

Strategy Selection and Function Learning in Decision Making

Hrvoje Stojić

TESI DOCTORAL UPF / ANY 2016

DIRECTOR DE LA TESI

Robin M. Hogarth, Departament d'Economia i Empresa



to Nikolina

Acknowledgments

A PhD is much more than a thesis; it is about being trained as a scientist. I was privileged to receive training from Robin Hogarth, an outstanding expert in judgment and decision making research and a wonderful person. Robin has been generous with his time, quick to respond to any request, and always had words of encouragement, accompanied with a smile. He has guided me in appraising ideas, taught me how to recognize the good ones and discard the bad ones. He believed in me and gave me a chance to become a scientist.

I am greatly indebted to several other people. To Gaël Le Mens for always finding time for me and inspiring me to explore reinforcement learning concepts. To Henrik Olsson, who worked with me on the challenging strategy selection problem, introduced me to the world of exemplar-based models and hosted me at Max Planck Institute for Human Development. To Pantelis P. Analytis, with whom I have delved into social decision making and dreamed up exciting new projects. To Maarten Speekenbrink and Peter Dayan for helping me to realize the potential of the feature-based multi-armed bandit project and hosting me at University College London on several occasions. My special thanks go to Omiros Papaspiliopoulos and Gabor Lugosi who gave me the opportunity to teach in the Master of Data Science program. These courses provided me with challenges that deepened my knowledge of machine learning, but also gave me opportunity to meet many talented students.

There are many other colleagues and professors who have helped me in my PhD mission: Thomas Störk, Andrea Matranga, Giovanni Giusti, Tom Schmitz, Andrei Potlogea, José Alejandro Martinez, Niklas Heusch, Francesco Cerignoni, Jagdish Tripathy, Philipp Schustek, Eric Schulz, Rosemarie Nagel, José Apesteguia, Larbi Alaoui, Mihalis G. Markakis, Christian Fons-Rosen, Kurt Schmidheiny, Mirta Galešić, Dražen Prelec, Konstantinos Katsikopoulos and Samuel Gershman. I am thankful to them for taking time to listen to my ideas and for their advices. I am particularly grateful to Marta Araque, Laura Agustí and Esther Xifré who helped me immensely in navigating administrative complexities of PhD life, and to Pablo López-Aguilar Beltrán for support in conducting experiments – without their support my mission would have been almost impossible. A special mention goes to Michael Anreiter, who is sorely missed.

Finally, I would like to thank my friends and family who often had to put up with my absence, and Nikolina without whom I would be truly lost – I love you.

Abstract

This thesis consists of three studies investigating the strategy selection problem and the role of function learning in human decision making. Chapter 1 examines how people learn which decision strategy to use when facing multiple environments. It provides evidence that people associate different decision strategies to different types of environments through a trial-and-error type of process and learn to flexibly switch between the strategies as needed. Chapter 2 aims to identify the source of inter-individual differences in strategy adoption. It suggests that such differences can be traced back to how fast people learn the relationships between cues and the criterion they are trying to infer. Finally, Chapter 3 focuses on how people simultaneously learn functional relationships between cues and alternative rewards and make decisions. It provides evidence for interactions between function learning and decision processes and proposes a Bayesian optimization framework for understanding these interactions.

Resum

Aquesta tesi consisteix en tres estudis que investiguen el problema de selecció de l'estratègia i el paper de la funció de l'aprenentatge en la presa de decisions humana. El capítol 1 examina com les persones aprenen quina estratègia de decisió utilitzar quan s'enfronten a múltiples entorns. Es proporciona evidència de que les persones associen diferents estratègies de decisió a diferents tipus d'ambients a través d'un tipus de procés d'assaig i error i aprenen a canviar de forma flexible entre les estratègies segons sigui necessari. El capítol 2 té com a objectiu identificar l'origen de les diferències interindividuals en l'adopció d'estratègia. Es suggereix que aquestes diferències es poden remuntar-se a quant ràpidament les persones aprenen les relacions entre els senyals i el criteri que estan tractant d'inferir. Finalment, el capítol 3 es centra en com les persones aprenen al mateix temps les relacions funcionals entre senyals i recompenses alternatives i prenen decisions. Es proporciona evidència de les interaccions entre l'aprenentatge funcional i els processos de decisió i proposa un marc d'optimització bayesiana per a la comprensió d'aquestes interaccions.

Preface

“Models are like toothbrushes – everyone should have one, but you would never dream of using someone else’s”

M. J. Watkins (1984)

This popular proverb summarizes the state of the affairs in research on human judgment and decision making. For any given decision problem one has a multitude of models to choose from. As a fresh scholar in the field I was struck by this fact. The decision models that researchers hold so dear, often but not always, aim to capture decision strategies people use when facing decision problems. My initial thought was that perhaps, there is no single true model, instead all of them are correct – people might be using more than one decision strategy. However, then the question becomes: how do we choose which strategy to use?

This was the first question I decided to tackle in my studies. I discovered that many other researchers pursued this question. It has been termed the strategy selection or “deciding how to decide” problem. While Beach and Mitchell (1978) and Christensen-Szalanski (1978) were one of the first to tackle it, the “Adaptive decision making” book by Payne, Bettman, and Johnson (1993) provided a more extensive treatment. It characterized a set of strategies and made explicit the connection to properties of external environments that decision situations emanate from. This early work advanced a cost-benefit approach for dealing with the strategy selection problem. According to it, people choose a strategy by trading the benefits of applying a strategy against its costs. The benefits are related to the strategy’s accuracy, while the costs are related to the time or cognitive effort of applying the strategy. More recently, Gigerenzer and colleagues from the Adaptive Behavior and Cognition group in Berlin developed a series of what they call fast-and-frugal heuristics (Todd & Gigerenzer, 2000). These heuristics are not general strategies that are meant to be applied in every situation, but instead they are highly adapted to solving a particular problem. The authors explicitly recognized that for the fast-and-frugal heuristics approach to work, the strategy selection problem has to be tackled – how does one know when to use which heuristic? The strategy selection problem does not appear only in these specific approaches to decision making, the problem is more widespread than one might think initially – it is implicitly present in many other areas of cognitive psychology. For example,

in category and function learning there is a consensus that people use both exemplar-based and rule-based processing to learn concepts and functions (Ashby & Maddox, 2005; Busemeyer, Byun, Delosh, & McDaniel, 1997).

In the first chapter, *Not everything looks like a nail: Learning to select appropriate decision strategies in multiple environments*, co-authored with Henrik Olsson and Maarten Speekenbrink, I take the learning approach to solving the strategy selection problem. The focus of this approach is the learning process by which those strategies that result in highest average rewards end up being used relatively more than other less rewarding strategies. This reinforcement learning approach was recently developed as an alternative to the cost-benefit one (Erev & Barron, 2005; Rieskamp & Otto, 2006). Previous research showed that when facing decision problems coming from a single environment, people can learn to select appropriate strategies. However, what happens when, as is typical outside the psychological laboratory, they face multiple environments? In the chapter I argue that the strategy selection problem is compounded with the category learning problem – people need to learn to select the strategy and learn in which environment to use it. In the experiments people face interleaved decision tasks, one from a linear and the other from a nonlinear environment, with qualitatively different strategies being more appropriate in each. The chapter provides evidence that people are able navigate such scenarios, adopting appropriate strategies in each environment. The reinforcement learning approach can account for this process – people associated different strategies to different types of environments through a trial-and-error type of process, and learned to flexibly switch between the strategies as needed.

Even though people generally learn to use the appropriate strategy, there are always those that keep using less effective strategies. What is the source of this inter-individual variability? This is the question I turn to in the second chapter, *Explaining inter-individual variability in strategy selection: A cue weight learning approach*, co-authored with Henrik Olsson and Pantelis P. Analytis. Past research has tried to explain the adopting of strategies like take-the-best (Gigerenzer & Goldstein, 1996) or weighted additive (Payne et al., 1993) using intelligence scores, working memory span or personality traits. Yet, they did not yield satisfying results. In this chapter, I propose that the puzzle could be explained by differences in the speed of learning. Adoption of the strategy should depend on the statistical properties of the environment – how the cues of the alternatives are related to the criterion they are judging. Since such properties have to be learned, my thesis is that commonly observed differences in learning might result in differences in strategy adoption. In mustering the evidence the exper-

iments involved both deciding between multi-cued alternatives as well as making predictions about alternatives' criterion values. This allowed us to jointly study decision making and learning the properties of the environment. Overall, the results presented in the chapter provide support for the thesis.

In chapter two, we observed how learning the properties of the environment can be coupled with decision making. Exploring interactions between these two processes can substantially improve our understanding of how peoples' knowledge of functional relationships and concepts is shaped, and how decision making in the wild operates. Unfortunately, these two processes have been mostly studied in isolation so far. On the one hand, researchers studying multiple-cue probability learning and function learning have investigated how people learn relationships between observed cues and an unobserved criterion value (e.g. Busemeyer et al., 1997; Hammond & Stewart, 2001). On the other, researchers working on reinforcement learning have been very successful in explaining how animals and humans learn to choose rewarding stimuli and avoid punishments (e.g. Niv, 2009; Schultz, Dayan, & Montague, 1997). In reality, however, the samples used to learn the functional relations are systematically skewed by making decisions that lead to accumulation of rewards. For instance, we might have a good knowledge of what makes a good restaurant but much poorer knowledge of what makes a bad one. Decision processes are also biased by our knowledge of functional relations – when weighing between alternatives, people can draw on such knowledge to predict the value of alternatives, even ones they have never seen before. For example, our previous experiences with dining in restaurants tells us that popularity strongly predicts the quality of the meal and we are likely to have a tendency to choose a restaurant with more patrons for our next outing.

The last chapter of the thesis, *Trials-with-fewer-errors: Feature-based learning and exploration*, co-authored with Pantelis P. Analytis, Peter Dayan and Maarten Speekenbrink, is one of the first studies that systematically investigates the interactions between learning of the functional relations and decision making. I consider it to be the most developed chapter in the thesis. It introduces a novel feature-based multi-armed bandit task to study the interactions in more detail. In this task rewards are a noisy function of the features, an important difference in comparison to reinforcement learning tasks often used to study how animals and humans learn to choose rewarding actions and evade punishments (Gershman & Daw, 2017; Gershman & Niv, 2010). The chapter also puts forward a Bayesian optimization framework for tackling such decision making problems (Shahriari, Swersky, Wang, Adams, & de Freitas, 2016). The framework relies on similarity-

based learning of functional relationships between features and rewards, and choice rules that use uncertainty to balance exploration and exploitation. Through a series of experiments it is demonstrated that people's exploration patterns exhibits clear signs of Bayesian optimization – simultaneous function learning and function maximization. Several other predictions obtained from the framework are supported as well. Overall, this chapter makes an important contribution by jointly studying two processes – function learning and decision making, that have been studied in isolation thus far. It charts the new territory by illustrating the interactions between them and advancing a theoretical framework for understanding the interactions.

Contents

List of Figures	xvi
List of Tables	xvii
1 Not everything looks like a nail: Learning to select appropriate decision strategies in multiple environments	3
1.1 Introduction	4
1.1.1 Strategy selection in multiple environments	6
1.1.2 Overview	9
1.2 Method	10
1.2.1 Participants	10
1.2.2 Materials	10
1.2.3 Procedure	13
1.3 Results	13
1.3.1 Behavioral analysis	13
1.3.2 Identifying the strategies adopted by the participants	16
1.3.3 Contextual strategy selection learning	23
1.4 Discussion	28
1.A Parameter Estimation and Model Selection	33
1.A.1 Identifying the strategies adopted by the participants	33
1.A.2 Contextual strategy selection learning	34
1.B Additional results	37
2 Explaining inter-individual variability in strategy selection: A cue weight learning approach	41
2.1 Introduction	42
2.2 Method	44
2.2.1 Participants	44
2.2.2 Stimuli and procedure	44
2.3 Behavioral results	46
2.3.1 Choices in the decision task	46

2.3.2	Predictions in the estimation task	47
2.3.3	Relation between decisions and predictions	49
2.4	Modeling	49
2.4.1	Modeling the cue weight learning	49
2.4.2	Modeling the choices	52
2.5	Modeling results	53
2.6	Discussion & Conclusion	56
3	Trials-with-fewer-errors: Feature-based learning and exploration	61
3.1	Introduction	62
3.2	Goals and Scope	64
3.3	Feature-based Multi-Armed Bandit Task	65
3.4	Function Learning Approach	69
3.4.1	Gaussian Process Regression	70
3.4.2	Upper Confidence Bound strategy	74
3.4.3	Mean Tracking Approach	77
3.5	Summary	78
3.6	Experiment 1A and 1B: Influence of feature information on learning and exploration	79
3.6.1	Method	80
3.6.2	Results and Discussion	84
3.7	Experiment 2A and AB: Are People Function Learners and the Hidden Dangers of Function Learning	92
3.7.1	Method	93
3.7.2	Results and Discussion	94
3.8	Experiment 3: Different Flavors of Exploration and Factors Affecting Exploration	102
3.8.1	Method	104
3.8.2	Results and Discussion	106
3.9	General Discussion	110
3.9.1	Implications	112
3.9.2	Limitations	115
3.9.3	Concluding Remarks	117
3.A	Ensuring data quality	119
3.B	Details on stimuli in the functional knowledge task	120
3.B.1	Items in the linear environments	120
3.B.2	Items in the nonlinear environments	122
3.C	Additional Results	124
3.D	Bayesian Models Performance and Parameter Overview	131
	Bibliography	133

List of Figures

1.1	Accuracy of participants' choices in blocks of trials in the training phase and in the test phase	15
1.2	Model performance in predicting individual choices in the test phase, presented separately for each environment . . .	22
1.3	Model performance in predicting choices in the test phase for strategy selection learning models.	27
1.4	Probability of choosing the GCM strategy by CSSL models over time in each environment in the training phase.	29
1.B.1	Choice accuracy in blocks in the training and the test phase.	37
1.B.2	Model performance in predicting choices in the test phase for strategy selection learning models, separately for the linear and nonlinear environment.	38
2.1	Screenshots of the tasks from the experiment.	45
2.2	There is a clear learning effect in the decision making task.	47
2.3	Participants have insight into their own learning of cue validity weights – they learned the ranking of cues and cue directions to a large extent.	48
2.4	Participants make good predictions in the estimation task. .	50
2.5	Relation between performance in the estimation task and the decision task.	51
2.6	Participants best fitted by the weighted additive model are learning cue validity weights much faster than participants best fitted by the take-the-best model or the random choice model.	55
2.7	Estimated learning and decay rate parameters for the Least mean squares network model with decaying learning rates.	56
3.1	Screenshot of the FMAB task from the experiment.	68
3.2	Illustrating the GP-UCB model with a single feature function.	75
3.3	Screenshot of the FK task.	82

3.4	Choice performance of Bayesian models and participants in Experiment 1A.	86
3.5	Exploration patterns of Bayesian models and participants of the bandit tasks in Experiment 1A.	88
3.6	The GP-UCB model is able to generalize and performs very well on the FK task in Experiment 1A.	90
3.7	Exploration patterns of Bayesian models in Experiment 2A.	95
3.8	Exploration patterns of participants in Experiment 2A in the first 10 trials of the bandit tasks and the last 10 trials. . . .	97
3.9	Allocation of choices in the feature space by participants in the FMAB condition in Experiment 2B.	100
3.10	Screenshot of the FL task from Experiment 3.	106
3.11	Differences in the exploration patterns in Experiment 3 between the fFMAB conditions and FMAB conditions.	108
3.12	Differences in exploration patterns in Experiment 3 between participants that completed either a short horizon FMAB task (30 trials) or a long horizon FMAB task (100 trials). . .	111
3.C.1	Exploration patterns of participants in all 100 trials of the bandit tasks in Experiment 1A.	124
3.C.2	Performance of GP-UCB model and participants in Experiment 1A on FK task, broken across item types.	125
3.C.3	Behavioral results of participants in the bandit task in Experiment 1B.	126
3.C.4	Choice performance of participants in Experiment 2.	127
3.C.5	Exploration patterns of two clusters of participants in the FMAB-ml condition in Experiment 2 in the first 10 trials of the bandit tasks and the last 10 trials.	128
3.C.6	Performance of Bayesian models and participants in FMAB-q condition in Experiment 2 on FK task.	129
3.C.7	Choice performance of participants in Experiment 3.	130

List of Tables

1.1	Overview of the stimuli characteristics in the Experiment. .	12
1.2	Overview of the models representing each type of strategy.	17
1.A.1	Summary of parameter estimates for models in identifying the strategies adopted by the participants in each environment.	35
1.A.2	Summary of parameter estimates for strategy selection models.	36
2.1	Overview of model fits.	53
3.1	Overview of experiments and characteristics of participants across experimental conditions.	81
3.D.1	Overview of performance and estimated parameters of the models on the same stimuli that participants have had in the experiments.	132

Stojić, H., Olsson, H., & Speekenbrink, M. (2016). Not everything looks like a nail: Learning to select appropriate decision strategies in multiple environments. PsyArXiv. DOI: [10.17605/OSF.IO/FMA3P](https://doi.org/10.17605/OSF.IO/FMA3P)

Chapter 1

Not everything looks like a nail: Learning to select appropriate decision strategies in multiple environments

Abstract

How do people choose which decision strategy to use? When facing single tasks, research shows that people can learn to select appropriate strategies. However, what happens when, as is typical outside the psychological laboratory, they face multiple tasks? Participants were presented with two interleaved decision tasks, one from a nonlinear environment, the other from a linear environment. The environments were initially unknown and participants had to learn their properties. Through cognitive modeling, we examined the types of strategies adopted in both tasks. Based on out of sample predictions, most participants adopted a cue-based strategy in the linear environment and an exemplar-based strategy in the nonlinear environment. A context-sensitive reinforcement learning model accounts for this process. Thus, people associated different strategies to different types of environments through a trial-and-error type of process, and learned to flexibly switch between the strategies as needed. This evidence further supports the strategy selection approach to decision making which assumes that people pick and apply strategies available to them according to task demands.

