%
Extended WSLS	45.52%
ϵ-greedy	8.52%
ϵ-decreasing	3.14%

Table 4.5: Comparison of proportion of participants using each model with priors from (Zhang & Lee, 2010a) hard-coded into models

In summary, it is my belief that the differences in posterior estimates of the psychological parameters made in this analysis and those reported by Zhang and Lee (2010a) are due to the difference in prior distribution placed on the psychological parameters of each model, as evidenced by the almost perfect replication of posterior estimates reported by Zhang and Lee (2010b) which used the same prior distribution

Model	Individual model	Mixture model
γ	0.705 (0.0966)	0.626 (0.116)
γ^w	0.806 (0.194)	0.891 (0.096)
γ^l	0.567 (0.216)	0.500 (0.206)
ϵ	0.314 (0.128)	0.340 (0.130)
ϵ'	0.989 (0.0135)	0.762 (0.0855)

Table 4.6: Comparison of means (and standard deviations) of each parameter after applying individual and mixture models

as is used in this analysis. I also believe that the differences in the proportion of participants using each model in the replication versus the original paper (Zhang & Lee, 2010a) is also due to the difference in prior distribution used. Finally, as the relative degree of participants to each model was similar in this analysis to those reported by Zhang and Lee (2010a), and the posterior estimates were nearly perfectly replicated when the same prior distribution as used by Zhang and Lee (2010b) was used for each model, I can be confident in the results of my subsequent analysis which use these models.

Chapter 5

Model Analysis

5.1 Motivation

Steyvers et al. (2009), the subject of the replication study in Chapter 4, mention in their conclusion that an "important way to extend the work we report here is to consider a richer and larger set of possible models as accounts of people's decision-making." Additionally, Zeigenfuse and Lee (2009) state that "Future bandit problem work should focus on evaluating numbers of different heuristic models, and partitioning participants into groups to capture variations in the way those models are applied, using accounts of individual differences similar to those presented [in this paper]". It is precisely this analysis that is described in this section: both a larger set of cognitive models were considered, as well as testing of newly developed models detailed in Chapter 3.

Lee et al. (2009) mention that a motivating factor for their research was to develop a model which captures latent states. The researchers believe that a latent state model is useful in bandit tasks, as this type of model has demonstrated its

usefulness in other tasks in cognitive science. With this as their motivation, they develop a latent state model which contains a mixture of two different states: exploration and exploitation. A motivation for my analysis was to develop a model which would extend the use of latent states to change the fundamental strategy by which participants make decisions - changing on both a per-game and per-trial basis - thus allowing for a richer account of how humans can vary in the strategies they use. An aim of this analysis was therefore, to determine whether this model gave a better account of human decision-making data than simple latent state models did. Another motivation was to test the finding of Lee et al. (2009) that the τ-switch model gave the best account of human data, by applying it to a larger data set. The authors also believed the τ-switch model acts as a good heuristic for generating optimal data. It is therefore an aim of this paper to test this finding, and test whether any of the newly proposed models also give a good approximation to optimal data.

This analysis is split into two parts: the first section deals with applying the framework outlined by Steyvers et al. (2009) to a larger class of cognitive models in order to test model identifiability, as well identifying individual differences in parameter and model choice after applying the larger class of models to human decision-making data; the second section follows the methodology for analysing hierarchical Bayesian models outlined in the literature (Zhang & Lee, 2010b, 2010a; Lee et al., 2011) in order to analyse individual differences in model and parameter choice - comparing these results to those found in the first stage of the analysis - as well as analysing agreement of various hierarchical Bayesian models with human and optimal data. Comparisons are also drawn between the posterior distributions

of key psychological parameters estimated by applying the models to both human and optimal decision-making data in order to draw conclusions about human and optimal performance, as well as determining whether any new models can be used as a computationally tractable heuristic to the optimal model.

5.2 Non-Hierarchical Models

Model Identifiability

The first stage of the analysis consisted of performing an artificial recovery study in order to test model identifiability. Testing model identifiability is important as an initial sanity check to ensure models are correctly implemented and will be useful: (Farrell & Lewandowsky, 2018) report that frequently, when a model is non-identifiable, it is rarely of use, particularly for cognitive models whose analysis involves the summarization of data via interpretation of key psychological parameters; if the parameters are not identifiable, the model is of limited value.

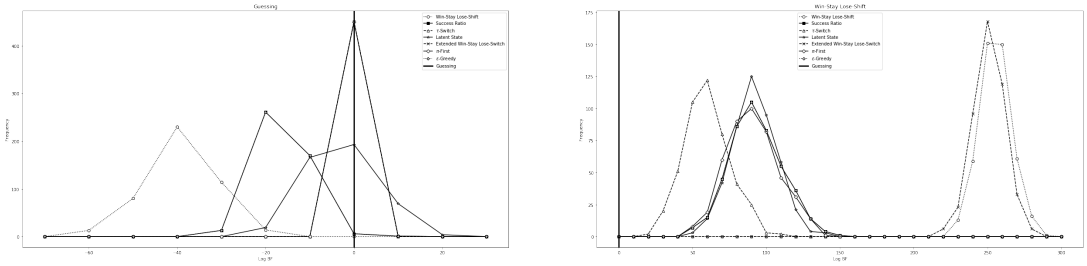
In order to conduct this analysis, the class of models implemented in Scala - detailed in Chapter 4 - was extended to include the π-first (Equation 3.6), τ-switch (Eq. 3.7), and Latent State models. For each model, an artificial data set was generated, consisting of 451 participants, each completing 20 games consisting of 15 trials in ordet to reflect the setup in the human decision-making data gathered by Steyvers et al. (2009). The marginal likelihood of that data set was then calculated under each model, with the results shown in Figure 5.1. Table 5.1 summarises the model recovery performance, showing the proportion of simulated participants inferred to belong to each model for each of the generating data sets. Although data was available for the Optimal model for some generating data sets - namely

Guessing, WSLS, ϵ-Greedy, and Optimal since these were generated for the analysis Chapter 4 - the model was too computationally expensive to run on the additional data sets, and so is excluded from this section of the analysis so that all results involve the same set of models.

Recovery Model	Guessing	WSLS	e-WSLS	ϵ-Greedy	π-first	Latent State	τ-switch
Guessing	61	0	1	0	2	35	1
WSLS	0	100	0	0	0	0	0
e-WSLS	0	100	0	0	0	0	0
ϵ-Greedy	0	0	0	100	0	0	0
π-first	0	0	0	0	100	0	0
Latent State	0	0	0	0	0	100	0
τ-switch	0	0	31	0	0	48	21
Optimal	0	0	0	100	0	0	0

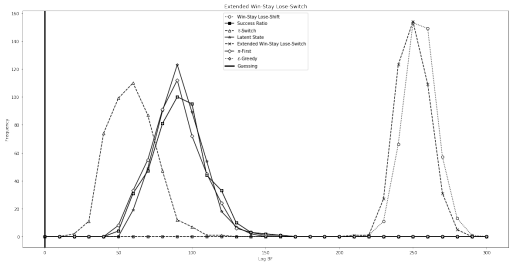
Table 5.1: Model recovery on artificial data

As can be seen from Figures 5.1 and Table 5.1, the ability of the log BF measure to identify the underlying decision model is mixed. In the case of the WSLS, ϵ-Greedy, π-First, and Latent State models, the correct model always had the most support. In the case of the e-WSLS model, the correct model was always assumed to be the WSLS model. I believe this can be explained due to the fact that the e-WSLS model is more flexible, and therefore more complex, than the simpler WSLS model: when calculating the marginal density of the e-WSLS model, accounting