1.1 Introduction

In the same way as a carpenter is able to choose between a hammer and a screwdriver to deal with a nail, the adaptive toolbox approach to judgment and decision making assumes that, when faced with a decision problem, a decision maker is able to choose an appropriate strategy from her toolbox of decision strategies (Gigerenzer, Todd, & the ABC Research Group, 1999; Payne et al., 1993; Scheibehenne, Rieskamp, & Wagenmakers, 2013). Entertaining the possibility that the mind carries such a toolbox, the question is then: how do we know which strategy to use in which situation? This question has been termed the strategy selection or “deciding how to decide” problem.

In the last two decades theoretical and empirical advances have been made in tackling the strategy selection problem. First theoretical attempts were cost-benefit approaches (Beach & Mitchell, 1978; Christensen-Szalanski, 1978; Lieder & Griffiths, 2015; Payne et al., 1993; Russell & Wefald, 1991). According to this approach, people choose a strategy by trading the benefits of applying a strategy against its costs. The benefits are related to the strategy’s accuracy, while the costs are related to the time or cognitive effort of applying the strategy. More recently, reinforcement learning approaches appeared as an alternative to cost-benefit analysis (Erev & Barron, 2005; Rieskamp & Otto, 2006). The focus of this approach is the learning process by which those strategies that result in highest average rewards end up being used relatively more than other, less rewarded strategies.

Despite these advances, there is at least one major problem not addressed theoretically or empirically. People navigate through *multiple environments* – classes of situations in which a certain strategy performs better than others. Not everything is a nail and situations differ – for example, when deciding between wines you might be better off using the take-the-best heuristic (Gigerenzer & Goldstein, 1996), while for choosing a cheese you might want to use a similarity based strategy (Nosofsky & Bergert, 2007). The strategy selection approach implies that people should treat different environments as such and adapt to each as needed. Moreover, they must be able to recognize a certain decision situation as belonging to an environment and flexibly shift between different strategies as they encounter one environment or the other. The empirical evidence thus far, however, mostly shows that people are able to select an appropriate strategy in a *single* environment. For example, experiments in Pachur and Olsson (2012), Rieskamp and Otto (2006) and Karlsson, Juslin, and Olsson (2007) employed between-subject designs where each participant faced

only one environment.¹ Hence, the question if participants can adaptively select strategies in tasks with multiple environments and decision situations is still unanswered.

Improvements can also be made in terms of evaluating formal models of strategy selection. Thus far empirical evaluations were based on environments where values of alternatives were linear functions of cues or attributes and information about the function in terms of cue validities was provided to the participants (Lieder & Griffiths, 2015; Rieskamp, 2006; Rieskamp & Otto, 2006, but see J. Hoffmann, von Helversen, & Rieskamp, 2014, for a recent exception). Exemplar-based strategies (Nosofsky, 1984; Nosofsky & Bergert, 2007) have not yet been included in such models.² Given the body of evidence for exemplar-based processing and that such strategies can also perform well in nonlinear types of environments, support for any strategy selection model is incomplete when only evaluating it in linear environments. Moreover, explicitly providing information about the statistical properties of the environment greatly facilitates solving the strategy selection problem. In more realistic situations these properties have to be discovered as well, and this important aspect of the strategy selection problem has thus far been ignored.

Our objective is to put the strategy selection approach to judgment and decision making to a stronger test by evaluating it in a multi-environment setting where participants face alternating instances of two different environments on a trial-to-trial basis. Moreover, one environment will be of a linear, while other of a nonlinear nature – requiring of participants to adopt qualitatively different strategies to perform well in them. Finally, the characteristics of the environments will be initially unknown and participants need to learn their properties.

We make two main contributions. First, we provide evidence that people can learn to flexibly use appropriate decision strategies on a trial-to-trial basis in initially unknown linear and nonlinear environments. This provides strong additional support for the strategy selection approach to decision making. Second, our contextual version of the reinforcement learning based strategy selection model (SSL Rieskamp & Otto, 2006) accounts for how people learn to associate different decision strategies to different en-

¹There are studies that examined dynamic environments, where there is a sudden shift in statistical properties and appropriate strategy (Bröder & Schiffer, 2006; Rieskamp, 2006). However, this is a change in properties of the *same* environment and there was no difference in observable features that would indicate the difference between the environments.

²In fact, Rieskamp and Otto (2006) considered it to be an alternative to their SSL model, instead of possibly another strategy in the toolbox.

vironments. In what follows, we first discuss the problem of strategy selection in multiple environments and examine how it fits in the landscape of existing theories of strategy selection. We then describe the design of our experiment, introduce the task and our qualitative predictions, and report the results. Then we describe the formal implementation of the contextual SSL model and assess how well it accounts for our results. We close with a discussion of our results and a call for further theoretical development with regards to the interaction between the categorization of environments and strategy selection.

1.1.1 Strategy selection in multiple environments

In a reply to a précis on fast-and-frugal heuristics (Todd & Gigerenzer, 2000), an influential work outlining a decision making framework where strategy selection has a strong role, Luce (2000) applauded the authors for presenting a different approach to studying judgment and decision making, and raised an issue of “how does one classify problems and decide upon which of several fast and frugal heuristic to employ?” (p. 758). In the same issue, Morton (2000) also noticed that classifying decision problems is a necessary component of the approach. Morton imagined an agent having a set of strategies and a database of previously encountered problems. The database contains the type of problem, which strategy was applied, and its performance. When a new problem is encountered, this database can be used to classify the problem and then to select between the strategies. Decision-making researchers took little notice of these early observations – the issue of how people classify problems has not been addressed explicitly yet.

Classifying problems does not look like a serious issue at first glance: everybody can trivially see that choosing between cheeses is a different situation than choosing between wines. But here is the catch: while such perceptual features can signal that a decision problem is different from another one, they may not be relevant at all for determining which strategy should be used in it.

Normative research has shown that important indicators for strategy performance are statistical properties of the environment. For example, in environments where the value of an alternative is a weighted additive (linear) function of cue values, features such as dispersion of cue weights or cue inter-correlations are good predictors of strategy performance (Hogarth & Karelaia, 2005a, 2005b, 2006a, 2006b, 2007; Martignon & Hoffrage, 2002; Martignon & Laskey, 1999). In such linear environments optimal cue

weights can have a compensatory or non-compensatory type of dispersion. A non-compensatory pattern is such that the cue with the greatest weight cannot be beaten by any pattern of values for the remainder of the cues. In a non-compensatory environment, a lexicographic strategy such as take-the-best (TTB; Gigerenzer & Goldstein, 1996), which focuses on the most important cues and ignores the rest, will perform well. In an environment where the optimal cue weights have a compensatory pattern, a strategy that integrates all the cues, such as the weighted additive rule (WADD; Payne et al., 1993), will perform well. Higher inter-correlations between the cues imply higher redundancy, that is, less information is obtained from knowing the value of each additional cue. Hence, lexicographic strategies do not lose much by ignoring most of the cues and might outperform strategies that integrate all cues (Hogarth & Karelaia, 2005a, 2006a).³

Cue weight dispersion and cue inter-correlations are not immediately available perceptual features. A compensatory and non-compensatory environment might be perceptually very similar. And two environments that are perceptually very different might both be of a compensatory nature, and thus should belong to the same category with respect to decision strategies. When faced with an unknown environment, how do people infer the statistical properties of that environment in order to choose which decision strategy to apply? Taking a reinforcement learning approach to strategy selection, such inferences are not actually required. What matters is that people can learn that certain features indicate that compensatory strategies are likely to be successful, and other features are predictive of the success of non-compensatory strategies. Nonetheless, strategy selection in multiple environments involves non-trivial complexities of mapping the decision situations to the space of strategies.

What are the potential solutions to this joint problem of selecting the strategy and classifying decision situations? Lieder and Griffiths (2015) propose a solution in the vein of the cost-benefit tradition, where one weighs the cost of applying each strategy against its estimated accuracy, and selecting the one that yields the best ratio. They propose using the statistical properties discussed above as features to predict the expected re-

³Other characteristics of environments have also been studied. The link between strategy effectiveness and properties like the number of observations, number of cues, and dominance relations is currently unclear (Gigerenzer et al., 1999; Martignon & Hoffrage, 2002; Martignon & Laskey, 1999). Under time pressure people use more frugal heuristic strategies like TTB (e.g. Rieskamp & Hoffrage, 2008). Cognitive effort also plays a role. People with better episodic memory have a stronger tendency to use exemplar-based strategies (J. Hoffmann et al., 2014), presumably because employing this strategy is less costly for these people.

ward of applying each strategy through linear or logistic regression. Such an approach can work well when decision makers know the properties and relevant features of the environment well. This is the situation in which Lieder and Griffiths (2015) evaluated their model – participants encountered compensatory and non-compensatory environments with the validity of each cue displayed. However, their model cannot be applied as easily in situations where such environmental properties are initially unknown. Nonlinear environments pose an even greater obstacle. While the statistical properties of linear environments have been identified that predict whether TTB or WADD will fare better, features that predict whether exemplar-based strategies are more appropriate, such as those related to the nonlinearity of environments, are not yet known (Pachur & Olsson, 2012).

In this paper we take a reinforcement learning approach to solving the dual problem of classifying decision situations and selecting the appropriate strategy within a situation. In the reinforcement learning approach, a strategy which accumulates more rewards when applied in a particular environment will be used more often. The SSL model (Rieskamp & Otto, 2006) has previously been used to describe strategy selection in single linear compensatory or noncompensatory environments with known cue validities. To deal with multiple environments, we extend SSL by assuming that decision makers use observable features to separate decision situations into different categories (e.g., cheeses and wines). Ignoring the latent statistical properties, this contextual version of SSL will run two separate reinforcement learning processes, one for each category, treating them as potentially different environments. If cheese and wine categories are indeed such that different strategies should be used in them, the model will eventually learn which strategy results in higher average reward. However, if they were such that the same strategy should have been used in both – for example, if both turned out to be compensatory such that WADD performs well – then the effort was duplicated. This is a slow and potentially wasteful mechanism, but it has the advantage that the decision maker does not have to know complex statistical features such as cue inter-correlations. This is particularly useful when facing nonlinear environments, where we only need to assume that the decision maker’s repertoire also contains strategies that can handle nonlinear environments, such as exemplar-based strategies (Nosofsky, 1984; Nosofsky & Bergert, 2007; Pachur & Olsson, 2012). In addition, we assume that the decision maker’s repertoire contains strategies that are able to learn, or approximate, a variety of functions that relate the cues to the value of decision alternatives. Whilst learning which strategy to use, a decision maker simultaneously adapts individual strategies to the

particulars of the environment. Hence, our contextual SSL model can work both in novel situations and in environments that decision makers know well. With sufficient experience, the individual strategies in the repertoire have adapted to the environment and it is clear which strategy will provide the maximum rewards.

1.1.2 Overview

We examined whether people are able to learn to use appropriate decision strategies when faced with multiple environments, flexibly shifting between them on a trial-by-trial basis. Participants in our experiment performed a paired comparison task where the goal was to pick the alternative with the highest criterion value. Each alternative was described by four cues and each paired comparison belonged to one of two types of environments – a linear or a nonlinear environment. In the linear environment, the task can be solved equally well by either a cue-based strategy that combines cue values in a linear fashion or an exemplar-based strategy. In the nonlinear environment, an exemplar-based strategy has a clear advantage over cue-based strategies as it can approximate the nonlinear function. The main prediction of a strategy selection approach to decision making is that in the linear environment the participants will adopt a strategy mix where cue-based strategies are used most often. In the nonlinear environment, the strategy mix should be dominated by exemplar-based strategies.

As outlined in the previous section, the reinforcement learning approach to strategy selection tackles this problem by partitioning the decision situations on the basis of perceptual information. We used two cover stories that were easy to visually differentiate – “bugs” and “comics” – that represented either the linear or nonlinear environment. If participants cannot adopt appropriate strategy mixes in this relatively simple situation, there is little hope they will be able to do so when faced with less perceptually differentiated environments.

Our analysis relies on two modeling approaches. After confirming that participants indeed learn over time in our task we first identify which strategy they have adopted in each environment. We accomplish this by fitting several cue-based and exemplar-based models separately to trials from each environment. We examine the extent to which participants have appropriately adopted different classes of strategies in each environment and narrow down the most representative strategies in both cue-based and exemplar-based class. Second, using the selected representative strategies as a strategy repertoire, we fit the contextual strategy selection learning

model to both environments simultaneously, with the aim of explaining how the strategy preferences develop over time. With the first modeling exercise, besides deriving inputs for the strategy selection modeling, we obtain evidence of strategy use that does not rely on the precise learning mechanism we assume in the strategy selection modeling.

1.2 Method⁴

1.2.1 Participants

Fifty-five participants (29 women, 26 men, $M_{age} = 21.4$, age range: 18–40 years) took part in the experiment. Participants were recruited from the Universitat Pompeu Fabra subject pool. They were paid a show-up fee of three euros and an additional performance-dependent bonus (5.8 euros on average). The experiment was run in groups of about 10 people in the BES laboratory at Pompeu Fabra University. The experiment lasted for one hour on average.

Six participants did not reach the required level of accuracy in the training phase and did not continue to the test phase. Two of these participants failed to reach the required level of accuracy in the nonlinear environment, while the other four did not perform well enough in the linear environment. These participants were excluded from the analysis completely. The final sample consisted of 49 participants (27 women, 22 men, $M_{age} = 21.6$, age range: 18–40 years).

1.2.2 Materials

On each trial in the learning and test phase, participants were presented with a pair of stimuli and had to choose the stimulus with the higher criterion value. The stimuli used were modified from Pachur and Olsson (2012) and Olsson, Enkvist, and Juslin (2006). Fifteen unique stimuli with four binary cues were used to construct choice pairs in both the linear and nonlinear environment. Table 1.1 shows the cue patterns of all the stimuli together with their criterion value in both environments. The criterion value in the linear environment, y_L , was a linear function of four cues, c_1, c_2, c_3

⁴Software, together with exact instructions and stimuli used in the experiment, is publicly available at the Open Science Framework website: <https://osf.io/3q5if/>. Raw data from the experiment is publicly available on Figshare: <http://dx.doi.org/10.6084/m9.figshare.1585822>.

and c_4 :

$$y_L = 0.1 + 0.4c_1 + 0.3c_2 + 0.2c_3 - 0.1c_4$$

An independent error term was added to both items in each pair, drawn from Normal distribution with a mean of 0 and standard deviation of 0.15. The noise was added to provide probabilistic feedback and further induce the usage of a cue-based strategy (Juslin, Jones, Olsson, & Winman, 2003).

Following Olsson et al. (2006), the criterion value in the nonlinear environment, y_{NL} , was a nonlinear function of the linear criterion values:

$$y_{NL} \approx 4.0508y_L - 0.0367y_L^2 - 110.8225$$

No noise term was added in the nonlinear environment.

The environments were randomly interleaved in the training and test phases. The purpose of the training phase was to allow participants to learn how to solve the tasks. The training phase consisted of four blocks, 84 trials in each block – 44 trials from the linear and 40 from the nonlinear environment – giving 336 trials in total. For the linear environment we created 44 pairs using 10 unique stimuli – all possible combinations except for one pair where the stimuli had identical criterion levels. For the nonlinear environment, we used five unique stimuli and created all possible pairs, 10 in total, and repeated these 10 pairs four times. The stimuli used in the training phase are marked as “Old” in Table 1.1. We used smaller number of unique stimuli in the nonlinear environment to induce people further to adopt an exemplar-based strategy (Olsson et al., 2006).

The purpose of the test phase was to more clearly assess the strategy mix adopted in each environment and to see the extent to which participants generalized what they learned in the learning phase. For the linear environment, we used five new unique stimuli together with old ones to create 18 pairs. Seven pairs with old stimuli from the training phase were repeated four times and the remaining nine pairs that included at least one new stimulus were repeated eight times, giving 116 trials in total. For the nonlinear environment, we selected from the pairs used in Pachur and Olsson (2012) those that maximized the discrimination between cue-based and exemplar based strategies. The resulting 17 pairs include eight new stimuli, together with old ones. Three pairs with old stimuli were repeated four times and the remaining 14 pairs with at least one new exemplar were repeated eight times, giving 124 trials in total. In the whole test phase there were 240 pairs. Participants did not receive feedback on their choices.

We used two different cover stories for the linear and the nonlinear task – poisonous “bugs” and dangerous “comics”. In the bugs story participants had to choose which bug was more poisonous, and in the comics story

Table 1.1 Cue patterns and continuous criterion values of the 15 exemplars used in linear and nonlinear environment in the Experiment.

ID	Cues				Linear env.		Nonlinear env.	
	Cue 1	Cue 2	Cue 3	Cue 4	y	Role	y	Role
1	0	0	0	0	0.10	old	0	new
2	0	0	0	1	0.00	new	0.35	old
3	0	0	1	0	0.30	old	0.62	new
4	0	0	1	1	0.20	old	0.82	old
5	0	1	0	0	0.40	old	0.82	new
6	0	1	1	0	0.60	old	–	–
7	0	1	1	1	0.50	old	–	–
8	1	0	0	0	0.50	new	0.94	old
9	1	0	0	1	0.40	old	1	new
10	1	0	1	0	0.70	new	0.97	new
11	1	0	1	1	0.60	new	0.88	new
12	1	1	0	0	0.80	old	0.88	old
13	1	1	0	1	0.70	old	0.71	new
14	1	1	1	0	1.00	new	0.47	old
15	1	1	1	1	0.90	old	0.16	new

Note. ID = exemplar identification number; env = environment; old = exemplar used in both training and test phase; new = new exemplar that occurs only in the test phase; y = criterion.

they had to choose which comic figure was more dangerous. The stimuli consisted of pictures of either bugs or comic figures, and both bugs and comic figures varied on four binary cues. In bugs – antennae, spots on the back, wings, and legs, were either present or absent. Similarly, in comic figures – hair, ears, nose, and stripes on the shirt, were either present or absent. Pictures of bugs and comics were a subset of those in J. Hoffmann et al. (2014).

The mapping of bugs or comics to the linear and nonlinear environment, and physical features (e.g., hair, ears) to the cues (c_1, \dots, c_4), was determined at random for each participant. For instance, for one participant the first trial might correspond to linear environment represented as a choice between bugs, where c_1 corresponded to the presence of antennae. For another participant, the first trial might correspond to the nonlinear environment represented as well as a choice between bugs, but c_1 corresponded to the presence of wings. Trials from both environments were randomly interleaved for each participant. Order of the trials was random-

ized within each block in the training phase and in the whole test phase. Position of the stimuli on the screen (left or right) was also randomized on each trial.

To proceed to the test phase, participants had to reach 70% accuracy in both environments in the last block of the training phase. When participants did not satisfy this criterion, we provided them with another block of trials and checked their accuracy again. Participants who failed to reach the required level of accuracy after two additional training blocks were not allowed to continue the experiment.

1.2.3 Procedure

Participants completed the experiment on desktop computers, using custom software written in Python and the PsychoPy library (Peirce, 2007). At the beginning of the experiment, participants completed an informed consent form. They then received on-screen instructions about the task and earnings. All instructions were presented in Spanish.

To motivate participants, we told them that while the task would initially be difficult, they could improve with practice. Moreover, depending on their performance they could earn additional money: on every trial they could earn experimental units (EU's) – they gained 10 EU's for a correct choice and lost 10 EU's for an incorrect choice. The exchange rate was 1 euro for 500 EU's. Participants started the experiment with zero EU's and they could see the running total during the training phase, but not the test phase.

We did not provide participants with information on the exact number of rounds in each phase, instead we told them that the experiment would take 60 minutes on average to complete. The test phase was announced at the beginning of the instructions but without specific details, which were provided only at the start of the test phase. Earnings in the test phase were computed in the same way as in the training phase.

1.3 Results

1.3.1 Behavioral analysis

We used the proportion of expected correct choices (choosing the alternative which is expected to have the highest criterion value) in a block of trials as performance measure.

Training phase

Figure 1.1 shows the performance in the learning phase for each environment. Participants performed substantially better than chance already in the first block, achieving a mean accuracy of 0.76 in the linear and 0.68 in the nonlinear environment. In fact, in the linear environment, performance was better than chance even on the very first trial. People tend to have strong prior beliefs that cue-outcome relations are positive and linear (Brehmer, 1974; Busemeyer et al., 1997; Olsson et al., 2006) and the linear environment is consistent with this belief; hence, initial guesses that a bug or comic with more features present is more poisonous or dangerous were correct on average. Overall, participants improved during the training phase, reaching a mean accuracy of 0.85 in the linear environment by the last block. A Wilcoxon signed rank test shows a significant difference in choice accuracy between the first and last block, $M_{diff} = .09$, $Z = 989$, $p \leq .0001$. A similar result holds for the nonlinear environment, where participants achieved 0.90 by the last block, which is significantly higher than performance in the first block, $M_{diff} = .22$, $Z = 1175$, $p \leq .0001$.

In the last training block, performance in the nonlinear environment reached a higher level than in the linear environment, as shown by a Wilcoxon signed rank test on the difference in choice accuracy between the environments, $M_{diff} = .049$, $Z = 262$, $p = .0003$. This indicates that the linear environment was more difficult to learn than the nonlinear environment, at least with the amount of training trials in our experiment. Although there is some evidence that people can learn nonlinear functions better than linear ones (J. Hoffmann et al., 2014; J. A. Hoffmann, von Helversen, & Rieskamp, 2013; von Helversen & Rieskamp, 2008), most studies show the opposite (e.g. Brehmer, 1994; Busemeyer et al., 1997). In our experiment, the small number of exemplars and deterministic feedback used in the nonlinear environment evidently facilitated learning compared to the linear environment.

Note that the results of block five and six are based on responses of a subset of participants who completed an additional one or two blocks in the training phase. Five participants completed two additional training blocks due to poor performance in the linear environment, while 13 participants completed one additional block (six of these due to poor performance in the linear environment). In Figure 1.B.1 in Appendix 1.B we illustrate choice accuracies separately for groups of subjects that did or did not require additional training blocks. While slower learners took more time, by the end of the training phase they achieved performance levels similar to the faster learners. For this reason here and in the rest of the article we plot the re-

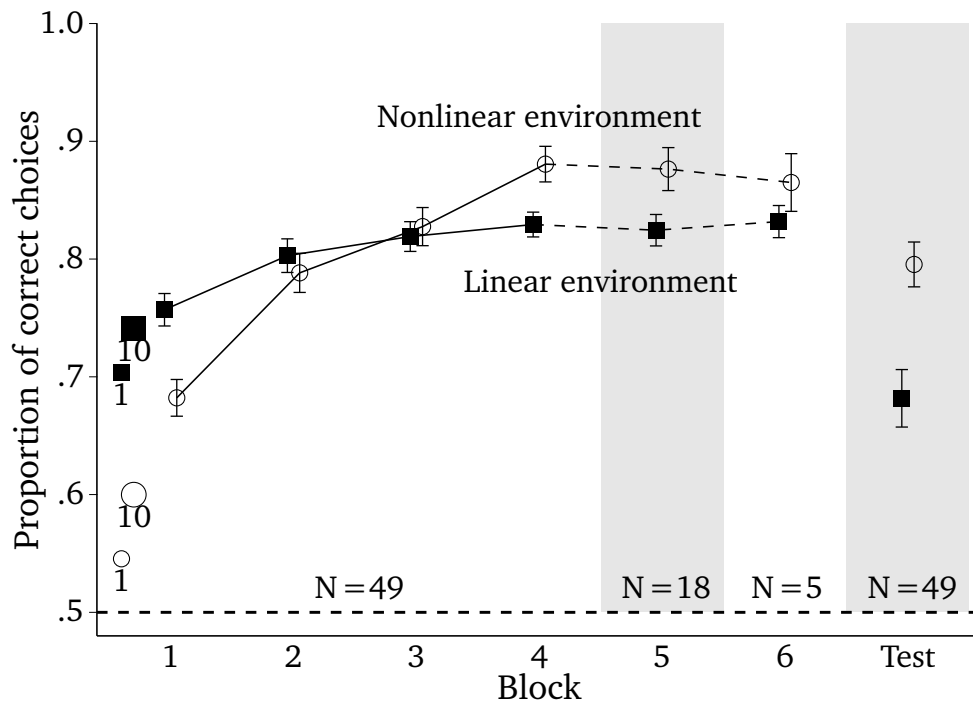


Figure 1.1 Accuracy of participants' choices in blocks of trials in the training phase and in the test phase. Training blocks consist of 44 linear and 40 nonlinear trials. Result for a block is a mean of individual mean accuracies across trials in a block. Results of block five and six come from a subset of participants that took additional one or two blocks in the training phase. Error bars represent standard errors of group means of each block of trials. Points are displaced horizontally to make them easy to distinguish. In addition, we display mean accuracy in the very first trial and across the first ten trials, marked with numbers one and ten, respectively.

sults of all participants together, but point out that some results (i.e., those in block 5 and 6) are based on a subset of participants.

Decrease in response time is another behavioral signature of learning. The time to make a choice in both environments almost halved by the last block in the training phase, from 4.61 to 2.85 seconds, $M_{diff} = 1.76$, $Z = 1225$, $p \leq .0001$. Moreover, on average participants took more time to make a choice in the linear environment, $M_{diff} = .38$, $Z = 990$, $p \leq .0001$ (Wilcoxon signed-rank test).

Test phase

How well did the participants generalize their knowledge from the training phase to the test phase? In the test phase participants encountered pairs with new stimuli and they did not receive feedback on their choices. Mean accuracy in the test phase dropped compared to the last training block: in the linear environment it decreased from 0.85 to 0.68 and in the nonlinear environment from 0.90 to 0.80. The difference in accuracy between nonlinear and linear environments found in the training phase persisted in the test phase, $M_{diff} = .114$, $Z = 290$, $p = .001$. Response times in the test phase were very similar to those obtained in the last training block.

The decrease in performance from training to test phase was expected as the pairs in the test phase contained many new items that participants had not experienced before. The somewhat larger decrease in the linear environment was partly due to the slower learners. As shown in Figure 1.B.1 in Appendix 1.B, those participants who needed two additional training blocks had particularly poor performance in the linear environment. Without these five subjects, the mean accuracy in the linear environment in the test phase increases to 0.71. Interestingly, their performance in the nonlinear environment did not suffer at all.

1.3.2 Identifying the strategies adopted by the participants

We used cognitive modeling to investigate which decision strategies participants relied on in the linear and nonlinear environments. We expected that participants would adopt an exemplar-based strategy in the nonlinear environment and a cue-based strategy in the linear environment. While both classes of strategies can perform well in the linear environment, we expected the probabilistic feedback and fewer repetitions of stimuli to tip the scale in favor of cue-based strategies.

We used several models from the literature as representatives of each type of strategy. The cue-abstraction model (CAM, Pachur & Olsson, 2012) and weighted additive (WADD, Bergert & Nosofsky, 2007; Payne et al., 1993) model are representative cue-based strategies. To represent the exemplar-based strategies we used two versions of the generalized context model (GCM, Nosofsky & Bergert, 2007) that were specifically adapted for pairwise comparison tasks as used here. We describe the models in more detail in the following sections, while the estimation procedure and overview of estimated parameters can be found in Appendix 1.A. In Table 1.2 we list the models we set out to investigate. We examined several

other variants of these models in an exploratory manner, their results and parameters are presented in Appendix 1.A, however, we do not focus on these in the main text.

We used more than one model per type of strategy as we are mainly interested whether a certain *type* of strategy has been adopted. Previous research has shown significant individual variation in which particular cue- or exemplar model describes behavior best in an environment. For instance, some people are better described by the WADD model and some by the CAM in a linear environment (e.g., Pachur & Olsson, 2012). Including several instantiations of cue- and exemplar-based models should reduce the chance of falsely rejecting our hypotheses due to the particular choice of model.

The weighted additive (WADD) model

The weighted additive model (Payne et al., 1993) and take-the-best heuristic (Gigerenzer & Goldstein, 1996) are popular models for describing the behavior in pairwise comparison tasks. We used the probabilistic generalization of these models developed by Bergert and Nosofsky (2007). In the WADD model the probability that A will be chosen over B is given by

$$P(A; A, B) = \frac{(\sum_{a \in FA} w_a)^\gamma}{(\sum_{a \in FA} w_a)^\gamma + (\sum_{b \in FB} w_b)^\gamma},$$

where $\gamma \geq 0$ is a free response scaling parameter and w_j ($0 \leq w_j \leq 1$) are the weights assigned to each individual cue, constrained to sum to 1. FA and FB denote the set of discriminating cues favoring alternatives A

Table 1.2 Overview of the models representing each type of strategy.

Strategy type	Model	# Par.
Cue-based	CAM_u	4
	$WADD$	4
Exemplar-based	$pGCM_{11}^\gamma$	5
	$jGCM_{11}^\gamma$	5

Note. # Par. = Number of free parameters in the model; CAM_u = Unconstrained cue abstraction model; WADD = Weighted additive model; GCM = Generalized context models.

and B , respectively.⁵ Generalized take-the-best (gTTB) is a special case with scaling factor γ set to 1. Although the predictions of two models are equivalent in that case, the implied psychological processes are different. In the main text we present the result for WADD model since gTTB is a special case of WADD. We report results specifically for gTTB in Appendix 1.A.

As these models are based on a linear combination of cues, they are especially well suited for linear environments. This gives them an edge in linear environments, but prevents them from performing well in nonlinear environments. Scaling parameter γ can additionally capture potential inter-individual differences in sensitivity to differences in evidence between alternatives.

Overall, the WADD model had four parameters – γ , w_1 , w_2 and w_3 , while the gTTB model had three parameters – w_1 , w_2 , and w_3 .

Cue-abstraction model (CAM)

The cue abstraction model (Juslin, Jones, et al., 2003; Pachur & Olsson, 2012) is another model that combines evidence in a linear way. Alternatives are evaluated jointly by looking at the difference of each cue value $\Delta c_j = c_{jA} - c_{jB}$, $j = 1, \dots, 4$. The importance of each cue difference is reflected in its cue weight $w_j \geq 0$. The higher the cue weights are, the more they will influence the choice. The probability that alternative A will be chosen over alternative B is given by

$$P(A; A, B) = \frac{e^{\sum_j w_j \Delta c_j}}{1 + e^{\sum_j w_j \Delta c_j}}$$

Essentially, CAM is a logistic regression model without an intercept. It is also similar to the WADD model; the main difference being that CAM transforms the evidence into choice probabilities through a logistic function and allows for more subjectivity in weights. Even though the models produce similar predictions, empirically researchers have found differences in terms of fit to choice behavior (Pachur & Olsson, 2012).

We tested two versions of the model. In CAM_c the weights are constrained to lie between 0 and 1 and to sum to 1, i.e. $0 \leq w_j \leq 1$, and

⁵In our environments some cues have a negative effect on the criterion and the sign of the difference between the cue values of two alternatives needs to be reversed (multiplied by minus 1) whenever the difference is not equal to zero. For each environment we fitted the WADD to the actual winning alternatives with all possible combinations of cue reversals. In the linear environment the WADD with fourth cue reversed performed the best, and in the nonlinear the WADD with second, third and fourth cue reversed was the best. When fitting the model to each individual we reversed the cues according to these results.

$\sum_{j=1}^4 w_j = 1$, while in CAM_u they are unconstrained. The constraint prevents the weights from becoming very large which can reduce overfitting and may help the model to generalize better. Because the constraint implied positive effects for all cues, we reversed the direction of some cues using the same procedure as for WADD and gTTB. We focus on the more general CAM_u and we examined CAM_c in an exploratory manner. Results of CAM_c are reported in Appendix 1.A. CAM_u had four parameters – w_1 , w_2 , w_3 and w_4 , while CAM_c had three parameters – w_1 , w_2 , and w_3 .

The generalized context model (GCM)

The generalized context model is a memory-based exemplar model widely used in category learning (Nosofsky, 1986), but also for continuous judgments (Juslin, Jones, et al., 2003; Speekenbrink & Shanks, 2010). GCM assumes that previous experiences are stored as instances in memory and when a new situation arises, a prediction is generated by combining exemplars stored in memory according to their similarity to the new situation. The similarity component allows the model to mimic both linear and non-linear functions, which is why it can perform well in both types of environment.

We used the GCM developed for pairwise comparison tasks by Nosofsky and Bergert (2007). The model compares the probe (the current pair of alternatives) to the previously encountered exemplars (pairs of alternatives) that are kept in the memory. The model determines how similar the probe p is to each exemplar i through an exponentially decreasing function of the distance $d(p, i)$ between the probe and exemplar

$$S(p, i) = e^{-\lambda d(p,i)^q},$$

where $0 \leq \lambda \leq 10$ is a sensitivity parameter and $q = 1$ for the exponential, and $q = 2$ for the Gaussian similarity function. The distance function is the generalized Minkowski distance

$$d(p, i) = \left[\sum_{j=1}^J w_j |c_{pj} - c_{ij}|^r \right]^{1/r}$$

with Minkowski parameter r being either 1 or 2. c_{pj} and c_{ij} are the cue values of probe p and exemplar i , respectively, for cue j . w_j , $0 \leq w_j \leq 1$, are attention weights assigned to each individual cue, constrained to sum to 1. The more closely the cue values of the probe and the exemplar correspond to each other, the smaller the distance between them and the greater the similarity.

Nosofsky and Bergert (2007) proposed two versions, depending on how the decision situation is represented. In what we call a “paired” representation, the model assumes that winning alternatives are stored as exemplars of a winners category, W , while losing alternatives are stored as exemplars of a losers category, L . Similarities to the winners and losers categories for alternative A are computed separately as

$$S(A, W) = \sum_{i \in W} s(A, i)$$

and

$$S(A, L) = \sum_{i \in L} s(A, i)$$

The relative evidence for alternative A is given by

$$G_A = \frac{S(A, W)^\gamma}{S(A, W)^\gamma + S(A, L)^\gamma}$$

where $0 \leq \gamma \leq 10$ is a free scaling parameter. Finally, the probability that alternative A is chosen is given by

$$P(A; A, B) = \frac{G_A}{G_A + G_B}$$

The “joint” representation version of the model similarly assumes that *pairs* of alternatives are stored as exemplars in winners and losers categories. If the feedback indicates that alternative A is a correct choice, then the pair AB is stored in the winners category as a vector where alternative B is concatenated to alternative A , while a vector BA , where A is concatenated to B , is stored in the losers category. The attention weights are the same for both alternatives and in this representation they are simply duplicated and concatenated to form a vector of the same length as pairs AB and BA . The probability that alternative A is chosen is given by

$$P(A; A, B) = \frac{S(AB, W)^\gamma}{S(AB, W)^\gamma + S(AB, L)^\gamma}$$

where $S(AB, W)$ and $S(AB, L)$ represent similarities of the pair AB to each exemplar in the winners and losers categories, respectively, based on the same distance and similarity computations as paired representation.

We focused on GCM versions with Minkowski distance parameter $r = 1$ and exponential similarity function $q = 1$, which we report in the main text. This model with the paired representation is denoted as $pGCM_{11}^\gamma$ and the

version with the joint representation as $jGCM_{11}^Y$. Both models had a total of five parameters: λ , γ , w_1 , w_2 and w_3 . Given the binary nature of features in our task, Minkowski and similarity parameters should not matter that much, but we explored both paired and joint versions with different combinations these parameters. In one variant we also set the scaling parameter to one. Results of all models are presented in Appendix 1.A.