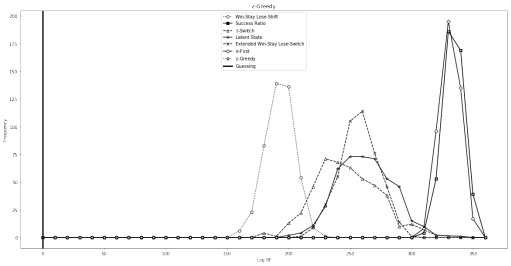


(a) Guessing

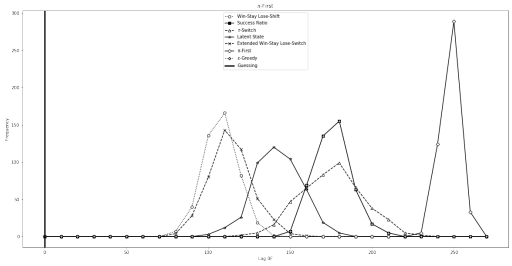
(b) WSLS



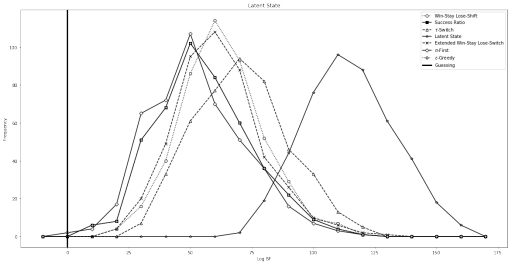
(c) e-WSLS



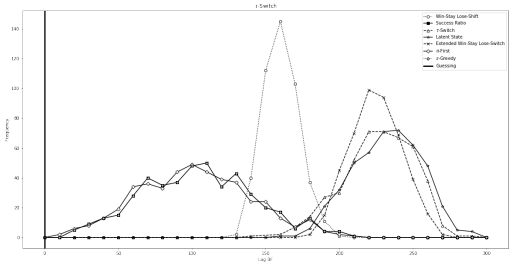
(d) ϵ-greedy



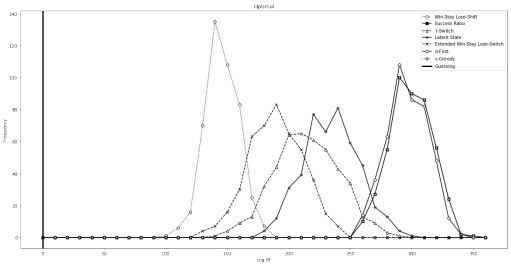
(e) π-first



(f) Latent State



(g) τ-switch



(h) Optimal

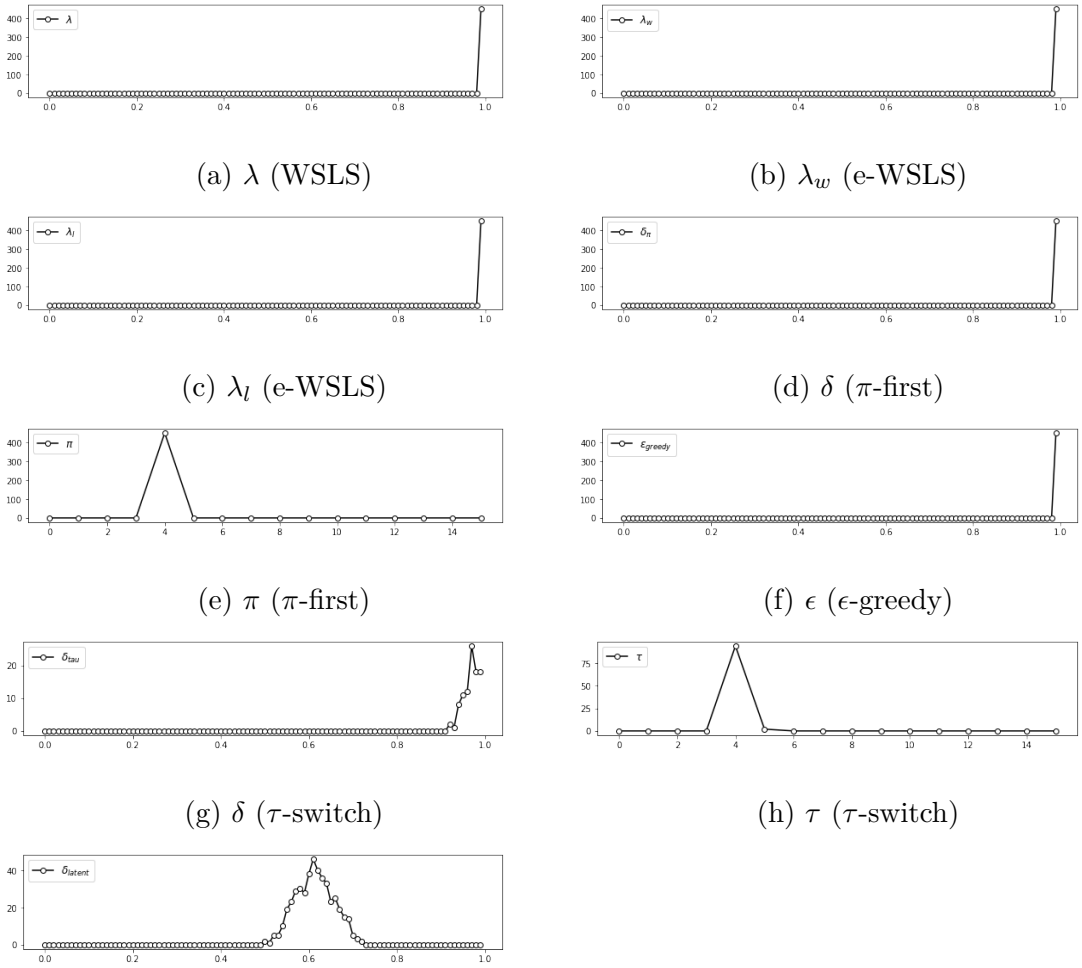
Figure 5.1: log BF measures for models applied to the artificial decision-making data

for the additional parameter would lead to a slightly lower marginal density. Since the log BF measures of the WSLS and e-WSLS are so similar, as can be seen in Figure 5.1b, I believe that the slight preference is given to WSLS due to its simpler

marginal density function. I believe the same argument holds true for the τ-switch model: 48% of artificial data points generated by the τ-switch model are assumed to have been generated by the latent state model. The τ-switch model and the latent state model use the same underlying decision-making method - with the τ-switch model being developed as a simplification of the latent state model. The latent state model is parameterised by a latent state variable z, and an exploration parameter δ; the τ-switch model is also parameterised by an exploration parameter δ, while the second free parameter controls the trial on which to switch from exploration to exploitation. During calculation of the marginal densities of each model, 40 values of δ are considered for each model, however, only two values are considered for z - since only two latent states are considered - whilst k values are considered for τ, where k equals the number of trials per game. I believe that the added complexity due to the larger domain of values to be considered for τ compared to z caused the latent state model to be given such a high proportion as the marginal density for the τ-switch model is calculate over a larger domain, leading to a smaller value; this reasoning is further supported by noticing the similarity in log BF measures of the two models in Figure 5.1g. As for the 31% of participants assigned to the e-WSLS model, I am unsure as to the reason for this high value, and can only speculate as to its occurrence, perhaps due to the flexibility of the model. The guessing model also showed poor recovery, with 35% of participants assigned to the latent state model. In the original paper, Steyvers et al. (2009) reported that the Guessing model was only recovered in 98% of cases, with 2% assigned to the Optimal model. They don't report a reason for this, nor can I, and I can only speculate that the sophistication of the Optimal and Latent State models allowed for a signal to be found in the noise of

the random Guessing data. Model recovery for the ϵ-greedy model is perfect, whilst the Optimal data is consistently thought to have been generated by the ϵ-greedy model. Since I had the results for applying the optimal model to the artificial data generated by the optimal model, I compared the log BF measures for the ϵ-greedy and Optimal models on this data set, and found that the log BF measure for the Optimal model was higher than that of the ϵ-greedy model for all 451 data points, indicating perfect recovery for the optimal data.

Identifying the correct underlying model was only one part of this identifiability analysis, with the arguably more important analysis being that of parameter recovery. Figure 5.2 shows the distribution of MAP parameter estimates over the simulated subjects, conditional on the correct model being recovered. It is clear that these parameters were perfectly recovered in all models except for the τ-switch and latent state models, however, the τ parameter of the τ-switch model is perfectly recovered, and the mean value of δ is 0.968, very close to the original value of 1.0. The value of δ in the latent state model is not recovered at all. This, coupled with the high degree of the Latent State model in the Guessing and τ-switch models in Table 5.1 lead me to have doubt in the reliability of the Latent state model. With these results in mind, as well as accounting for the similarity in the latent state and τ-switch models, and the good recovery of the τ-switch model, the latent state model was omitted from the final analysis on the real participant data, with the high identifiability of the other models allowing me to have confidence in the results of said analysis.



(a) λ (WSLS)

(b) λ_w (e-WSLS)

(c) λ_l (e-WSLS)

(d) δ (π-first)

(e) π (π-first)

(f) ϵ (ϵ-greedy)

(g) δ (τ-switch)

(h) τ (τ-switch)

(i) δ (Latent State)

Figure 5.2: MAP Estimates on artificial data

Individual Differences

The next stage of the analysis involved applying the models to the real participant data in order to determine whether the new models gave a better account of human decision-making than the models considered by Steyvers et al. (2009). The log BF measures are shown in Figure 5.3, with Figure 5.4 detailing the proportion of participants deemed to be using each model. Note that in Figures 5.3 and 5.4, data from the optimal model has been added to this analysis, since this data was gathered during the replication conducted in Chapter 4, and allows for a more

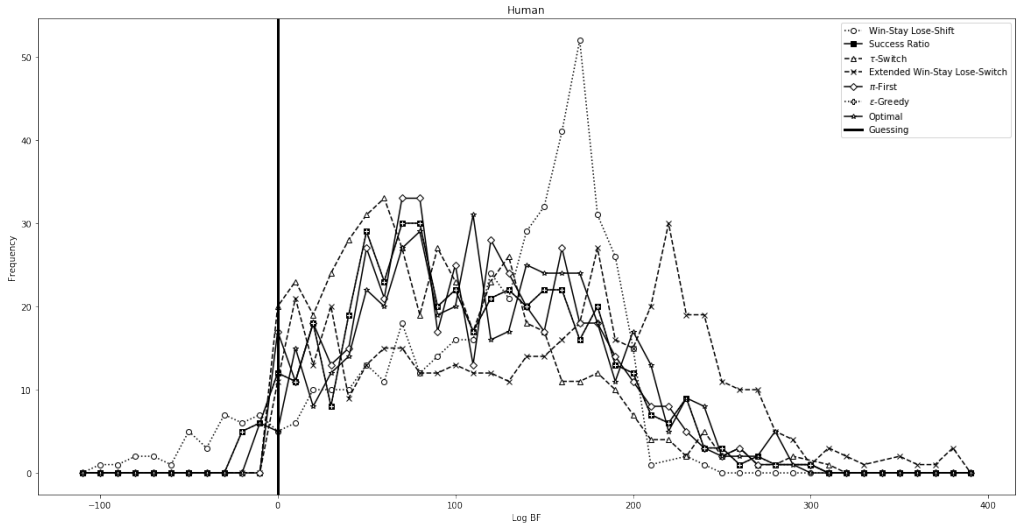


Figure 5.3: log BF measures on Human data

complete analysis of human decision-making.

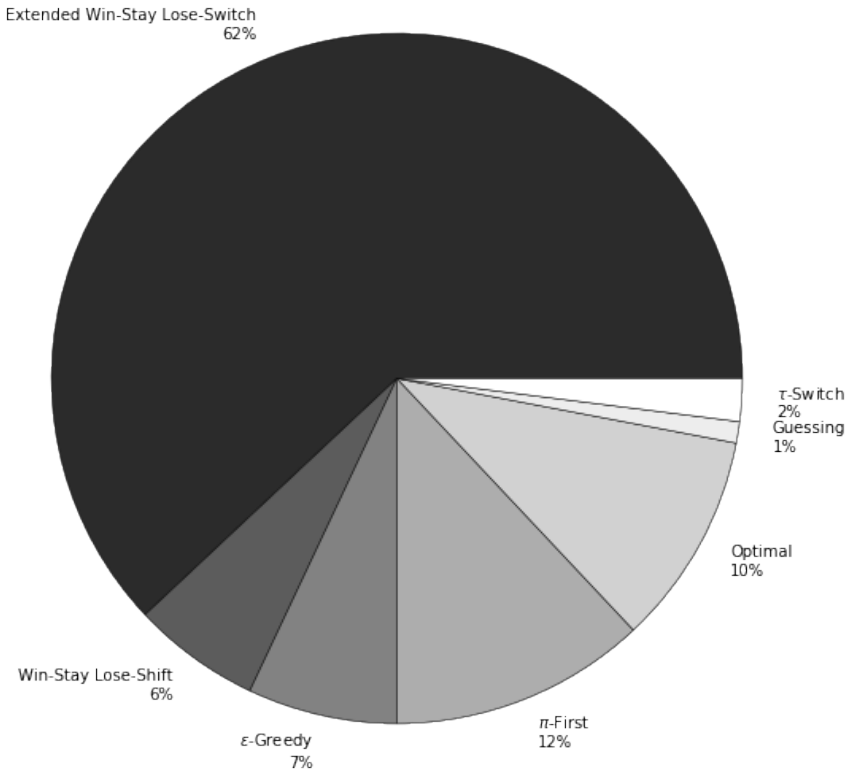


Figure 5.4: Proportion of participants using each model according to maximum value of log BF measure

In the original paper, decisions of 47% of participants were best described by the WSLS model. Results in Figure 5.4 show that the newly included e-WSLS model was thought to account for decision-making data for 62% of participants, with 6% now subscribing to the original, simpler, variant. The number of participants subscribing to the optimal policy has now dropped from 30% to 10%, and whereas in the original paper 22% of participants were using the ϵ-greedy - dubbed the Success Ratio model in the original paper by (Steyvers et al., 2009) - the extended analysis determines only 7% of participants using that model. The addition of the π-first model proved to be worthwhile, with 12% of participants deemed to be using it. One final point worth noting is that only 2% of participants were thought to be using the τ-switch model. Lee et al. (2009) report that the τ-switch model gave a good account of human decision making for most participants in their study, and whilst it is clear from Figure 5.3 that this model was able to give a good account of participants behaviour, it by no means gave the best account, as evidenced by the results of Figure 5.4. There multiple reasons for this: first and foremost, Lee et al. (2009) by no means claim that this model gives the best account of human decision making, just that it gives a "good account" - with not all participants in the original study being best described by this model - which has been shown here; secondly, the original τ-switch model was developed for application in two-armed bandit tasks, with the original researchers concluding that "the τ-switch model is a useful addition to current models of finite-horizon **two-armed bandit problem decision-making**"[1] (Lee et al., 2009), whereas the variant used in this analysis was developed to work with an arbitrary number of arms, and applied to