Best fitting models in each environment

Figure 1.2 summarizes the test set generalization results of the selected models. Following Wagenmakers and Farrell (2004) we computed log likelihood (LL) weights for each of the four models in our candidate set, separately for each environment. LL weights allow for better interpretation of observed relative differences in model performances. Weight can be interpreted as the probability that a particular model is the best model, given the data and the set of models in the comparison set. See Appendix 1.A for more details on LL weight computation.

As can be seen in the figure, on average, CAM_u predicted participants' choices in the linear environment best, while in the nonlinear environment $jGCM_{11}^Y$ and $pGCM_{11}^Y$ performed about equally well, with CAM_u closely trailing behind. In the linear environment CAM_u has the greatest probability of being the best model among the four (0.57), being more than two times more likely than $jGCM_{11}^Y$ (0.16) and $pGCM_{11}^Y$ (0.23). $WADD$ fared poorly, having only 0.04 probability of being the best model.

For the nonlinear environment the results are less clear. Evidence is favoring $jGCM_{11}^Y$ and $pGCM_{11}^Y$, with probabilities of 0.37 and 0.35 respectively, but only with a small margin over CAM_u with probability of 0.28. The finding that CAM_u performed well also in the nonlinear environment shows that a subset of participants did not adapt well and tried to apply a cue-based strategy in the nonlinear environment too.

Classifying individuals according to the strategy used

Average results do not tell us exactly how well adapted the participants are. We classified participants as users of those strategies that best predicted their choices in the test phase, separately for the linear and nonlinear environment (numbers denoted with N in Figure 1.2). In the linear environment most participants were best described by one of the cue-based strategies. In the nonlinear environment most participants were best described by one of the exemplar-based strategies, although the number of participants best described by the CAM_u model was also large. Thus, for

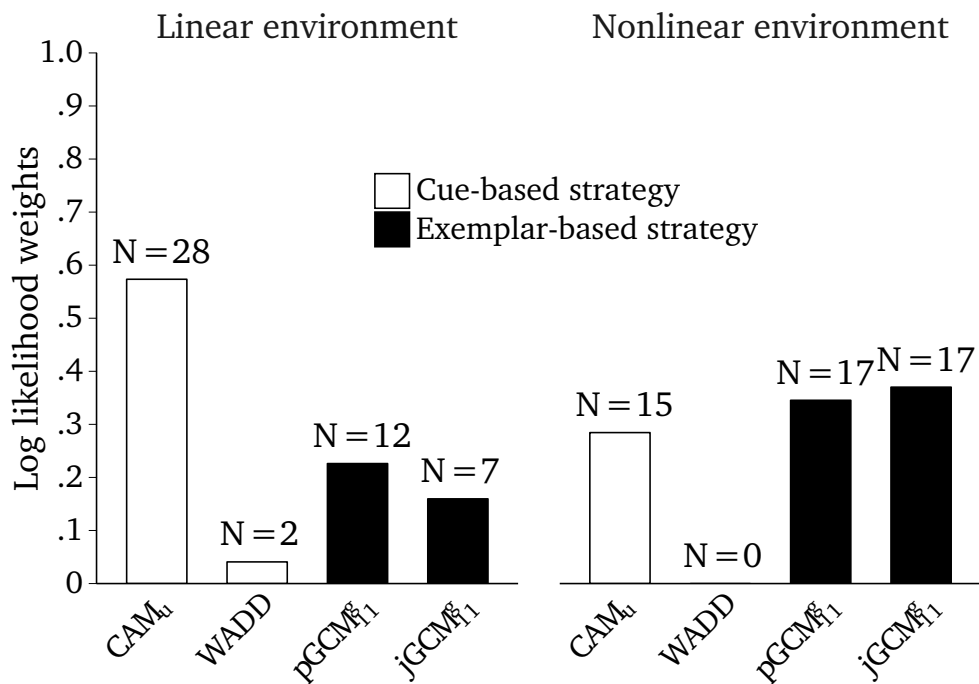


Figure 1.2 Model performance in predicting individual choices in the test phase, presented separately for each environment. Performance is expressed as mean log likelihood weight across participants, computed for these four models in the comparison set. Numbers above the bars represent number of participants whose choices in the test phase were best predicted by each of the models. Most people were best predicted with cue-based strategies in the linear environment, and with exemplar-based strategies in the nonlinear environment.

a majority of the participants, we found evidence that they were able to adaptively switch between strategies from trial to trial, as they encountered different environments.

In the linear environment, 19 participants were best described by one of the exemplar-based strategies. Recall that in the linear environment both classes of strategies can achieve good performance, while in the nonlinear environment only exemplar-based strategies can achieve good performance. In the nonlinear environment there were 15 participants that were best predicted by CAM_u . These participants either did not learn to select a more appropriate strategy for the nonlinear environment or failed to separate decision situations into two different environments.

Overall, 12 participants used exemplar-based strategy in both environments, while 8 used cue-based strategy exclusively. 29 participants adopted

exemplar-based strategy in one environment and cue-based in the other, 7 of which used them in unexpected fashion – exemplar strategy in the linear and cue-based strategy in the nonlinear environment.

1.3.3 Contextual strategy selection learning

The previous analysis showed that the majority of participants successfully adopted a cue- or exemplar-based strategy in the linear environment, and an exemplar-based strategy in the nonlinear environment. However, that analysis did little to inform how the associations between strategies and environments were learned.

To describe this process in our experiment, we used a contextual version of the reinforcement learning based SSL model (Rieskamp & Otto, 2006). In contrast to the original SSL, we assume that people form two categories of situations based on perceptual features – the “bugs” and the “comics” category. Our version of SSL then learns which strategy is more successful separately for the “bugs” and the “comics” category. Moreover, our strategies are probabilistic and the repertoire contains exemplar-based strategies that can perform well in the nonlinear environments. We fitted the model to the training phase for each individual, examined what strategies were adopted and how well the choices in the test phase are predicted with the particular strategy mix adopted in the training phase.

The model

The original SSL model (Rieskamp & Otto, 2006) assumes that people have a repertoire of strategies they can apply to the decision problem at hand. A crucial assumption in the model is that rewards obtained from the choices reinforce the strategies instead of specific alternatives. The main implication of the model is that the strategy that on average leads to higher rewards will be chosen more often.

In the contextual SSL (CSSL) we assume that the decision problem that is encountered can be a member of one of E environments. The first step is then to categorize the problem as belonging to one of the environments, $e \in 1, \dots, E$. We assume there is a vector of contextual features \mathbf{x} and that there exist a mapping, $f(\mathbf{x})$, from contextual features to environment categories, $e \in 1, \dots, E$. The contextual features can take any form, for example the time available for making a decision, cue weights, (non)compensatoriness of the cue weights, or simply perceptual features of the alternatives. In light of our discussion in the introduction, what is relevant is that problems are differentiated – there is no further meaning

ascribed to any of the categories. Our experiment was designed so that the mapping function, $f(\mathbf{x})$, is particularly simple; we made it highly likely that participants use perceptual features – bugs and comics – to partition the problems into two categories. And as this indeed is a useful way to partition the problems, they are likely to stick with it. Hence, for the purposes of the present experiment, we assume that the model employs a simple deterministic function from a single contextual feature, $x \in \{\text{bugs, comics}\}$, to two environments, $e \in 1, 2$.

In the second step, the model chooses a strategy from the repertoire where strategy expectancies are conditional on the environment. Expectancy is a measure of preference for a certain strategy in an environment. The probability of choosing strategy s from repertoire S in environment e at trial t is defined as

$$P_t(s|e) = \frac{Q_t(s|e)^\theta}{\sum_s Q_t(s|e)^\theta},$$

where $Q_t(s|e)$ is the expectancy of strategy s in environment e at trial t and θ is a sensitivity parameter. When $\theta = 1$ we obtain Luce's (1959) choice rule. Initial expectancies are defined by

$$Q_1(s|e) = r_{max} w \beta_s,$$

where $0 < w < 10$ is an initial association parameter, r_{max} is the maximum reward that can be obtained with a correct decision in the task (10 experimental points in our case), and the β_s parameter describes the initial bias toward a certain strategy (with $0 \leq \beta_s \leq 1$, and $\sum_s \beta_s = 1$). In addition, if $Q_t(s|e)$ falls below some minimum level ρ due to negative payoffs, it is set to $\rho = 0.0001$.

After applying the selected strategy a reward is obtained and this reward is the basis for updating the expectancies of the strategies:

$$Q_t(s|e) = Q_{t-1}(s|e) + I_{t-1}(s|e)r_{t-1}(s|e)$$

where $I_{t-1}(s|e)$ is an indicator function, and $r_{t-1}(s|e)$ is the reinforcement.⁶ In our case reinforcement is the payoff that the strategy produces, either 10 or -10 experimental points. We implemented two types of indicator function: deterministic and proportional. The deterministic indicator function equals 1 if the strategy s was applied, and 0 if it was not. How do we infer that the strategy was chosen? If the strategy prediction coincides with

⁶The update equation looks different than the usual delta learning rule (Rescorla & Wagner, 1972). This works equally well as in this context the absolute value of the strategy expectancy does not matter much, only relative values play a role. This learning rule might then obviate the need for the temperature parameter in the choice rule above.

the participant’s choice (that is, if the probability of choosing the alternative is greater than 0.5), and if other strategies predict a different choice, we assume that the participant has chosen that strategy. If more than one strategy prediction coincides with the participant’s choice, we assume that $I_{t-1}(s|e)$ equals the probability with which the model predicts the selection of those strategies in a given environment, $P_t(s|e)$. In this case, the strategy preferences do not change as ratio of expectancies will remain constant.

In the original SSL model only a deterministic indicator function was used since the authors considered only deterministic strategies. The proportional indicator function takes the probability with which each strategy predicts the participant’s choice and produces a weight normalized by the sum of the probabilities. This mechanism provides a more gradual strategy learning process. Since this mechanism would lead to smaller relative differences between the strategy expectancies, we used proportional indicator function in combination with a sensitivity parameter θ in the choice rule as a free parameter. Since we do not directly observe which strategy was employed, the proportional indicator function makes a more reasonable choice than the deterministic ones.

We assumed there are two strategies in the repertoire – a representative of exemplar-based strategies and a representative of cue-based strategies. Following the results of modeling the test phase choices, we chose $jGCM_{11}^Y$ to be the representative of exemplar-based strategies, and CAM_u as representative of cue-based strategies. Strategies also have free parameters. This is another deviation from the original SSL model, besides partitioning according to the observable features and proportional indicator function. In CAM_u cue weights are free parameters, and in $jGCM_{11}^Y$ γ , λ and attention weights are free parameters. Hence, learning occurs on multiple levels – adapting the strategy mix at the strategy selection level, and adapting the strategies themselves to each environment.⁷

Overall there were three parameters on the strategy learning level, the initial association parameter w , initial strategy bias parameter β_s and sensitivity parameter θ . We varied whether a deterministic or proportional indicator function was used, marked with prefix d and p respectively. When a deterministic indicator function was used we fixed θ to one, reducing the number of parameters by one.

⁷Note however that, for the sake of simplicity, the models that we use to represent the strategies are not learning models that adapt their parameters on a trial-by-trial basis. Instead, for each individual we estimate the parameters of each model and environment separately, and then use them in the CSSL model.

Results of modeling the strategy selection learning

Modeling results in terms of projective fit in the test phase are depicted in Figure 1.3. Details of the fitting procedure can be found in Appendix 1.A, and estimated parameters in Table 1.A.2 in the same Appendix. We compared the context sensitive CSSL model with the original SSL model (for details, see Rieskamp & Otto, 2006) containing the same two strategies in the repertoire and governed by the same strategy selection parameters. We also fitted single strategies CAM_u and $jGCM_{11}^Y$ to choices from both environments, to investigate how the strategy selection models compare to simpler explanations using single strategies.

We can see that the contextual versions of SSL fared better than the original SSL and single strategy models. $dCSSL$ model with deterministic indicator function predicted participants' choices in the test phase the best, reaching probability of 0.26 of being the best model among the six we have considered. $pCSSL_\theta$ performed worse, reaching probability of 0.17, but still better than $dSSL$ and $pSSL_\theta$ models that have probabilities 0.11 and 0.16 respectively of being best models. Interestingly, the version with the deterministic indicator function had a worse performance in this case. Numbers above the columns indicate the number of individuals best fitted with the model. These show that 21 participants are best described by one of the CSSL models, while 14 are best described with one of the SSL models. Although CSSL models predict participants' choices better, the advantage over simpler SSL models does not look immediately impressive. However, the advantage is considerable given that CSSL models are more complex, effectively having twice as many parameters (when strategy-specific parameters are taken into account) and still perform well on the held-out sample.

With respect to the single strategy models, given that choices of many participants in the nonlinear environment were best predicted with the CAM_u model, we expected CAM_u to perform well when fitted to the whole data. Indeed, CAM_u has probability of 0.19 of being the best model, second only to the $dCSSL$ model, and nine participants were best predicted with this model. $jGCM_{11}^Y$ model performed the worst, reaching probability of 0.11 and predicting choices of five participants the best. Overall, single strategy models performed as well as the SSL (but not CSSL) models.

Which strategies do CSSL models adopt in each of the environments? Figure 1.4 shows the evolution of probability of choosing the exemplar-based strategy (as represented by the $jGCM_{11}^Y$ model) over blocks of trials, presented in terms of averages across the participants. As we expected, by the end of the training phase the exemplar strategy was the preferred one in the nonlinear environment; $dCSSL$ and $pCSSL_\theta$ models ended up with

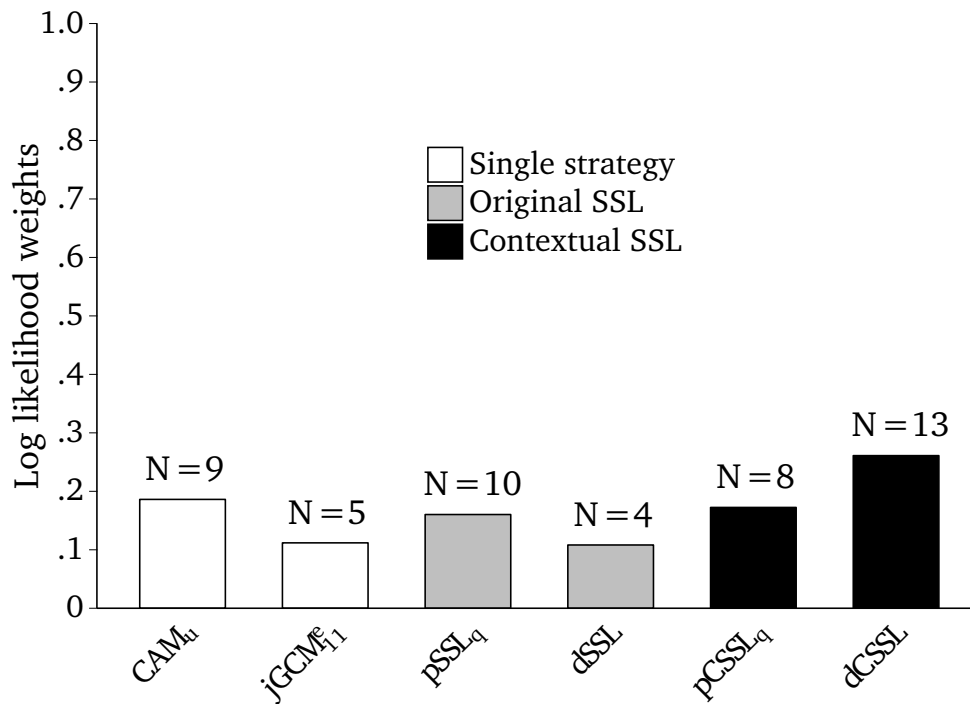


Figure 1.3 Model performance in predicting choices in the test phase for contextual strategy selection learning models (CSSL), original strategy selection learning models (SSL), and two single strategy models – CAM_u and $jGCM_{11}^Y$. Performance is expressed as mean log likelihood weight across participants, computed for these six models in the comparison set. Numbers above the bars represent number of participants whose choices in the test phase were best predicted by each of the models.

probabilities of 0.73 and 0.71 of choosing the exemplar strategy. There is very little difference between the models in terms of evolution of strategy preferences as well. Inspecting the end-of-training strategy mixtures for both CSSL models, most participants can be described as having a higher probability to use the cue-based strategy in the linear environment and the exemplar strategy in the nonlinear – 31 for $dCSSL$ and 39 for $pCSSL_{\theta}$ model. Fewer participants are described with mixtures that favor exemplar strategies (12 for $dCSSL$ and 5 for $pCSSL_{\theta}$) or cue-based strategies (6 for $dCSSL$ and 4 for $pCSSL_{\theta}$) in both environments. Only one participant was described by the $pCSSL_{\theta}$ model as preferring the exemplar strategy in the linear and the cue-based strategy in the nonlinear environment.

The parameters for the initial preference toward a strategy, w and β , were shared across environments. For most participants parameter values

indicate a weak initial preference for the cue-based strategy as both CSSL models started with a weak initial preference for the CAM_u model. In the linear environment this preference was kept more or less constant throughout the training phase (ending at probabilities of 0.35 for $dCSSL$ model, and 0.39 for $pCSSL_\theta$). In the nonlinear environment the change in strategy expectancies was strong and steered rapidly in favor of the exemplar strategy. There are substantial deviations in the fifth and sixth block of the training phase. This is due to several slower learners on which these data points are based, that had different evolution of strategy mixtures.

The difference between the environments in which strategy is mostly adopted is the source of improvement offered by CSSL in comparison to the SSL and single strategy models. It results in a weak preference for the cue-based strategy in the linear environment and strong preference for the exemplar-based strategy in the nonlinear environment. In contrast, SSL can learn only a single strategy mixture that works best on average over all environments and here both SSL models develop a strong preference for the exemplar-based strategy. These differences can be seen more clearly in Figure 1.B.2 in Appendix 1.B, where model performance is shown separately for the environments. We can see that because the CSSL models predict choices in the linear environment much better than the SSL models, whose performance suffers in the linear environment.