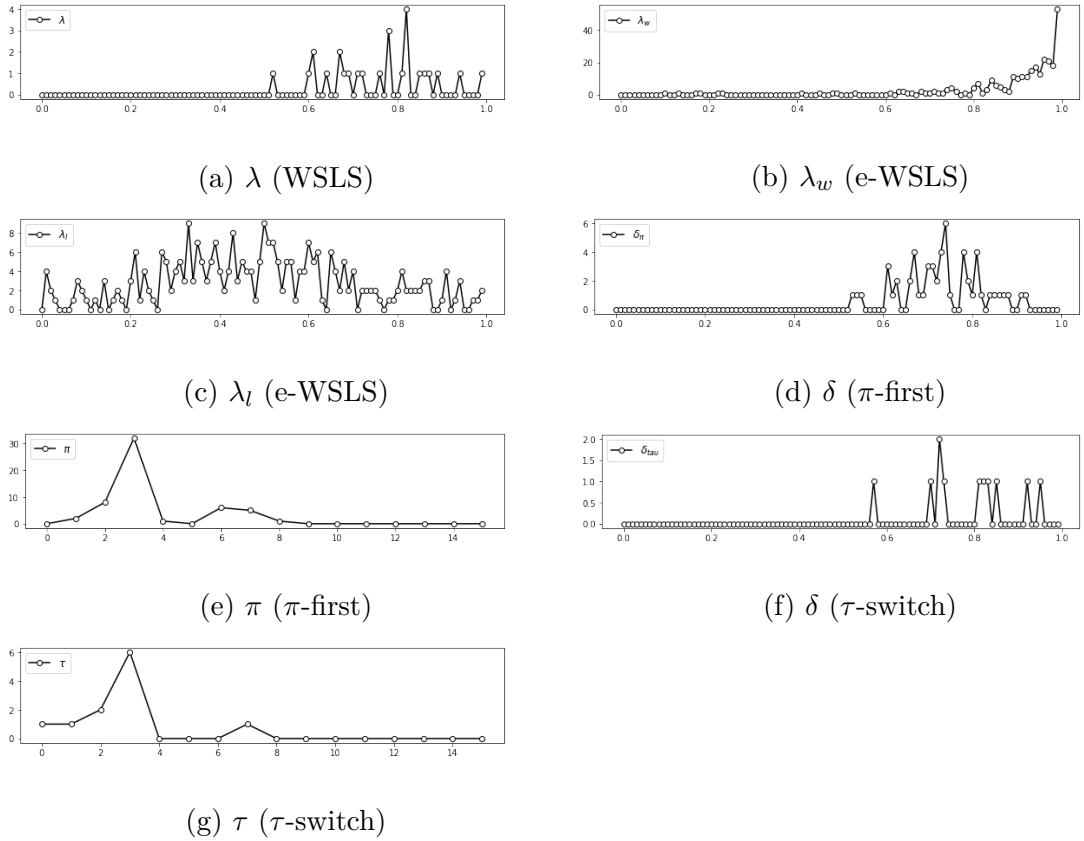
[1] bold text added by me

a data set of decisions made on four-armed bandit tasks, not two as in the original study. Finally, the original paper applied their data to a set of 10 participants who completed 50 bandit problems across six different setups: the setup varied the number of trials per game, using 8-trial and 16-trial setups, and varied the reward rate to simulate "plentiful" ($Beta(4, 2)$), "neutral" ($Beta(1, 1)$), and "scarce" ($Beta(2, 4)$) environments. In contrast, the data used in this analysis consisted of 451 participants, each of whom completed 20 bandit problems with 15 trials, all of which were distributed according to a $Beta(2, 2)$ reward rate. These differences may explain why the number of participants using the τ-switch model in this analysis were of a different magnitude than in the previous study, and highlights the need for studies such as this which apply previously developed models to new data.

Figures 5.5 show the distribution of MAP parameter estimates, conditional on the model being the one with the largest support from the log BF measure. The results show that, with the exception of the the λ_l parameter of the e-WSLS model, the inferred accuracy of execution was generally quite high.

5.3 Hierarchical Models

In the second stage of the analysis, the graphical models detailed in Chapter 3 were implemented in JAGS. In each analysis, models were fit on decision-making data gathered by Steyvers et al. (2009), and 1,000 samples from 2 chains were collected, after a burn-in period of 1,000 samples. The justification for this decision was two-fold. Firstly, the same set up was used by Zhang and Lee (2010a), and since all of the models considered in that paper were also considered in this analysis, using the same setup allowed for a more direct comparison of results. Secondly, the graph



(a) λ (WSLS)

(b) λ_w (e-WSLS)

(c) λ_l (e-WSLS)

(d) δ (π-first)

(e) π (π-first)

(f) δ (τ-switch)

(g) τ (τ-switch)

Figure 5.5: MAP Estimates on human data (estimates for ϵ-greedy are missing since it wasn't found to account for human data)

size of later hierarchical mixture models was very large - some surpassing 34 million nodes, with a training time of over 24 hours[2] for certain models on a 32-core Eddie node. The size of the model after fitting was also as large as 6GB, and so memory limitations were also of concern. It would, of course, be ideal to have used a larger sample size and more chains, but due to the time limitations of this project, this was not feasible. Postprocessing of MCMC samples was carried out in R, again completed on nodes of the Eddie compute cluster to overcome initial memory errors encountered during analysis on my local machine.

[2]clock time

Parameter Differences

The first result worth reporting is a summary of the individual differences in posterior estimates of parameters used in each of the cognitive models; these results are summarised in Table 5.2.

Parameter	Individual Model	Mixture Model	Extended Mixture Model
ϵ-Greedy	0.314 (0.0633)	0.340 (0.130)	0.196 (0.0731))
ϵ-Decreasing	0.989 (0.0135)	0.762 (0.0855)	0.687 (0.104)
WSLS	0.705 (0.0966)	0.626 (0.116)	0.670 (0.102)
Extended-WSLS Win	0.806 (0.194)	0.891 (0.096)	0.906 (0.0775)
Extended-WSLS Lose	0.567 (0.216)	0.500 (0.206)	0.492 (0.202)
π-first Pi	6.30 (4.97)	-	8.00 (4.34)
π-first Gamma	0.179 (0.145)	-	0.229 (0.102)
Latent State	0.699 (0.140)	-	0.450 (0.149)
τ-switch Gamma	0.676 (0.162)	-	0.550 (0.132)
τ-switch Tau	8.58 (4.89)	-	7.96 (4.33)

Table 5.2: Means and standard deviations of the group distributions for each parameter in each hierarchical model

As was seen in section 4.2, combining the heuristic models in a mixture model had a large effect on posterior estimates of psychological parameters. This is better explained when viewed in conjunction with the differences in group estimates shown in Table 5.3: if a participant is thought to subscribe to the WSLS model alone, then the parameter must reflect the best estimate for all participants, whereas estimates from the original Mixture Model and Extended Mixture Model where posterior es-

timates are based upon decisions made only by participants who are inferred to using a given model. In later sections of this report, the posterior predictive agreement with model and human data is calculated, which gives a better idea of which parameter estimates giver a better account of human data.

For readers unfamiliar with the concept of Bayesian posterior estimates, it may seem natural to want to compare the results in Table 5.2 with the MAP estimates found in section 5.2, however, these estimates are generated according to entirely different methodologies. MAP estimates are generated by estimating the parameter values which best explain a set of data, whereas Bayesian posterior estimates are generated by considering the range of possible values, and weighing those according to their likelihood of occurring, informed by the data seen. In this manner, Bayesian posterior estimates give a truer estimate of the range of possible values which are likely to be seen based on prior information. Uniform priors were placed on parameters in the Bayesian models, and researchers have pointed out that this may not be a plausible approach (Steyvers et al., 2009). It would be interesting to place priors which assume participants are likely to adhere to a policy with high accuracy of execution - as well as other prior values - and examine the effect this has on posterior estimates.

Model Differences

In contrast to the previous section of this report which examined the differences in posterior estimates of model parameters, this section will discuss findings in model differences i.e. the degree to which participants vary in the fundamental decision-making process they use. To examine this, the mixture models outlined in Figures 3.7 and 3.8 were applied to the human decision-making data, and inferences on

the posterior values of the mixture parameters were drawn in order to estimate the proportion of participants assigned to each model. These results are shown in Table 5.3.

Model	Mixture Model	Extended Mixture Model
WSLS	2.72%	3.07%
e-WSLS	68.23%	65.1%
ϵ-Greedy	27.7%	8.57%
ϵ-Decreasing	1.32%	0.568%
π-First	-	17.2%
Latent State	-	0.350%
τ-Switch	-	5.10%

Table 5.3: Proportion of participants using each model in two mixture models

It is interesting to note the the WSLS and e-WSLS models give almost identical accounts of human behaviour in both mixture models, indicating the presence of a group of participants who follow the WSLS model and its variant. If it is to be believed that one group follows the WSLS family of models, then another can be thought to subscribe to the ϵ mode of thinking: in the mixture model consists of four heuristics, 29% of participants adhered to one of the ϵ models, with 26% subscribing to one of the ϵ models - now including the π-first model - in the Extended Mixture Model. The remaining 5.45% of participants are thought to be using on the newly included latent state models, with the majority of participants opting for the simple τ-switch alternative. These results show that π-first and τ-switch models are useful models for the study of sequential decision making.

In addition to this, the extended mixture models detailed in section 3.12 of this report were also applied to the data, and the same posterior estimates on mixture parameters were drawn. The degree to which these models agree with human decision-making data is detailed in the next section of this report, and for now model differences are summarised in Table 5.4.

Model	Mixture Model (participant)	Mixture Model (game)	Mixture Model (trial)	Extended Mixture Model (participant)	Extended Mixture Model (game)	Extended Mixture Model (trial)
WSLS	3%	6%	0%	3%	3%	0%
e-WSLS	68%	64%	72%	65%	61%	71%
ϵ-Greedy	28%	12%	26%	9%	1%	12%
ϵ-Decreasing	1%	18%	2%	1%	9%	11%
π-First	-	-	-	17%	19%	0%
Latent State	-	-	-	0%	0%	0%
τ-Switch	-	-	-	5%	7%	6%

Table 5.4: Proportion of participants using each model in extended mixture models allowing for latent model switching

The results in Table 5.4 show that certain models may be used more than initially thought. The results in Table 5.4 show that ϵ-greedy model was used by 9% of participants in the Extended Mixture Model which does not allow for latent model switching (column 5 of Table 5.4), whereas allowing participants to use it for a

limited number of games/trials - shown in the final 2 columns - shows that it is used by 1% and 12% of participants, respectively. This result can be seen across the range of models analysed. Whereas simpler accounts of individual differences might say that a participant never uses a certain model, extended models such as these allow for descriptions of decsision making such as "a single participant follows the latent state strategy in 80% of games, however, we can see that now and again they opt for the simpler guessing approach". I believe these models better capture the hierarchical structure described by (Mehlhorn et al., 2015), and allows for other factors to be considered as is asked for by (Cohen et al., 2007): perhaps the small set of problems where a guessing strategy is used are the final trials in an hour long experiment, and the participant has grown tired. In any case, a more flexible understanding of human performance will allow for more factors to be considered, and more meaningful conclusions to be drawn.

Results in Table 5.4 show the degree to which participants use a certain model, however the natural question to ask is whether these extended models give a better account of human decision-making than previous heuristic and hierarchical mixture models do. This is the focus of the next section of this chapter.

Characterization of human decision-making

Throughout this analysis, many methods and modes of analysis have been applied to human data in an attempt to explain human decision making. One suggestion was to marginalise over key parameters in the likelihood functions of the various cognitive models, and use this marginal density to generate a Bayes Factor that a participant is subscribing to one model rather than another; by comparing the log BF values, we were able to draw inferences about which model was most likely

used by each participant. In the second stage of this analysis, a hierarchical mixture model was constructed which allowed participants to vary in the model used. By analysing the posterior estimates of the mixture parameter, we were again able to draw inferences on which model was most frequently used. Finally, the previous section discussed extensions of two mixture models which allowed participants to change their model choice on a per-game and per-trial basis, thus allowing for an extensive analysis into the individual differences in model choice for a group of individuals. With all the results in place, an interesting question worth asking is which of the models considered gave the best account of human decision making. In order to answer this question, the posterior predictive agreement of model decisions with human data was calculated. Posterior prediction is used here as it is a standard approach in assessing the descriptive ability of Bayesian models, with wide use in statistics (Gelman et al., 2013) and recent adoptions in the cognitive sciences (Shiffrin et al., 2008). The method compares observed data to a posterior predictive distribution over the entire data space, which is calculated by taking the predictions made by a model at all possible parameter settings, and weighing those predictions by the posterior probability of each parameter setting. As highlighted by Lee et al. (2011) "An important property of posterior predictive methods is that they automatically balance goodness-of-fit with model complexity in their evaluation". This is due to the fact that flexible - and complex - models are able to adapt the parameters used to give a good account of the observed data, however, as the posterior predictive methods takes predictions made at all possible parameter settings, and characterises the average behavior of the model, the method naturally controls for model complexity.

The results of this analysis are summarised in Table 5.5 which shows test statistics on the posterior predictive agreement of each model considered thus far with human decision-making data. Figure 5.6 shows the proportion of models which have the highest value for posterior predictive agreement with human data.

Model	Minimum	Maximum	Mean	Standard Dev.
ϵ-Greedy	0.25	0.84	0.57	0.13
ϵ-Decreasing	0.24	0.83	0.57	0.12
WSLS	0.13	0.83	0.56	0.13
Extended-WSLS	0.25	0.96	0.62	0.17
π-First	0.25	0.83	0.55	0.11
Latent State	0.24	0.86	0.59	0.13
τ-Switch	0.24	0.87	0.59	0.13
Mixture Model (per participant)	0.26	0.96	0.64	0.15
Mixture Model (per game)	0.27	0.96	0.66	0.14
Mixture Model (per trial)	0.28	0.95	0.67	0.15
Extended Mixture Model (per participant)	0.25	0.96	0.64	0.14
Extended Mixture Model (per game)	0.27	0.95	0.67	0.14
Extended Mixture Model (per trial)	**0.32**	**0.94**	**0.68**	**0.15**

Table 5.5: Summary of test statistics from posterior predictive agreement analysis

It can clearly be seen from the results above that the Extended Mixture Model which allows for participants to change their decision-making strategy on a per-trial basis gave the best account of human decision-making data: 51.9% of participants were inferred to be using this model, and the model gave the largest posterior predic-