1.4 Discussion

We presented an experiment where participants were asked to solve two interleaved choice tasks. In one task (the linear environment), a cue-based strategy was more appropriate while in the other (the nonlinear environment), an exemplar-based strategy was more appropriate. During the training phase, participants learned to solve the tasks well. Their choices in the test phase, where they also encountered previously unseen alternatives, were critical for our modeling approach. In our first modeling analysis, using an out-of-sample prediction criterion, we found that on average the cue-based CAM_u model predicted participants' choices in the linear environment best, while the exemplar-based $jGCM_{11}^Y$ predicted choices best in the nonlinear environment. This modeling evidence does not rely on assumptions of how strategy preferences are learned. Thus, our results show that majority of the participants in our experiment have appropriately adopted a cue-based strategy in the linear environment and an exemplar-based strategy in the nonlinear environment, and were able to flexibly shift between them as they encountered a decision problem from one environment to

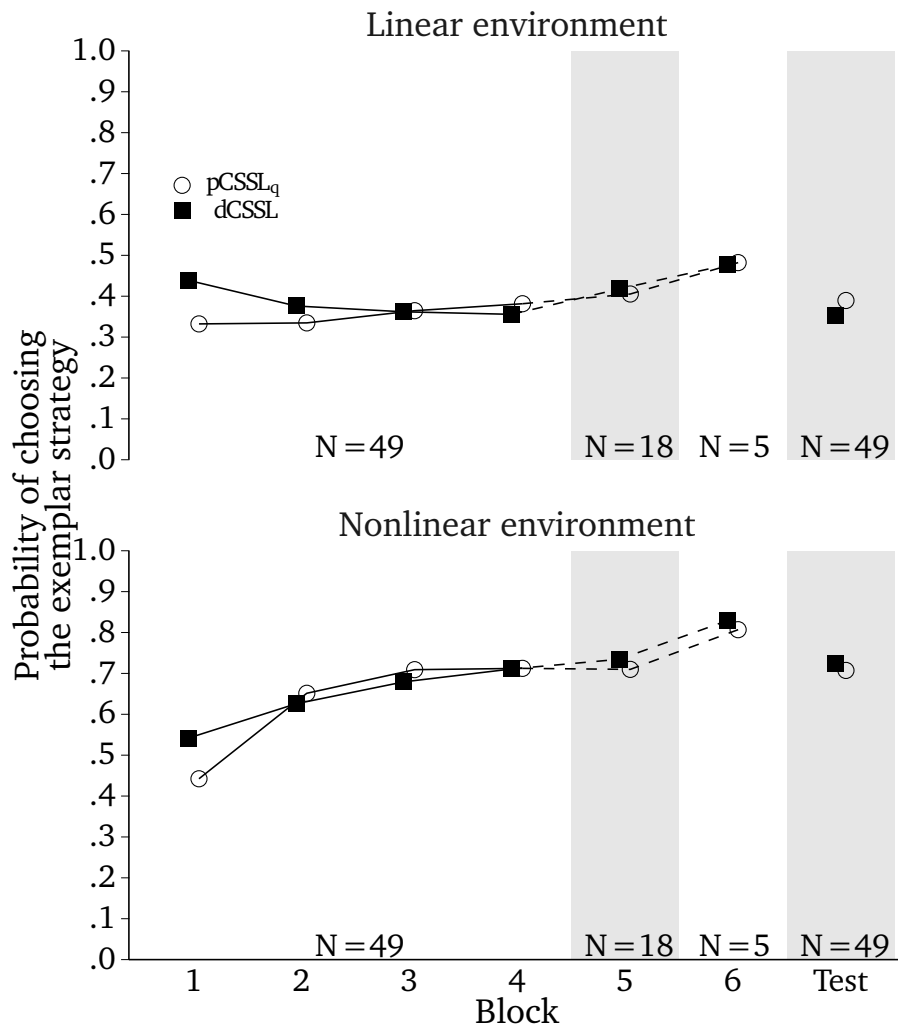


Figure 1.4 Strategies adopted by CSSL models over time in each environment in the training phase, expressed in probabilities of choosing the exemplar strategy, $jGCM_{11}^Y$. Probability that the CAM_u model is selected is one minus probability of choosing the exemplar strategy. Test phase strategy mixture is simply the mixture from the last trial in the training phase. Points are averages across participants, where for each participant an average across the block was taken. Results of block five and six come from a subset of participants that took additional one or two blocks in the training phase.

another. However, a substantial number of participants appeared to use a cue-based strategy (CAM_u) in the nonlinear environment. These participants either failed to separate the two environments and adopted the

same strategy in both, or they simply failed to adapt adequately to the nonlinear environment (Brehmer, 1974; Busemeyer et al., 1997; Olsson et al., 2006). We favor the latter explanation. Stojić, Olsson, and Analytis (2016) find that differences in speed of learning could account for the inter-individual variation in strategy adoption *within* conditions. Hence, we believe these participants were slow learners who would have adopted an exemplar-based strategy given sufficient experience.

We found that participants' choice accuracy in both environments decreased substantially from training to test phase. This drop was not expected in the linear environment. One of the advantages of cue-based strategies over exemplar-based strategies is their ability to accurately extrapolate outside the range of experienced exemplars (Busemeyer et al., 1997). If participants truly used a cue-based strategy in the linear environment they should have no difficulty generalizing their knowledge to the new items in the test phase. However, there were important differences in difficulty between the environments, so this makes the comparison harder. Moreover, cue-based strategies can be poor at extrapolation as well, depending on the specifics of the learning process (McDaniel & Busemeyer, 2005) and if the weights have not been learned sufficiently well.

In our second modeling analysis we have fitted a contextual version of the strategy selection learning (CSSL) model, with CAM_u and $jGCM_{11}^Y$ in the strategy repertoire, representing cue-based and exemplar-based strategies. The model implements a trial-and-error mechanism by which participants learn over time to associate environments to the strategy which works best within it. The CSSL model predicted the behavior of the participants better than simpler explanations in the form of the original SSL and single strategy models. The evolution of strategy expectancies in the CSSL was consistent with our earlier findings identifying which strategy was used in each environment. Our CSSL model shows an initial preference for a cue-based strategy, as also found in previous studies (Rieskamp & Otto, 2006). In the linear environment this preference is maintained, while in the nonlinear environment it changes substantially throughout the training phase in favor of an exemplar-based strategy.

In this modeling analysis, many of the participants best fitted with SSL or single strategy models were also the ones that incorrectly used a cue-based strategy in the nonlinear environment, as shown in the first modeling analysis. However, there are inconsistent classifications as well, e.g., participants classified as adaptive on the basis of the first modeling exercise that were not best predicted by a CSSL model in the second modeling analysis. Such differences are most likely due to using only CAM_u and $jGCM_{11}^Y$ in CSSL's strategy repertoire. This was necessary for practical reasons, but it

resulted in forcing these two strategies on all participants, while in the first modeling analysis participants were fitted with several models from both cue- and exemplar-based class. Some differences were also expected since the two modeling approaches differ substantially.

In several previous studies it was shown that people adopt different strategies in different environments (e.g., Karlsson et al., 2007; Pachur & Olsson, 2012). Crucially, however, their experiments employed between-subject designs such that single participants were not exposed to multiple environments. Consequently, they were concerned less with the mechanisms through which strategies are adopted, focusing instead on identifying the dominant strategy adopted by participants. (J. Hoffmann et al., 2014) is one of the rare studies that used a within-subject design. In their experiments participants performed a multiple cue probability learning task belonging to a linear environment or multiplicative environment, although participants were exposed to them in separate blocks. They found that participants' responses in the linear environment were best described with a linear regression model, while responses in the multiplicative environment were best described with an exemplar model. However, they investigated the role of episodic memory in strategy adoption and did not examine the influence of environment classification on strategy selection, or attempt to model the mechanism behind adopting the strategies in multiple environments.

Lieder and Griffiths (2015) also used a within-subject design, aiming to shed more light on the strategy selection mechanism. Based on their results they concluded that their feature-based cost-benefit model described participants' behavior better than the reinforcement learning approach of the SSL model (Rieskamp & Otto, 2006). In their experiments they used two similar environments – compensatory and noncompensatory, both of which are linear. Moreover, they have presented cue validities to the participants that made it easy to estimate the accuracy of each strategy. On the other hand, our study favors the reinforcement learning approach as it can deal with much more complex situations, where the importance of features to classify environments still has to be learned. A cost-benefit based strategy selection approach such as the model proposed by Lieder and Griffiths (2015) would find it difficult to explain how people solve the strategy selection problem in this setting.

In contrast to the cost-benefit model developed by Lieder and Griffiths (2015), the CSSL model does not require predetermined features to classify environments. For example, the CSSL model does not need to know the statistical properties of the environment to classify it as one in which an exemplar strategy would work best. All that is needed is that decision

situations are separated into different categories, which strategy works best in that category can be learned. In our experiment we used a very clear visual feature that participants could use to partition the situations into two groups – one situation was always represented as deciding between “bugs”, and the other between “comic” figures. One could argue that we have made the partitioning task too easy and the task lost much on its external validity. The present study can be thought of as a proof of principle – if participants had difficulty with associating different strategies to two easily distinguishable environments, there would be little hope that they would be able to do it in more complex realistic scenarios. In future work, we aim to test the model in situations where the features distinguishing environments are more subtle.

Another concern relates to the scalability of the reinforcement learning approach to such situations. When there are many potential features to distinguish between environments, there is a danger of identifying too many categories. Such over-categorization is wasteful as it reduces the amount of experience with each category, so that learning which strategy works best for that category is difficult. A direction we aim to explore in the future is to combine reinforcement learning with a similarity-based mechanism to generalize over categories. For example, if one learns to prefer a certain strategy when deciding between apples, then based on some similarity measure you might start with a similar strategy when deciding between oranges, but perhaps not when deciding between televisions.

Finally, it is important to note that the issue of categorizing environments extends to any “cognitive toolbox” theory that assumes the existence of a repertoire of mechanisms that can be selected. Such theories are gaining in popularity and can be found in many areas in psychology, from developmental psychology to categorization (for a recent overview, see Scheibehenne et al., 2013). Dual system theory can also be seen as a toolbox type of theory (e.g. Kahneman, 2011; Shiffrin & Schneider, 1977), where there are two tools in the toolbox – System 1 and System 2 – and the question is how you choose which one to apply when facing multitudes of problems. The problem of categorizing environments is intimately connected to the strategy selection problem, and as we argued above, solving it requires more than a straightforward extension of strategy selection in a single environment. Without successfully addressing both categorization and strategy selection, the toolbox approaches to cognition will be found lacking.

Appendix

1.A Parameter Estimation and Model Selection

1.A.1 Identifying the strategies adopted by the participants

All choice models were fitted to each individual participant's choices in the last two blocks in the training phase, separately for trials in the linear environment and nonlinear environment. Parameters were found by minimizing the log likelihood of the data given the choice probabilities predicted by the model. The likelihood of the data set, L , of model i is given by

$$L(data|M^i) = \prod_{t=1}^T P(M_t^i = C_t) \quad (1.1)$$

where T is total number of trials being modeled, and $P(M_t = C_t)$ is probability of model making the same choice as participant made in trial t . Number of trials was 88 for the linear and 80 for the nonlinear environment. Optimization was done on the log transformed likelihood, $-\ln(L(data|M^i))$, using the Nelder–Mead simplex algorithm implemented in the `optim` function in R (R Core Team, 2015).

For model selection we used a version of generalization criterion (Busemeyer & Wang, 2000) – for each model we used parameters estimated on the training data from one environment and predicted choices in the test phase of the same environment that were designed to discriminate better between the CAM and GCM models. As a measure of model performance we used log transformed likelihood, while for model comparison we used log likelihood weights (LL weights), following Wagenmakers and Farrell (2004). Similar to AIC or BIC weights, LL weights is a simple transformation of raw log likelihood scores that can be directly interpreted as conditional probabilities for each model. From the differences in log likelihoods

we obtain an estimate of the relative likelihood L of the model i by

$$L(M_i|data) \propto \exp\{\ln(L(data|M^i)) - \ln(L(data|M^{min}))\} \quad (1.2)$$

where $L(data|M^{min})$ is the likelihood of the model in our comparison set with the minimum likelihood, i.e. the best model. Then we normalize the relative model likelihoods to obtain the LL weights

$$w_i(LL) = \frac{L(M_i|data)}{\sum_{k=1}^K L(M_k|data)} \quad (1.3)$$

where K is the number of models in the comparison set. This makes the weights dependent on models that are being compared, stressing the relative aspect of the model comparison. We have always compared four models – CAM_u , $WADD$, $pGCM_{11}^Y$ and $jGCM_{11}^Y$, that we set out to investigate as primary models, even though we did fit more than these four. Importantly, LL weights allow for better interpretation of observed differences in model performances. Weight w_i can be interpreted as the probability that M_i is the best model, given the data and the set of models in the comparison set.

1.A.2 Contextual strategy selection learning

The fitting procedure is the same as in identifying the strategy used by participants in the linear and nonlinear environment, however the models were fitted to all blocks in the training phase and both environments jointly. When estimating parameters for strategy selection models, SSL and CSSL, we fixed the strategy parameters – for SSL models to the ones estimated for single strategies (CAM_u and $jGCM_{11}^Y$ fitted to both environments), and for CSSL models the parameters estimated according to the procedure from the previous section. This was implemented on individual level. Model comparison followed the procedure described in the previous section, but here six models comprised the comparison set – CAM_u , $jGCM_{11}^Y$, $dSSL$, $pSSL_\theta$, $dCSSL$ and $pCSSL_\theta$.

Table 1.A.1 Summary of parameter estimates for models in identifying the strategies adopted by the participants in each environment together with their performances in the test phase as indicated by their negative log transformed likelihood, $-\log(L)$. We report means and standard deviations in parenthesis for each parameter. For all models except CAM_u only three weight parameters were free parameters, the fourth was constrained by those three.

Envir.	Model	#	$-\log(L)$	γ	λ	w_1	w_2	w_3	w_4
Linear	CAM_u	4	81 (95)	-	-	12.88 (55.38)	7.11 (13.92)	4.64 (17.19)	1.03 (6.53)
	CAM_c	3	72 (9)	-	-	0.43 (0.36)	0.55 (0.35)	0.02 (0.06)	0 (0)
	$WADD$	4	530 (454)	∞	-	0.44 (0.38)	0.48 (0.4)	0.06 (0.09)	0.02 (0.03)
	$gTTB$	3	532 (455)	-	-	0.43 (0.34)	0.47 (0.36)	0.08 (0.09)	0.02 (0.03)
	$jGCM^j_1$	5	69 (28)	11.16 (8.21)	7.98 (7.83)	0.22 (0.21)	0.22 (0.23)	0.37 (0.32)	0.19 (0.27)
	$jGCM^j_2$	5	71 (23)	10.69 (8.01)	9.61 (8.29)	0.23 (0.17)	0.25 (0.22)	0.36 (0.29)	0.16 (0.19)
	$jGCM^j_{21}$	5	70 (23)	12.15 (8.22)	7.91 (7.98)	0.27 (0.3)	0.23 (0.28)	0.37 (0.36)	0.13 (0.25)
	$pGCM^j_{11}$	4	72 (7)	-	20 (0)	0.32 (0.11)	0.33 (0.14)	0.25 (0.12)	0.1 (0.11)
	$pGCM^j_1$	5	72 (27)	10.47 (7.07)	14.51 (7.7)	0.39 (0.3)	0.18 (0.24)	0.24 (0.28)	0.19 (0.26)
	$pGCM^j_{11}$	5	71 (27)	11.46 (6.87)	14.26 (7.55)	0.32 (0.2)	0.24 (0.22)	0.27 (0.25)	0.18 (0.19)
	$pGCM^j_2$	5	77 (36)	10.76 (7.06)	12.76 (7.59)	0.42 (0.38)	0.18 (0.29)	0.23 (0.31)	0.17 (0.29)
	$pGCM^j_{21}$	4	175 (147)	-	-	9.32 (8.32)	-10.92 (9.13)	-0.35 (0.77)	-2.92 (3.17)
	CAM_c	3	73 (9)	-	-	0.76 (0.28)	0.07 (0.11)	0.15 (0.23)	0.02 (0.08)
	$WADDt$	4	538 (360)	∞	-	0.65 (0.38)	0.12 (0.1)	0.14 (0.23)	0.09 (0.17)
	Nonlinear	$gTTBt$	3	500 (322)	-	-	0.6 (0.27)	0.21 (0.11)	0.13 (0.21)
$jGCM^j_1$		5	181 (185)	6.61 (7.16)	12.77 (7.47)	0.16 (0.22)	0.37 (0.31)	0.25 (0.24)	0.22 (0.23)
$jGCM^j_2$		5	239 (304)	6.46 (7.53)	15.13 (6.78)	0.18 (0.2)	0.32 (0.23)	0.31 (0.18)	0.19 (0.2)
$jGCM^j_{21}$		5	184 (373)	7.19 (7.88)	12.39 (7.36)	0.22 (0.33)	0.42 (0.38)	0.21 (0.31)	0.14 (0.21)
$pGCM^j_{11}$		4	81 (20)	-	19.87 (0.48)	0.1 (0.18)	0.34 (0.15)	0.3 (0.25)	0.26 (0.19)
$pGCM^j_1$		5	116 (72)	11.08 (7.11)	10.45 (7.53)	0.18 (0.21)	0.37 (0.35)	0.34 (0.29)	0.11 (0.14)
$pGCM^j_2$		5	142 (100)	10.32 (6.6)	12.96 (6.46)	0.16 (0.19)	0.23 (0.26)	0.41 (0.18)	0.2 (0.16)
$pGCM^j_{21}$		5	112 (102)	13.52 (7.43)	10.45 (7.54)	0.13 (0.23)	0.55 (0.42)	0.25 (0.35)	0.06 (0.13)

Note. CAM_u = Unconstrained cue abstraction model; CAM_c = Constrained cue abstraction model; $WADD$ = Weighted additive model; $gTTB$ = generalized take-the-best model; GCM = Generalized context model, prefix p and j denote paired and joint representation respectively, first number in subscripted suffix denotes Minkowski distance parameter while the second denotes similarity function parameter, superscript suffix j denotes whether scaling parameter was used as well; Envir. = Type of environment; # = Number of parameters; w_{1-4} = Weight parameters, for GCM models these are attention parameters.

Table 1.A.2 Summary of parameter estimates for strategy selection models. We report means and standard deviations in parenthesis for each parameter. *SSL* models used strategy parameters for CAM_u and $jGCM_{11}^\gamma$ for which means are reported in the first two rows of this table, while *CSSL* models used environment-specific parameters for which means are reported in Table 1.A.1. For $jGCM_{11}^\gamma$ only three weight parameters were free parameters, the fourth was constrained by those three.

Model	#	w	β	θ	γ	λ	w_1	w_2	w_3	w_4
CAM_u	4	-	-	-	-	-	1.79 (0.67)	0.34 (0.61)	-0.03 (0.5)	0.28 (0.6)
$jGCM_{11}^\gamma$	5	-	-	-	2.3 (0.96)	5.78 (2.26)	0.3 (0.17)	0.29 (0.15)	0.21 (0.14)	0.19 (0.09)
<i>dSSL</i>	2	5.45 (3.76)	0.83 (0.22)	-	-	-	-	-	-	-
<i>pSSL</i> $_\theta$	3	7.86 (3)	0.91 (0.11)	-	-	-	-	-	-	-
<i>dCSSL</i>	2	6.44 (3.42)	0.41 (0.21)	-	-	-	-	-	-	-
<i>pCSSL</i> $_\theta$	3	7.31 (3.12)	0.56 (0.28)	6.01 (14.58)	-	-	-	-	-	-

Note: CAM_u = Unconstrained cue abstraction model; $jGCM_{11}^\gamma$ = Generalized context model with joint representation, Minkowski and similarity parameters equal to 1 and free scaling parameter; *SSL* = Context-free strategy selection learning model, prefix *d* and *p* denote deterministic and proportional update, while the suffix θ denotes additional free scaling parameter; *CSSL* = Contextual strategy selection learning model, prefix *d* and *p* denote deterministic and proportional update, while the suffix θ denotes additional free scaling parameter; # = Number of parameters; w_{1-4} = Weight or attention parameters.

1.B Additional results

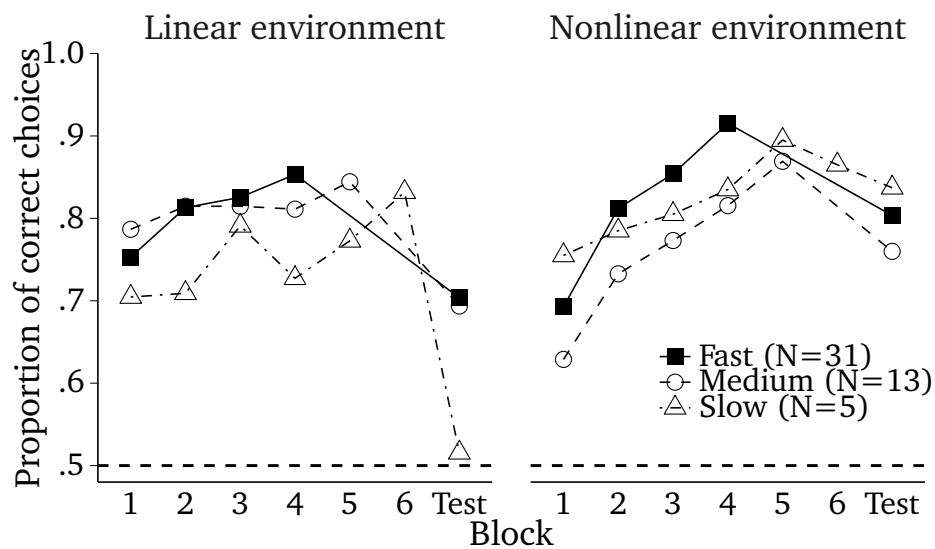


Figure 1.B.1 Choice accuracy in blocks in the training and the test phase. Participants that took additional one or two blocks in the training phase are illustrated separately, they are marked as slow or medium speed of reaching the accuracy level.

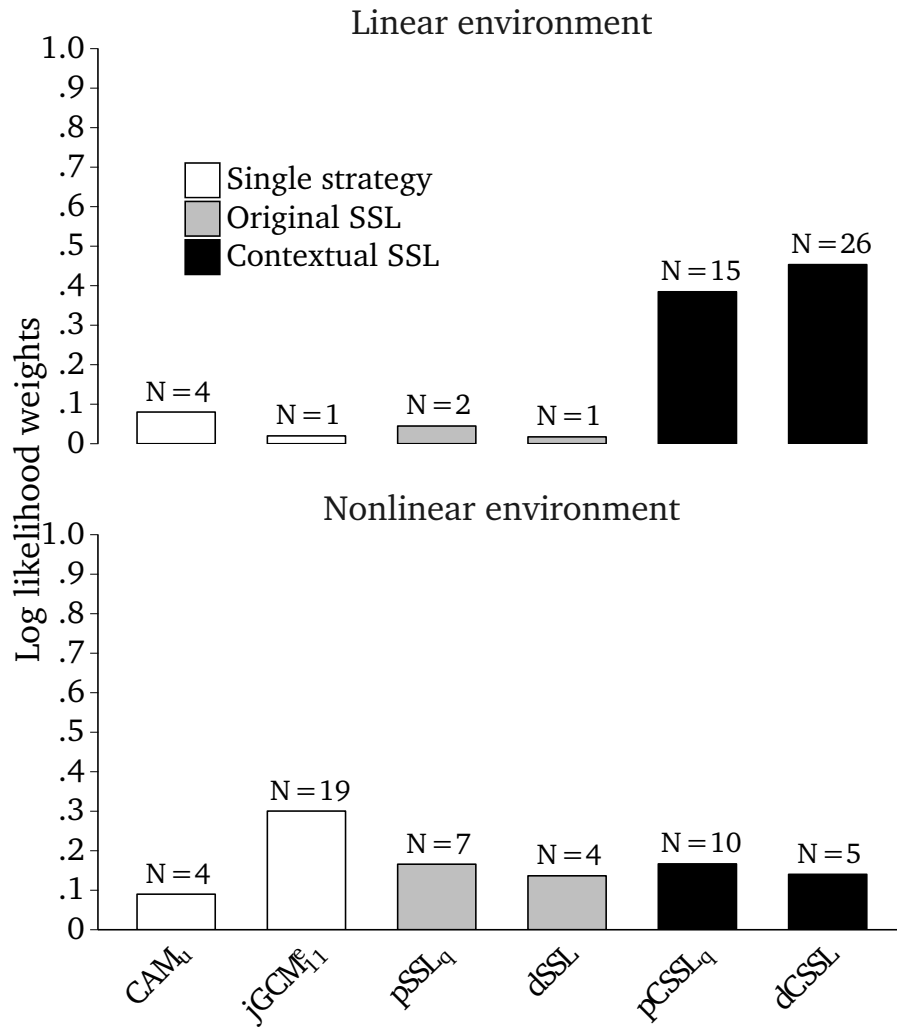


Figure 1.B.2 Model performance in predicting choices in the test phase for contextual strategy selection learning models (CSSL), original strategy selection learning models (SSL), and two single strategy models – CAM_u and $jGCM_{11}^Y$, computed separately for trials in the linear and nonlinear environment. Performance is expressed as mean log likelihood weight across participants, computed for these six models in the comparison set. Numbers above the bars represent number of participants whose choices in the test phase were best predicted by each of the models.

Stojić, H., Olsson, H., & Analytis, P. P. (2016). Explaining inter-individual variability in strategy selection: A cue weight learning approach. In D. Ritter & F. E. Ritter (Eds.), *Proceedings of the 14th International Conference on Cognitive Modeling* (pp. 144–150). Penn State. Retrieved from acs.ist.psu.edu/iccm2016/proceedings/stojic2016iccm.pdf

Chapter 2

Explaining inter-individual variability in strategy selection: A cue weight learning approach

Abstract

Do people integrate all the information at hand when they make choices or do they employ heuristics that ignore some of it? Recent research indicates that people's behavior should and does depend on the statistical properties of the environments within which cognition operates. However, in a single environment there are always decision makers who rely on less effective strategies. The source of this inter-individual variation has not been identified yet. In this article we postulate that it can be largely explained by differences in the speed of learning. We designed an experiment where participants first made choices between three multi-cue alternatives and received feedback about their quality. In a second stage, they predicted the quality of alternatives without receiving feedback. The quality was a linear combination of cue weights and cue values. To employ heuristics the participants had to learn at least weight directions and ranks, while for the integrative strategy they needed to learn the cue weights. We find that participants who showed evidence of learning cue weights rather than the ordering performed well in the estimation task that followed decisions, with cue weight knowledge being strongly related to decision performance. Further, we find that differences in how fast participants learn the cue weights explain the variability in regards to what strategy they adopted within an environment.

2.1 Introduction

Consider the following problem: you want to decide which hotel to book for your next vacation and you have access to information such as the facilities of the hotel, average reviews, cleanliness etc. To make an educated choice you could weight and add all the information at hand for each alternative and then choose the one that achieved the highest score. This is a weighted additive strategy (WADD; Payne et al., 1993). Alternatively, you could compare the hotels according to the most important cue and choose the one with the largest cue value. If some alternatives are tied on the first cue, you could move to the next cue in the ranking until you reach a decisive cue and stop your search. This corresponds to a heuristic strategy called take-the-best (TTB; Gigerenzer & Goldstein, 1996). On average take-the-best would ignore most of the information, as your decision would often be based on a single cue. Researchers have investigated theoretically the conditions under which it is well-advised to rely on integrative strategies such as WADD or heuristic strategies like TTB (e.g., Hogarth & Karelaia, 2005a, 2007; Martignon & Hoffrage, 2002). Empirically, however, there is a large inter-individual heterogeneity and substantial proportion of people still seem to use an inferior strategy (Bröder, 2003; Pachur & Olsson, 2012; Rieskamp & Otto, 2006).

Strategy performance primarily depends on the statistical properties of the relationship between cues and alternative quality. TTB fares well in comparison to WADD when the most informative cues are much more valuable than the less informative ones (Hogarth & Karelaia, 2007), or when the cue inter-correlations are high (Hogarth & Karelaia, 2005a). In environments with binary cue values, when the weights of the cues with higher weight rankings are larger or equal to the sum of weights of the cues with lower rankings, TTB cannot be outperformed by WADD. When this property does not hold, a WADD model with well-calibrated weights is expected to outperform TTB. The former environments are called non-compensatory and the latter compensatory (Martignon & Hoffrage, 2002).

Several experiments have demonstrated that over time most people converge to the best performing strategy. For example, people tend to adopt TTB in non-compensatory environments and WADD in compensatory environments (Bröder, 2003; Rieskamp & Otto, 2006). Similarly, in non-linear environments, when none of the aforementioned two strategies performs well, many people employ memory-based exemplar strategies (Pachur & Olsson, 2012). Further, people prefer heuristic strategies over integrative strategies when they are under time pressure or when the cost of learning cue values is high (Rieskamp & Hoffrage, 2008).

Within a single environment, however, there is always a substantial portion of participants that use inferior strategies. For example, in a non-compensatory environment there are always participants that continue using WADD, or TTB in the compensatory environment. The source of this inter-individual variation has not been identified yet, although it is widely reported (e.g., Brehmer, 1994; Bröder, 2003; Einhorn, 1970; Rieskamp & Otto, 2006). Bröder (2012) provides a summary of existing research on inter-individual differences in adoption of TTB and WADD strategies. The only variable that shows some correlation is the intelligence score. TTB users in the non-compensatory environment tend to score higher on an intelligence test than WADD users, although the effect is rather small. None of the personality measures, such as the “Big Five”, show a substantial correlation with strategy adoption. Similarly, motivational variables, cognitive styles, working memory capacity, and working memory load do not seem to influence adoption of TTB or WADD. Hence, the variation *within* an environment remains largely unexplained.

In this article we propose a solution to this puzzle. Strategies like TTB and WADD rely on cue weights. While in some experiments participants are given the cue validity weights directly (e.g., Rieskamp & Otto, 2006), in most of them participants have to learn the weights (e.g., Bergert & Nosofsky, 2007; Bröder, 2003). Hence, besides figuring out which strategy to use, they also need to learn the statistical properties that are input to the strategies. Importantly, strategies differ with respect to the amount of knowledge they require about the validity weights. While WADD requires exact quantitative estimates, TTB only requires the ranking and directions. Under reasonable theoretical assumptions, heuristic strategies like TTB are largely insensitive to the gap between estimated and objective validity weights, while performance of WADD is heavily affected (Hogarth & Karelaia, 2007; Katsikopoulos, Schooler, & Hertwig, 2010). As a result, in many environments people can leverage WADD’s improved performance only after some learning has occurred, and the estimated weights are relatively close to the objective ones. When coupled with usual individual differences in speed of learning, this explanation can address the observed variability in strategy selection. For example, in an environment favoring WADD, this leads to the prediction that slower learners will stick longer to the TTB heuristic, while faster learners will have more precise knowledge about the cue validity weights and will adopt WADD in greater numbers.

Our article suggests a novel approach in the study of decision making strategy by examining decision processes and cue weight learning in tandem. In our experiment, participants complete two tasks, a decision making and an estimation task. By adding an estimation task where par-

participants make predictions about values of alternatives we can model their cue weight learning and infer the evolution of their knowledge about cue weights. Thus, we can identify the role of cue weight learning in strategy selection and test the predictions made above.

2.2 Method¹

2.2.1 Participants

Seventy-eight participants (49 women, 29 men, $M_{age} = 21.8$, age range: 17–54 years), recruited from the Universitat Pompeu Fabra subject pool, took part in the study. They were paid a show-up fee of five euros and a performance dependent bonus of 6.8 euros on average. The experiment lasted 43 minutes on average.

2.2.2 Stimuli and procedure

The experiment consisted of two tasks: the participants first completed a decision making task and then an estimation task. In the decision task they repeatedly faced three alternatives, each described by the same four cues (Figure 2.1, left). The task was presented as a cheese game. Each alternative represented a cheese, the cues were “Lactic”, “Acetic”, “Casein” and “Texture”, while the alternative values represented enjoyment units (EU).

The criterion value, Y , of each alternative was a noisy linear combination of cue values and cue validity weights

$$Y = \sum_{i=1}^4 x_i v_i + e,$$

with weights v_i fixed at 4,-3,2 and -1. These cue validity weights strongly favor WADD over TTB. Cue values were sampled from uniform distribution $U(10, 90)$. A normally distributed error term, $e \sim N(0, 30)$, was added to each alternative. We created 480 unique alternatives in this manner and allocated them randomly across 160 trials, three alternatives per trial. Cue inter-correlations were zero on average. The stimuli were drawn only once and all participants received the same stimuli. The earnings were

¹The raw data is publicly available on Figshare: <http://dx.doi.org/10.6084/m9.figshare.1609680>.

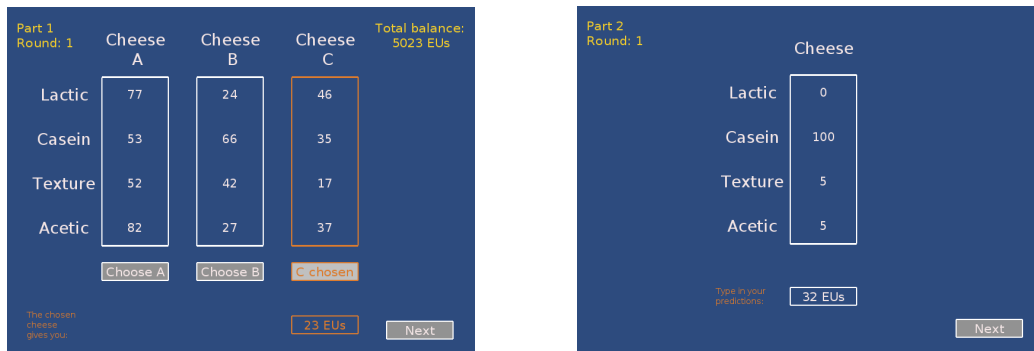


Figure 2.1 Screenshots of the tasks from the experiment. Left panel shows how the decision task appeared to the participants, while right panel shows the estimation task.

determined by the criterion value Y of the chosen alternative, which was also shown as feedback in each trial.

After every 40 trials in the decision task participants answered questions that probed their knowledge about the cue weights. Following Speekenbrink and Shanks (2010), we asked them to rate the strength of the relation between each cue and the value of the cheese on a scale from -10 (highly negative) to 10 (highly positive). Questions for all four cues were shown on the same screen, in the same order that was used to present the stimuli.

In the estimation task participants received a single alternative in each trial and their task was to predict the criterion value (Figure 2.1, right). No feedback was provided. We incentivized truthful reporting by computing the payoff as a function of a difference between the prediction P and the criterion value, $200 - |P - Y|$.

The stimuli for the estimation task were generated with the same cue validity weights as in the decision task. We generated 20 alternatives for interpolation trials by drawing cue values from the same range as in the decision task, $U(10, 90)$, and multiplying them with weights. We generated extrapolation trials in an analogous way by drawing cue values from two intervals at the extreme ends, $U(0, 10)$ and $U(90, 100)$ that have not been experienced during the decision task. After a single draw was made, trials were randomly ordered and all participants received the same set of stimuli.