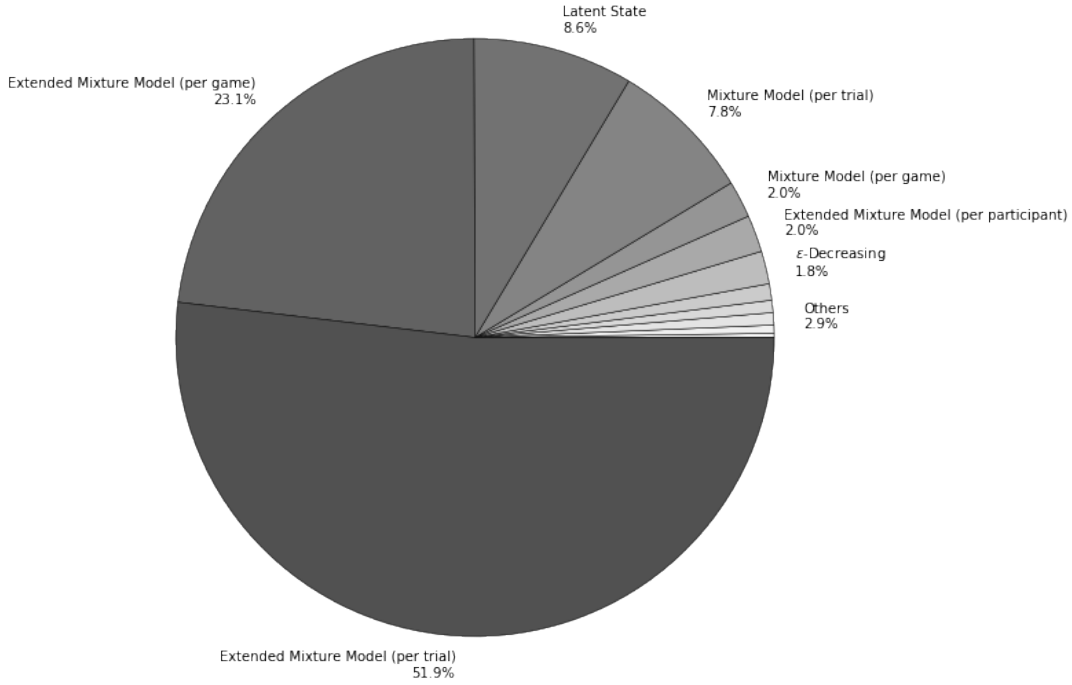


Figure 5.6: Percentage of participants whose decisions had largest posterior predictive agreement with a given model

tive agreement, on average, with human decision-making data. What is particularly interesting to note from Table 5.5 is that in both the Mixture Model and Extended Mixture Model, each subsequent increase in the ability of the model to capture individual differences - from a per-participant, to per-game, and finally to a per-trial basis - lead to an increase in the average posterior predictive agreement values with the data. Finally, it is worth noting that in Figure 5.6, 95.3% of participants decisions are accounted for by six models, all of which are a type of latent state model.

Characterization of optimal decision-making

The final analysis conducted was an investigation into the possibility of using the results found to teach humans how to behave in various tasks, and to determine if any models could be used a computationally tractable heuristic to the optimal model. To do this, models were applied to optimal decision data - generated using the Scala code detailed in section 4.2 of this report - and inferences on parameter estimates were drawn. The first part of this analysis is concerned with determining which models give the closest account of optimal data. To determine this, the posterior predictive agreement of decisions made under each model with optimal decision data was calculated, and results shown in Table 5.6 and Figure 5.7.

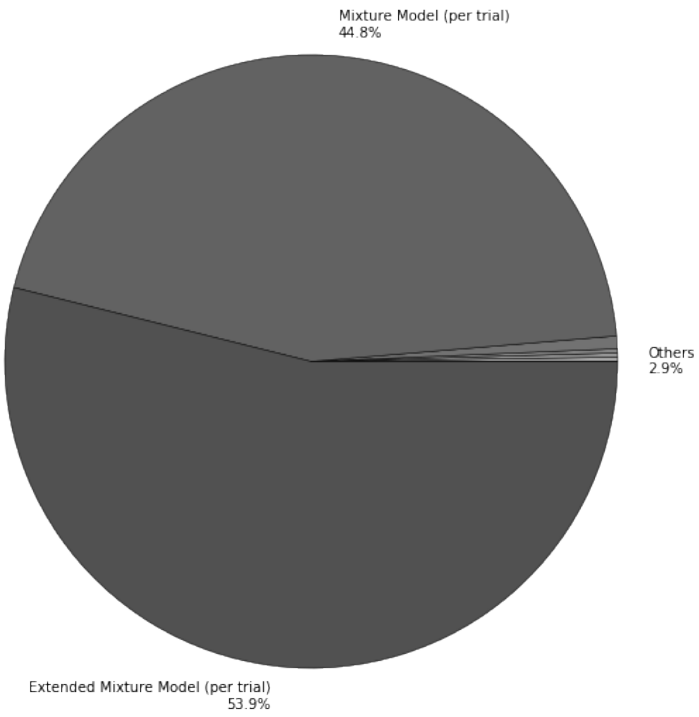


Figure 5.7: Percentage of optimal data points whose posterior predictive agreement was largest under a model

Lee et al. (2009) reported that the the τ-switch model could be used as a sur-

Model	Minimum	Maximum	Mean	Standard Dev.
ϵ-Greedy	0.34	0.69	0.52	0.06
ϵ-Decreasing	0.33	0.69	0.53	0.06
WSLS	0.57	0.71	0.64	0.02
Extended-WSLS	0.66	0.83	0.76	0.03
π-First	0.4	0.68	0.54	0.05
Latent State	0.43	0.72	0.59	0.05
τ-Switch	0.41	0.72	0.58	0.06
Mixture Model (per participant)	0.66	0.83	0.76	0.03
Mixture Model (per game)	0.66	0.83	0.76	0.03
Mixture Model (per trial)	0.7	0.84	0.78	0.03
Extended Mixture Model (per participant)	0.66	0.83	0.76	0.03
Extended Mixture Model (per game)	0.65	0.83	0.76	0.03
Extended Mixture Model (per trial)	**0.7**	**0.84**	**0.78**	**0.02**

Table 5.6: Summary of test statistics from posterior predictive agreement analysis on optimal data

rogate to the computationally expensive optimal recursive decision process. This suggestion was perhaps motivated by the high posterior predictive agreement reported between the τ-switch and optimal data: with mean values around 0.9 across a range of experimental setups in which the reward rates and numbers of trials are varied. The average posterior predictive agreement reported with the τ-switch model here is only 0.58. The reason for this difference could be due to the difference in the two models considered in each analysis: in this analysis, the N-armed variant

of the τ-switch model was used, whereas the two-armed variant was used by Lee et al. (2009). Additionally, as described previously in this section, the experimental set-up differed greatly between analysis, and it is unclear what form the optimal data generated for the analysis by Lee et al. (2009) was. It would be worthwhile to conduct a replication study of Lee et al. (2009) in order to test these theories.

The models which gave the highest posterior predictive agreement with optimal data were the original mixture model by Zhang and Lee (2010a), and the newly proposed mixture model which allow for latent switching between models on a trial-by-trail basis. It is interesting to note that these models give the best account of both human and optimal data, and suggest that flexible latent state models should be considered in future analyses.

Comparison of human and optimal parameter estimates

"Three basic challenges in studying any real-world decision-making problem are to characterize how people solve the problem, characterize the optimal approach to solving the problem, and then characterize the relationship between the human and optimal approach." - (Lee et al., 2011)

Finally, Table 5.7 shows the posteriors estimates of psychological parameters gathered after applying the model with the highest posterior predictive agreement with both human and optimal data - the newly proposed Extended Mixture Model - in order to draw parallels between parameter settings which give the best performance on each data set.

For many models above, the posterior estimates for both human and optimal data are almost identical, particularly the WSLS, e-WSLS, and latent state models; it is the ϵ variants which seem to differ most. The table above highlights the bene-

Parameter	Human Data	Optimal Data
ϵ-Greedy	0.158 (0.180)	0.197 (0.153)
ϵ-Decreasing	0.328 (0.112)	0.473 (0.144)
WSLS	0.589 (0.127)	0.520 (0.279)
Extended-WSLS Win	0.823 (0.229)	0.818 (0.228)
Extended-WSLS Lose	0.518 (0.284)	0.520 (0.279)
π-first Pi	8.00 (4.32)	8.00 (4.32)
π-first Gamma	0.265 (0.125)	0.413 (0.241)
Latent State	0.663 (0.142)	0.646 (0.138)
τ-switch Gamma	0.884 (0.045)	0.883 (0.0490)
τ-switch Tau	7.94 (4.32)	7.93 (4.32)

Table 5.7: Means and standard deviations of posterior parameter estimates of the Extended Mixture Model

fit of considering psychologically interpretable decision policies in analysing human decision-making: not only are we able to identify models which give good accounts of both human and optimal data, but for other models we are able to identify where human performance is falling short of optimal For example, when using the π-first model, participants are following the model to a lower degree of accuracy than the optimal data would suggest: perhaps participants are exploring more than the optimal data would suggest they should. An interesting avenue of future research could be to be gather results such as these, use them to provide feedback to participants, and examine whether the intervention improved decision making in a subsequent set of trials.

Chapter 6

Conclusion

6.1 Discussion of replications

One goal of this project was to replicate the findings of Steyvers et al. (2009) in order test their framework for model analysis and confirm the presence of individual differences in their data set. As detailed in section 4.1, the original results were reproduced almost perfectly, with only slight differences in results relating to the optimal model which can, I believe, the different ways in which the optimal was implemented between the two papers. These results affirmed the reliability of the framework and data used, allowing me to conduct similar analysis in section 5.2. Section 4.2 described another replication study, which unified the separate findings of two papers on the application of hierarchical Bayesian models to human data gathered through a series of bandit tasks. These results were less conclusive than the study in section 4.1, and possible explanations for why this occurred were detailed in section 4.2. One possibility is that conclusions drawn from the replication of models from one paper - those using a Beta prior in (Zhang & Lee, 2010b) - were

compared to results from models which used different priors - those reported in (Zhang & Lee, 2010a). Whilst this was acknowledged in Chapter 4, it is worthwhile repeating that these results should be analysed with caution. It was also made clear that the results from Zhang and Lee (2010a) were used only as a rough guideline for comparison, and not exact figures compared, with the replicated results, however, I would like reaffirm that here. Despite these caveats, I still believe that the similarity in magnitude of participant distributions between the replicated results and the original results by Zhang and Lee (2010a), and the almost perfect replication of parameter estimates of the models by Zhang and Lee (2010b), which used the same prior on parameters as the replicated models, provides good evidence for the reliability of the original findings. The findings of this project could have been made more reliable by replicating exactly the models in (Zhang & Lee, 2010a); this was not possible because, according to one of the authors of the original paper, the code was no longer available. In addition, this paper drew different numbers of posterior samples than Zhang and Lee (2010b) did, with numbers of samples and number of chains mirroring the results used by Zhang and Lee (2010a). I acknowledge that by mixing the models, methods, and experimental set-ups, and by attempting to replicate and unify the findings of two papers at once, I may not have done each paper justice. I conducted this analysis in this manner due to the missing - and seemingly inappropriate - priors used in (?, ?), as well as the incomplete model definitions in (Zhang & Lee, 2010b). In that sense, Section 4.2 is better viewed, not as a replication study, but as an extension of analysis upon previous similar works - the analysis has been described as a replication to ensure that the authors of the original papers receive their due credit. Future work which replicates each

of these papers more thoroughly than I have done may provide clearer insights into the original results presented.

6.2 Contributions of work

This paper has extended the τ-switch model originally reported by (Lee et al., 2011), with a likelihood function and graphical model for the hierarchical Bayesian variant of the model provided in Chapter 3. Lee et al. (2011) report that the model gives a good account of human data, and is a useful addition to the set of current models of finite-horizon two-armed bandit tasks. By extending this model to account for decisions made by N-armed bandit tasks, This model can be applied in the future to a larger set of bandit problems. The first application of the model to N-armed bandit tasks was presented in this project in section 5.2, and showed that the model was able to give a good account for human data, and could capture individual differences in human performance. Additionally, a hierarchical variant of the model was applied to human and optimal data in section 5.3. The results from this analysis, particularly the posterior predictive agreement presented in Tables [5.5] and [5.6] demonstrate that the model agreed with both data sets, giving the second best account in terms of the maximum value of each models posterior predictive agreement with both human and optimal data - values of 0.87 and 0.72, respectively - of any of the heuristic models, second only to the e-WSLS model. Also of significance is the likelihood function for models I have provided, which have previously only been described in written language, such as π-first, thus allowing future researchers to understand the assumptions I have made and reproduce my findings. Another key contribution of this project is the development of the Extended Mixture Model

detailed in Figure 3.8. Previous work (Zhang & Lee, 2010a) has constructed mixture models with four cognitive models considered. This model consists of a mixture of seven cognitive models, making it the most flexible mixture model applied to bandit tasks to date. Analysis of this model's ability to account for human decision-making by calculating the posterior predictive agreement between the model and human data showed that, on average, it gave a better account of data than any of the heuristic models considered, as well as outperforming the mixture model developed by Zhang and Lee (2010a). This was, of course, only one analysis, and the difference in performance of the model between its competitors was modest. In addition to this, due to the relatively expensive nature of the model, analysis was carried out on only 1000 samples from 2 chains - after a burn-in of 1000 samples. Applying this model to other data sets, and increasing the number of samples generated from this model after fitting on a data set, would be of great value in testing the findings of this analysis. Finally, I have developed two extensions of the two mixture models described in Chapter 3, both of which I believe make intuitive psychological sense. An analysis applying these models to human and optimal data found that the more flexible models which allowed for participants to change their decision making strategy on a per-game and per-trial basis gave a better account of both human and optimal data than the less flexible mixture models. The more flexible trial-by-trial model with seven mixture components gave the best account of both data sets seen across all models. It would be worthwhile to apply these models to a broader range of data from different types of bandit tasks, and to increase the number of samples drawn in each analysis, to produce more reliable results. I believe, however, that the results of this paper hint at how developing future hierarchical Bayesian models

with a larger number of component models and an increased ability to account for the flexibility of human decision-making might allow us to better understand how those decisions are made.

Chapter 7

Future Work

Findings from this analysis have shown that increasing the number of components in a hierarchical Bayesian model gave a better account of human and optimal data than simpler models. One such avenue for future research would, therefore, be to develop further mixture models with more components. One model of particular interest is the optimal Bayesian model detailed by Steyvers et al. (2009). This model was excluded from the mixture model in this project due to its computationally expensive nature and the time limitations in the first year of the project. Results from Steyvers et al. (2009) show that the optimal model is gives a good account of human behaviour, and a large proportion of the subjects in the data set used were inferred to be using this model. The first goal of the second year of this project will therefore be to include this model to the extended hierarchical Bayesian model detailed in this report, and to test this model on the data gathered by Steyvers et al. (2009).