In the decision task the participants were informed about the cues and the range of values they could take, and that they could use this information in making their choices. They were not told about the functional relationship between cue values and value of the cheese, nor that the weights differ for different cues. It was stressed that in each trial they would get

three new cheeses that differ in their cue values. The estimation task was announced at the beginning in the instructions, but without specifying details.

We told participants that it takes 60 minutes on average to complete the experiment. Each participant was presented with a unique random order of alternatives and cues. The four cue labels were also randomly attached to underlying cues separately for each participant.

2.3 Behavioral results

2.3.1 Choices in the decision task

Participants' performance, measured as percentage of correct choices per block, improved over time (Figure 2.2). Choice accuracy is much higher than the random level of 0.33 already in the first five trials (marked with number five in the figure), with 46% accuracy. People have a strong prior for positive linear relationships (Brehmer, 1994), which matches well the function that we used to construct the stimuli. Participants achieved a mean accuracy of 0.48 in the first block and by the end of the training phase they were close to choosing correctly the alternative with the highest criterion value two out of three times, 0.63. Although mean choice accuracy is similar to the accuracy achieved by TTB with ideal knowledge, 0.59 on average, the variance in individual choice accuracy curves is quite large. The shaded region around the mean performance indicates the range of accuracies, from 10th to 90th percentile. Hence, there are many individuals with accuracies far above what could be achieved with TTB.

Insight questions provide us with a first indication of how well participants have learned the cue validity weights. Previous research using such questions has shown that people have good insight into what they have learned (Speekenbrink & Shanks, 2010). Figure 2.3 shows mean ratings for all four cues. Participants got the relative ordering and directions right on average already after 40 trials and it got clearer as the training progressed. They learned that the second cue has a larger weight (although negative) than the third cue only at the end, and failed to detect that the fourth cue had a small negative weight. This is not surprising as negative linear relationships are more difficult to learn than positive linear ones (Brehmer, 1994). Although insight questions use an arbitrary scale and it is difficult to identify exact cue weights that participants have acquired, they do suggest that people learn more than ordering and directions. This is supported by changes in ratings over the course of the decision task, even

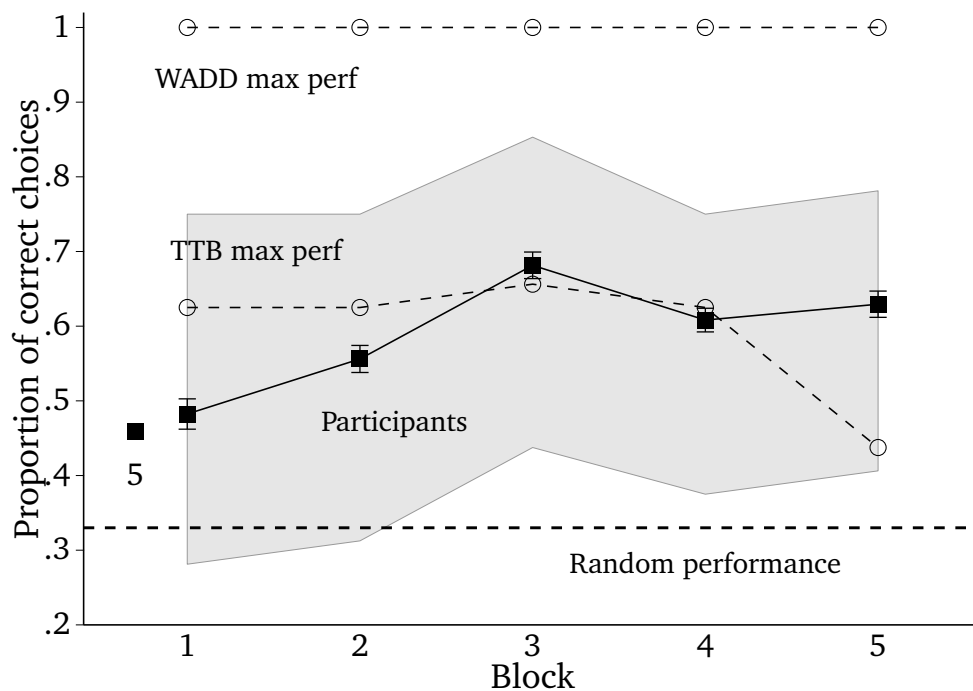


Figure 2.2 There is a clear learning effect in the decision making task. Figure shows increase in mean accuracy of participants' choices over blocks of trials. In addition, we displayed mean accuracy in the first five trials, marked with number five. Result for each block is a mean of individual means across 32 trials, and error bars represent standard errors of group means in each block. Shaded region around the curve indicating the mean accuracy is the range from 10th to 90th percentile of accuracy in each block. We also illustrated the performance of TTB and WADD model with perfect knowledge about the environment structure.

though the ordering and directions were mostly established already after first time participants answered the insight questions.

2.3.2 Predictions in the estimation task

We can also assess knowledge about cue validity weights by examining the performance in the estimation task. We computed mean absolute deviation (MAD) and correlation between participants' predictions and criterion values as a measure of performance. Results are shown in Figure 2.4B. Mean MAD across participants is 120 (SD = 30.8), which means that on average predictions were 120 EU's away from criterion values. Mean (median) Spearman correlation is 0.63 (0.70; SD = 0.24). The participants are do-

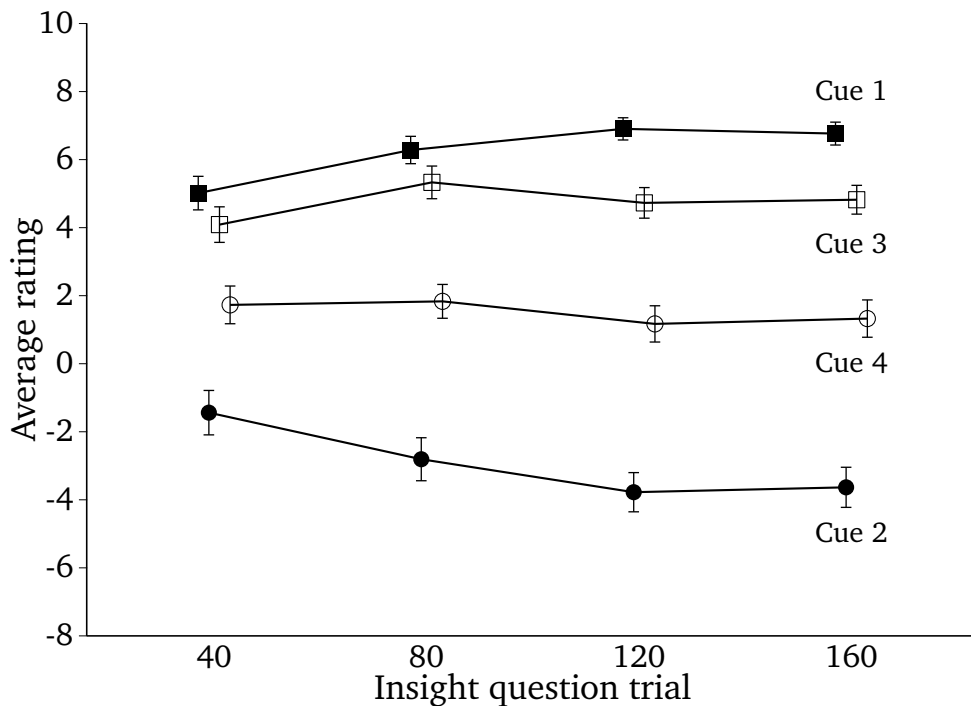


Figure 2.3 Participants have insight into their own learning of cue validity weights – they learned the ranking of cues and cue directions to a large extent. Points at trials 40, 80, 120 and 160 in the decision task are mean ratings. To allow for easier discrimination, ratings have been slightly displaced horizontally and connected by lines. Error bars represent standard errors of means across participants.

ing a good job in predicting criterion values of test items, but as expected, inter-individual variation in learning is substantial, with MAD ranging from 51 to 189. While most people are doing quite well, having very high correlations and low MAD's, some people do very poorly.

How would a decision maker that only learned the ranking of cues fare in the estimation task? Such a decision maker could take a mean of the criterion values experienced in the decision task and use it as a fixed prediction for all items in the estimation task. This is our baseline prediction performance. The MAD between baseline predictions and criterion values was 172, much larger than for observed MAD.

We get more complete insight by examining mean predictions across participants for each of the 40 items in the estimation task. Figure 2.4B shows that in the range of item values from about zero to 200, mean predictions correspond very closely to the criterion values. More deviations occur for more extreme values, with somewhat poorer predictions for ex-

trapolation items than interpolation items. Importantly, predictions correspond much better to criterion values than baseline predictions. Thus, most participants do acquire more precise knowledge about cue validity weights, rather than only the ordering and directions.

2.3.3 Relation between decisions and predictions

We examine the relationship between individual performances in the two tasks to obtain model-free evidence that cue weight learning plays an important role in strategy selection. We find a strong relationship between choice accuracy in the decision task and MAD in the estimation task, as indicated by a Spearman correlation of -0.78 (Figure 2.5). This suggests that participants with good prediction performance know the cue weights well, which allowed them to employ WADD and achieve good decision performance. Surprisingly, many participants who had poor prediction performance also had decision performance far below 0.59 which is possible to achieve with very little knowledge for TTb in this environment. They either relied on WADD in spite of their poor knowledge or those participants simply paid less attention and performed close to random in both tasks.

2.4 Modeling

Next we turn to identifying the strategies used by each participant in the decision task. We first describe the cue weight learning model that will produce trial-by-trial predictions of participants' knowledge of cue weights. These weights will in turn be used in fitting TTb and WADD models to participants choice data. Finally, we will examine whether participants that were best fitted by TTb have less developed knowledge of cue weights than those best fitted by WADD, as predicted.

2.4.1 Modeling the cue weight learning

We used a least mean squares model to model the cue weight learning process (Gluck & Bower, 1988). The LMS model predicts the criterion value of an alternative on trial t as

$$P_t = \sum_{i=1}^4 x_{i,t} u_{i,t},$$

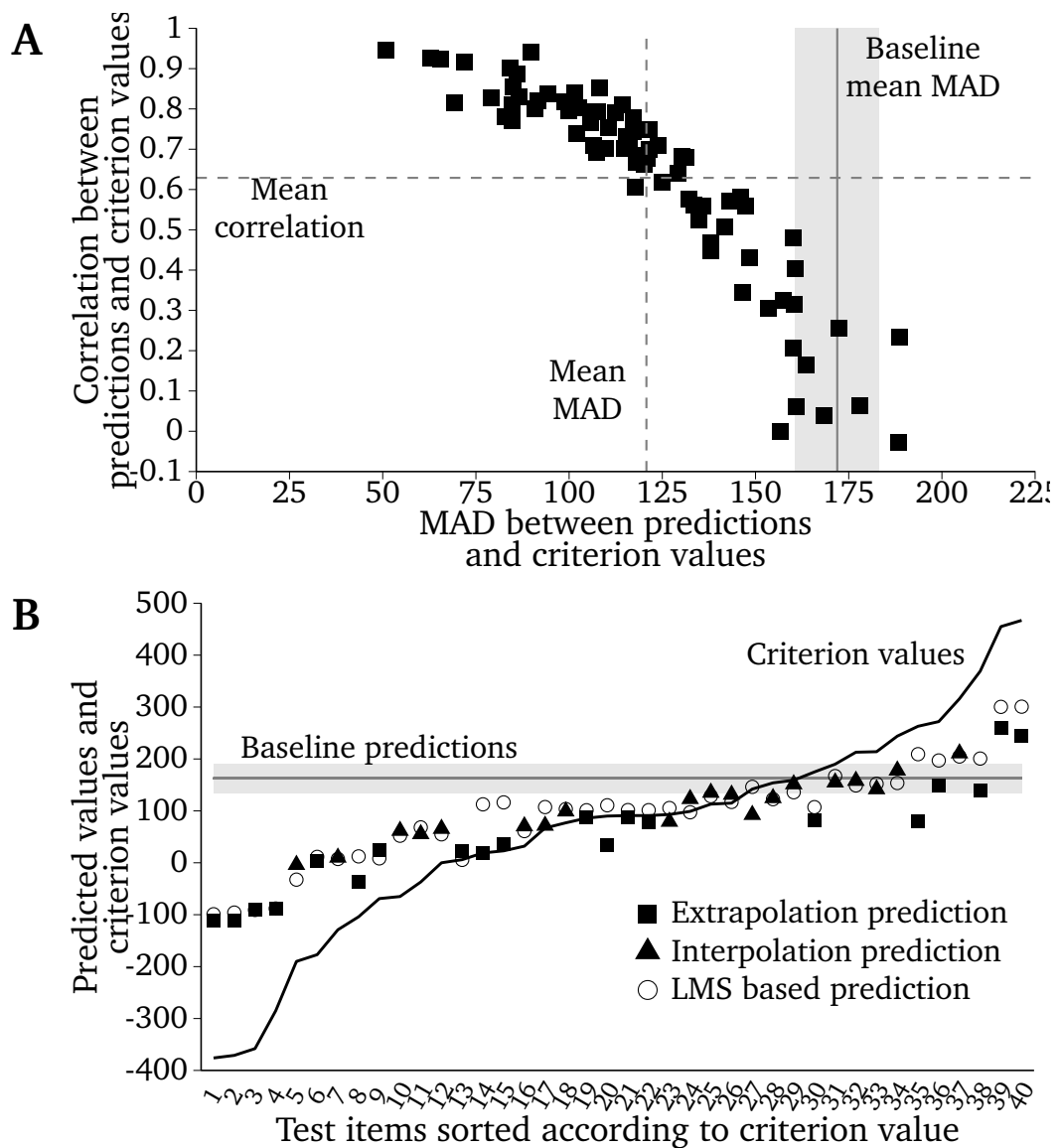


Figure 2.4 Participants make good predictions in the estimation task. (A) Prediction performance for each participant on two measures – mean absolute deviation (MAD) and correlation between predictions and criterion values. A single point is a result for one individual, while dashed lines are means across participants. Vertical line is the mean absolute deviation between baseline predictions (see text) and criterion values. (B) Mean predictions for each of the 40 items (20 extrapolation and 20 interpolation) in the estimation task. Diagonal black line represents the criterion value of the items. The farther the predictions are from this line the worse the predictions. Gray horizontal line is the baseline prediction – mean value of the items experienced in the decision task. We also denoted mean predictions based on the Least mean squares network model (LMS_d) fitted to each participant.

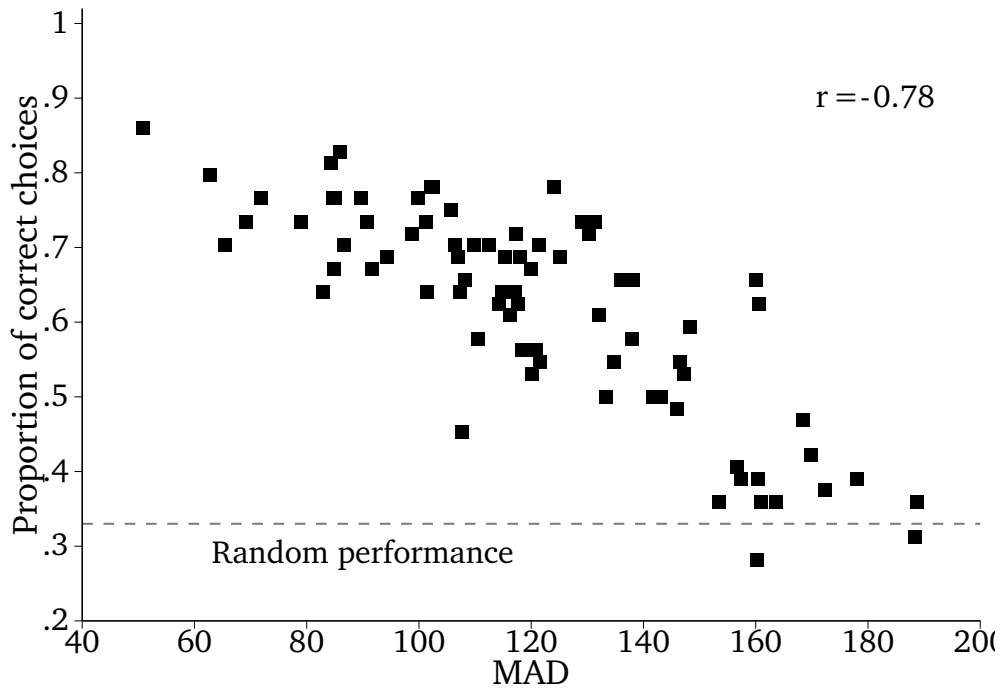


Figure 2.5 Performance in the estimation task and the decision task is highly correlated. Performance in the estimation task is expressed in terms of mean absolute deviation (MAD) between predictions and criterion values, where lower numbers indicate better performance. For performance in the decision task we used proportion of correct choices in the last two blocks. In the upper right corner we have shown Spearman correlation between these two variables.

where $u_{i,t}$ are cue utilization weights and $x_{i,t}$ are cue values of cue i in each trial t . Utilization weights are updated in every trial through the delta rule, based on a prediction error defined as the difference between the predicted criterion value, P_t and the true criterion value, Y_t , that a participant receives as a feedback in the decision task

$$u_{i,t+1} = u_{i,t} + \frac{\eta}{t^\gamma} (Y_t - P_t) x_{i,t},$$

where $0 \leq \eta \leq 1$ is a learning rate parameter shared by all four cues and $\gamma \geq 0$ is a decay parameter. We initialized the weights to $u_{i0} = 0$, $i = 1, \dots, N$. Note that the cue weight learning process is based only on the alternative for which participants receive feedback, the rest is ignored by the LMS model.