In addition to the optimal model, (Zhang & Yu, 2013) outline another heuristic model, dubbed the Knowledge Gradient model, and report that it gives a better

account of human data than various models including the Win-Stay Lose-Shift, ϵ-greedy, and optimal model. When extending the hierarchical Bayesian model to include the optimal model, it will also be worthwhile to include this model in that analysis.

Results from this analysis demonstrated the benefit of applying previously developed models to new data in order to draw conclusions about model and parameter differences. It would be of great value, therefore, to apply the newly proposed models detailed in this report to a new set of data, in order to verify and test the findings discussed here. Whilst many papers (Zhang & Lee, 2010a; Lee et al., 2011) discussed in this report gathered data from bandit tasks which may be available for analysis, it may be of use to conduct an experiment of my own in order to gather new data. One advantage of gathering new data, rather than using a previously curated data set, is that I would be able to vary the set up of the experiments in order to test human behaviour in various settings - such as the "plentiful" and "scarce" environments described by (Lee et al., 2011). Steyvers et al. (2009), whilst gathering bandit decision-making data from 451 participants, also gathered psychometric data on cognitive abilities and personality traits, and found correlations between these traits and adherence to certain models. It would be interesting to gather similar data during an experiment, in order to test whether there are any correlations between cognitive abilities, or personality traits, and the newly proposed models. Finally, one reason for the use of cognitive models in analysing human performance on cognitive tasks is that the psychological parameters of those models can be analysed in order to understand how humans performed on those tasks. One purported benefit of such an analysis is that this information can be used to explain to subjects

where their performance fell short of optimal, and instruct them on how they should perform on future tasks. It would be interesting, therefore, to test whether the conclusions drawn from the parameters of cognitive models can be used to improve human decision-making on bandit tasks by gathering a group of subjects and asking them to complete a series of bandit problems. Once data has been gathered and posterior estimates of model parameters drawn, instruction could be given to the original subjects on how they could improve performance, and a new set of bandit problems presented to them in order to determine whether or not the instruction improved performance.

References

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, *28*, 403–450.

Berry, D. A., & Fristedt, B. (1985). Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, *5*, 71–87.

Blanco, N. J., Otto, A. R., Maddox, W. T., Beevers, C. G., & Love, B. C. (2013). The influence of depression symptoms on exploratory decision-making. *Cognition*, *129*(3), 563–568.

Burtini, G., Loeppky, J., & Lawrence, R. (2015). A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942.

Eliassen, S., Jørgensen, C., Mangel, M., & Giske, J. (2007). Exploration or exploitation: life expectancy changes the value of learning in foraging strategies.

Oikos, *116*(3), 513–523.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, *20*(5), 578–585.

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

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 148–164.

Gottschling, J., Spengler, M., Spinath, B., & Spinath, F. M. (2012). The prediction of school achievement from a behavior genetic perspective: Results from the german twin study on cognitive ability, self-reported motivation, and school achievement (cosmos). *Personality and Individual Differences*, *53*(4), 381–386.

Heathcote, A., Brown, S., & Mewhort, D. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*, *7*(2), 185–207.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive science*, *30*(1), 3–41.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., Group, C. S. R., et al. (2015). Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, *19*(1), 46–54.

Hu, T., Zhang, D., & Wang, J. (2015). A meta-analysis of the trait resilience and

mental health. *Personality and Individual Differences*, *76*, 18–27.

Hung, Y.-C. (2012). Optimal bayesian strategies for the infinite-armed bernoulli bandit. *Journal of Statistical Planning and Inference*, *142*(1), 86–94.

Hunt, E., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In *Psychology of learning and motivation* (Vol. 7, pp. 87–122). Elsevier.

Jern, A., Lucas, C. G., & Kemp, C. (2011). Evaluating the inverse decision-making approach to preference learning. In *Advances in neural information processing systems* (pp. 2276–2284).

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237–285.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental science*, *10*(3), 307–321.

Lee, M. D. (2008). Three case studies in the bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

Lee, M. D. (2011a). How cognitive modeling can benefit from hierarchical bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7.

Lee, M. D. (2011b). Special issue on hierarchical bayesian models. *Journal of Mathematical Psychology*, *55*, 1–118.

Lee, M. D., & Newell, B. R. (2011). Using hierarchical bayesian methods to examine the tools of decision-making.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical*

course. Cambridge university press.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2009). Using heuristic models to understand human and optimal decision-making on bandit problems. In *Proceedings of the ninth international conference on cognitive modeling—iccm2009. manchester, uk.*

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164–174.

L Griffiths, T., Kemp, C., & B Tenenbaum, J. (2008). Bayesian models of cognition.

Lindley, D. V., & Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(1), 1–18.

Mata, R., Wilke, A., & Czienskowski, U. (2013). Foraging across the life span: is there a reduction in exploration with aging? *Frontiers in neuroscience*, *7*, 53.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191.

Merkle, E. C., Smithson, M., & Verkuilen, J. (2011). Hierarchical models of simple mechanisms underlying confidence in decision making. *Journal of Mathematical Psychology*, *55*(1), 57–67.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using dirichlet processes. *Journal of mathematical Psychology*, *50*(2), 101–122.

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

Parker, A. M., De Bruin, W. B., & Fischhoff, B. (2007). Maximizers versus satisficers: Decision-making styles, competence, and outcomes. *Judgment and Decision making*, *2*(6), 342.

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

Pooley, J. P., Lee, M. D., & Shankle, W. R. (2011). Understanding memory impairment with memory models and hierarchical bayesian analysis. *Journal of Mathematical Psychology*, *55*(1), 47–56.

Pothos, E. M., Perry, G., Corr, P. J., Matthew, M. R., & Busemeyer, J. R. (2011). Understanding cooperation in the prisoner's dilemma game. *Personality and Individual Differences*, *51*(3), 210–215.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of personality and social psychology*, *83*(5), 1178.

Shen, W., Wang, J., Jiang, Y.-G., & Zha, H. (2015). Portfolio choices with orthogonal bandit learning. In *Twenty-fourth international joint conference on artificial intelligence.*

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science*, *32*(8), 1248–1284.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *Winbugs user manual.*

version.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction.

Todd, P. M., Hills, T. T., & Robbins, T. W. (2012). *Cognitive search: Evolution, algorithms, and the brain* (Vol. 9). MIT press.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the bart: Assessment and application. *Journal of Mathematical Psychology*, *55*(1), 94–105.

Zeigenfuse, M. D., & Lee, M. D. (2009). Bayesian nonparametric modeling of individual differences: A case study using decision-making on bandit problems. In *Proceedings of the 31st annual conference of the cognitive science society, austin, tx: Cognitive science society* (pp. 1412–1417).

Zhang, S., & Angela, J. Y. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems* (pp. 2607–2615).

Zhang, S., & Lee, M. D. (2010a). Cognitive models and the wisdom of crowds: A case study using the bandit problem. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 32).

Zhang, S., & Lee, M. D. (2010b). Optimal experimental design for a class of bandit problems. *Journal of Mathematical Psychology*, *54*(6), 499–508.

Zhang, S., Lee, M. D., & Munro, M. (2009). Human and optimal exploration and exploitation in bandit problems. *Ratio*, *13*(14), 15.

Zhang, S., & Yu, A. (2013). Cheap but clever: human active learning in a bandit setting. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).