We fitted two different versions of LMS model. LMS_d where both η and γ are free parameters and LMS where γ is set to 0. Parameters were

initialized at the beginning of the decision task and in each trial cue values and criterion of the chosen alternative were used to update the weights. The weights from the last trial were used to make model based predictions in the estimation task. To estimate the model parameters we minimized the mean squared error between the participant's and model's predictions. The LMS model was fitted separately from the choice models.

2.4.2 Modeling the choices

Random Choice Model

We used a random choice model (RCM) as a baseline. RCM predicts the same probability, .33, for each alternative.

WADD Model

Our version of WADD linearly combines the cue utilization weights learned by the LMS model with cue values to produce predicted value of each alternative k in trial t

$$R_t^k = \sum_{i=1}^4 x_{i,t}^k u_{i,t},$$

where $u_{i,t}$ are cue utilization weights learned by the LMS model based on trials $1 : t - 1$. WADD then deterministically decides by maximizing among the alternatives. To fit WADD to data we assume an additional "tremble" error. If a strategy produces a probability that alternative k is chosen, $P(C = k)$, then the probability of choosing k after taking into account the tremble error, ϵ , is given by

$$P(C_t = k; \epsilon) = (1 - \epsilon) \times P(C_t = k) + \frac{\epsilon}{3}$$

TTB Model

Our version of TTB uses the cue weight information from the LMS model, $u_{i,t}$, to order the absolute value of the weights from the largest weight to the lowest, producing a ranking r_t . The ranking is done on absolute values because a strong negative weight is as predictive as a strong positive weight. TTB then chooses an alternative with the largest cue value of the most predictive cue according to ranking r_t . If values of the first cue according to the ranking are the same for all alternatives², TTB inspects the

²Ties are rare in environments with continuous cue values, making this version of TTB quasi-equivalent to a single-variable strategy, which uses only the most important cue.

Table 2.1 Mean Bayesian Information Criterion (BIC) scores of models (standard deviation in the parenthesis), number of participants best fitted the model and mean parameter values.

Model	#	BIC	N	η	γ	ϵ
<i>LMS</i>	1	368 (25)	39	2e-5	-	-
<i>LMS_d</i>	2	366 (24)	39	2e-4	.61	-
<i>WADD</i>	1	283 (40)	56	-	-	.62
<i>TTB</i>	1	302 (40)	11	-	-	.74
<i>RCM</i>	0	355 (2)	11	-	-	-

Note. # = Number of parameters in the model; N = number of participants best fitted by the model; *LMS* and *LMS_d* = Least mean squares network model, with constant learning rate and decaying learning rate respectively; *WADD* = Weighted additive model; *TTB* = Take-the-best model; *RCM* = Random choice model; η = learning rate in the LMS model; γ = decay rate in the LMS model; ϵ = tremble error.

second cue and so on, until it finds a cue that discriminates between the alternatives. If no cue discriminates, a choice is made at random. If the deciding cue had a negative weight according to the u_t , cue values of all three alternatives were multiplied with -1 , to maintain the correctness of the rule of choosing the alternative with larger cue value. Same as in the WADD model, we add a “tremble” error term to arrive at the final choice probability, $P(C_t = k; \epsilon)$.

2.5 Modeling results

Table 2.1 shows the mean Bayesian Information Criterion (BIC) score across participants for LMS models and choice models. Both *LMS* and *LMS_d* fit the predictions equally well, both in terms of mean BIC (368 and 366) and number of participants best fitted (39 for both). However, *LMS_d* fits results better in a qualitative sense. It emulates better the insight questions results where most people acquire ordering and directions very fast. Hence, we used weights from *LMS_d* in the choice models. Moreover, *LMS_d* based predictions for estimation task items correspond closely to participants’ predictions (Figure 2.4B).

In terms of choice models, as expected, WADD has better mean BIC score (283) than TTB (302). Similarly, most participants were best fitted by WADD (56), followed by TTB (11) and RCM (11). As has been widely observed in previous studies, although it pays better to adopt WADD, and

indeed most people do so, there is substantial inter-individual variability. There are substantial differences between the three groups. As expected, WADD users reached the highest accuracy, they were choosing the best alternative on average in 0.63 proportion of trials. TTB users performed worse, having a choice accuracy of 0.55. Although RCM users were the worst, reaching mean accuracy of 0.42, their performance is somewhat higher than the random level and they do exhibit some learning by the end of the training phase.

Next we examine our prediction that participants best fitted with TTB are those that learn slower and did not manage to arrive at sufficiently good utilization weights to switch to WADD. We plot the evolution of utilization weights estimated with the LMS_d model, separately for participants best fitted with each model (Figure 2.6). We see that WADD users have a well developed knowledge of all four cues, while TTB users have less developed knowledge. Notably, TTB users have very good estimates for the most important cue and do not distinguish that well between the other three cues. Their adoption of the TTB strategy is well justified by their subjective knowledge of the cue weights. RCM users' knowledge is very poor, capturing unmotivated or inattentive participants.

We can also examine estimated learning rate parameters of the LMS_d model. Learning rates are higher for WADD users than TTB users, and lowest for RCM users (Figure 2.7). Median learning rate for WADD users was 0.00015, while for TTB users it was lower for an order of magnitude, 0.000027. Median decay rates are correspondingly higher for the WADD users, 0.69, than for the TTB users, 0.54. Performance of TTB users in the estimation task ($M_{MAD} = 133$) was expectedly worse than that of WADD users ($M_{MAD} = 112$), but importantly, substantially better than of RCM users ($M_{MAD} = 155$) or baseline ($M_{MAD} = 172$). Similar differences can be seen in the insight questions results, with knowledge of TTB users evolving over time. This suggest that even a TTB user learns more than just the ordering and the direction of cues.

Finally, we conducted a logistic regression with mean absolute difference between LMS obtained utilization weights in the last block and objective weight as a predictor of strategy use. We obtained a negative coefficient, as predicted, at a value of -1.466 (95% $CI[-2.743, -0.367]$), $p = 0.0139$ (WADD users were coded as 1 and TTB users as 0, while RCM users were not included). In odds ratio terms, for one unit increase in mean difference, the odds of using WADD decrease by 76%. Odds of using WADD for the perfect knowledge (zero difference) is very high, 50.85, which amounts to a probability of 0.975. Although this outcome was already suggested by behavioral results illustrated in Figure 2.5, this analysis

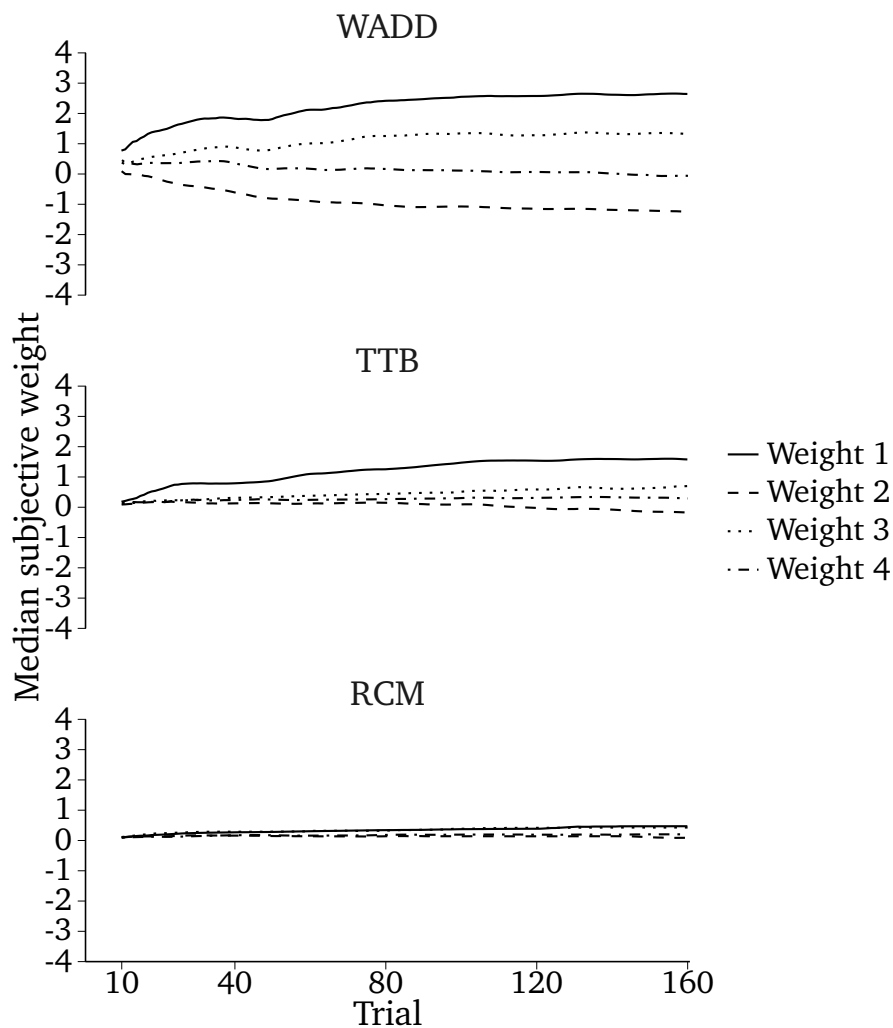


Figure 2.6 Participants best fitted by the weighted additive (WADD) model are learning cue validity weights much faster than participants best fitted by the take-the-best (TTB) model or the random choice model (RCM). In the figure we show evolution of cue utilization weights in the decision task according to the least mean squares network model (LMS_d) fitted to each participant. Weights are median cue weight across participants for each trial, smoothed with a moving average of ten trials. The objective weights used to construct the stimuli in the task were: $v_1 = 4$, $v_2 = -3$, $v_3 = 2$ and $v_4 = -1$.

establishes the link between the knowledge of cue weights and strategy selection more clearly, in a model based manner. Since WADD users achieve greater decision performance, it explains the large correlation between estimation and decision performance seen in Figure 2.5.

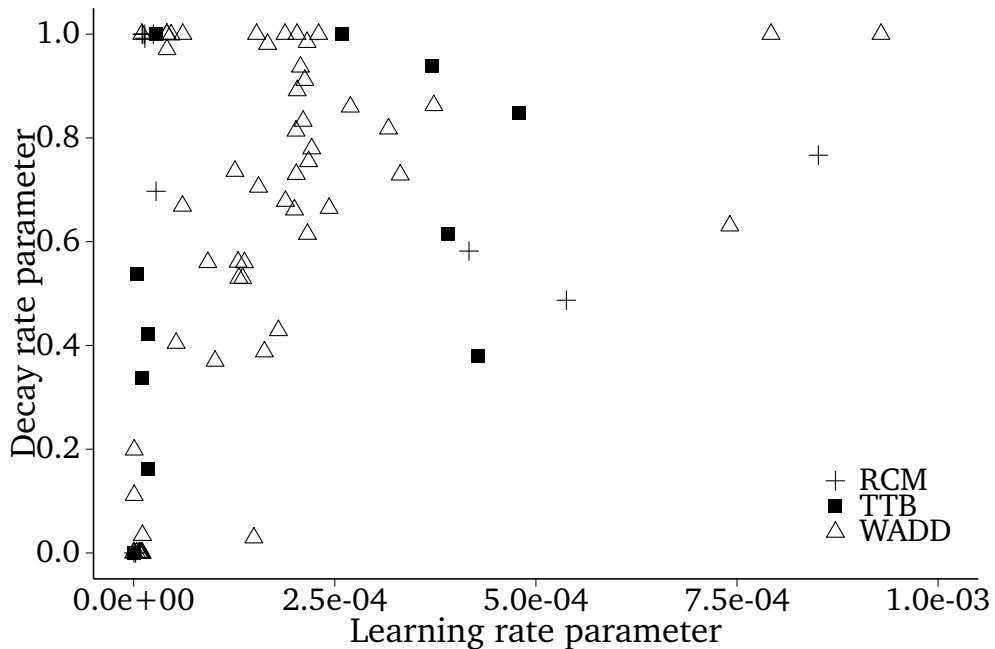


Figure 2.7 Estimated learning η and decay rate γ parameters for the Least mean squares network model with decaying learning rates (LMS_d). Learning and decay rates for the weighted additive (WADD) model are on average higher than for the participants best fitted by the take-the-best (TTB) model.

2.6 Discussion & Conclusion

In our experiment participants differed in how fast they acquired knowledge of cue weights, and we predicted this heterogeneity to be responsible for the variability in strategy selection. Our results showed support for our predictions – WADD users had better developed knowledge of cue weights than TTB users and the performance in the estimation task is consistent with the strategy adoption. Our learning rate account suggests that, given time, TTB users would learn the weights sufficiently well and switch to the better performing WADD strategy.

Where do the inter-individual differences in learning rates come from in the first place? These differences might be akin to traits like intelligence or personality factors investigated by Bröder (2012). This would require the learning rates to be stable across time and tasks within people. To our knowledge, there is no study that examines the stability of learning rates and is difficult to generalize beyond our task.

In our study we set out to test a specific hypothesis and to inform the debate on whether people are better described by the WADD or TTB model. We have to note that the models do not perform particularly well in our task. This can be witnessed in the high values of the ϵ parameter in Table 2.1, meaning that models on average predict the choices of the participants half of the time. Given our modest goals we did not try to look for models that would explain behavior even better. Our results, however, indicate that we should look for such models within the probabilistic rather than deterministic class of models (Bergert & Nosofsky, 2007).

Our results could be also explained if some participants first adopted TTB and as a consequence learned cue weights differently. With our current experimental design we cannot, unfortunately, determine the direction of the causal arrow. However, our evidence indicates that TTB users acquire more than ordinal information about cue weights and that this knowledge becomes more precise over time. This suggests that, if such interdependence exists, at most it slows down the learning. This evidence comes from three sources – the insight questions, the estimation task and the joint modeling of cue weight learning and decision making. The continuous evolution of our participants' knowledge of cue weights goes against the frugality and robustness justifications of TTB. The argument against using cue weights hinges on their vulnerability to overfitting – relying on ordinal information instead leads to better generalization. From our perspective, TTB and other heuristic strategies are used either due to cognitive limitations or when the structure of the environment is known better and these strategies are the rational thing to do (also see Davis-Stober, 2011; Davis-Stober, Dana, & Budescu, 2010).

In this decision-making paradigm, our evidence suggests that learning the properties of the environment is predominant, and strategy selection is influenced by it. Different decision making tasks, however, may lead to distinct linkages between cue weight learning and decision making processes. Exploring the nature of these interactions opens an exciting direction for future research (see Stojić, Olsson, & Speekenbrink, 2016; Stojić, Analytis, & Speekenbrink, 2015).

Stojić, H., Analytis, P. P., Dayan, P., & Speekenbrink, M. (n.d.). Trials-with-fewer-errors: Feature-based learning and exploration. (Unpublished working paper).

Stojić, H., Analytis, P. P., & Speekenbrink, M. (2015). Human behavior in contextual multi-armed bandit problems. In D. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. Jennings, & P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2290–2295). Austin, TX, US: Cognitive Science Society. Retrieved from mindmodeling.org/cogsci2015/papers/0394/paper0394.pdf.

Chapter 3

Trials-with-fewer-errors: Feature-based learning and exploration

Abstract

Reinforcement learning algorithms have provided much insight into human and animal learning and decision making. However, the traditional algorithms perform poorly when faced with real world situations characterized by multi-featured alternatives and contextual cues. In this paper, we propose a Bayesian optimization framework for tackling such decision making problems. The framework relies on similarity-based learning of functional relationships between features and rewards, and choice rules that use uncertainty to balance exploration and exploitation. To test this new approach, we designed a series of novel multi-armed bandit experiments where alternative rewards are noisy functions of two features. We evaluate human behavior in these problems and compare it to solutions prescribed by the Bayesian models. We find that people's exploration is guided by both their prior expectations and learning about the function in the task. However, there are notable inter-individual differences and a sizeable proportion of participants ignores the feature information and relies on mean rewards only. More importantly, we show that people's exploration patterns show clear signs of Bayesian optimization – simultaneous function learning and function maximization. Even though most people do not perform as well as the models, they do allocate a portion of choices to learning the function and take uncertainty about their knowledge into account when choosing. However, paying attention to context can impede performance if prior expectations about reward functions do not correspond to the actual function encountered in the environment. We illustrate the fertility of the paradigm and highlight several exciting lines of future research.

3.1 Introduction

Real-world decision situations are characterized by informative features – we check the number of patrons in a restaurant and ratings at online websites, while job candidates display their resumes filled with credentials. By making decisions, agents not only reap rewards or avoid punishments, they collect observations from which they learn the relations between features and outcomes. The arrow points to other direction as well – when weighing between alternatives, organisms can draw on their knowledge about relations and predict the value of alternatives, whether restaurants or job-market candidates.

Cognitive scientists have followed a divide-and-conquer strategy in studying this problem, isolating the process of learning the functional relationships, and the process of learning about rewarding courses of action. However, to understand decision making in the wild, we should not ignore the evident interactions between function learning and decision making. We propose a theoretical framework to capture these interactions and evaluate it in a series of experiments with human subjects.

The first process – learning to choose rewarding alternatives – has been studied in detail under the umbrella of reinforcement learning (RL). RL provides a normative framework that describes how agents learn to predict and acquire rewards through trial-and-error. Originating in early studies on the “Law of Effect” and conditioning (Rescorla & Wagner, 1972; Thorstone, 1927), modern RL is one of the success stories of psychology and neuroscience. Rescorla-Wagner and Temporal Difference learning models (Rescorla & Wagner, 1972; Sutton & Barto, 1998) have fostered a major breakthrough in understanding how animals and humans learn to make choices over time to reap rewards and avoid punishments. These models provide good descriptions both at the behavioral (e.g. Barron & Erev, 2003; Denrell, 2007; Denrell & Le Mens, 2007; Erev & Barron, 2005) and the neural levels – for example, they predict patterns of dopaminergic activity (e.g. Houk, Adams, & Barto, 1995; Niv, 2009; Schultz et al., 1997). Moreover, this research explicitly takes into account the fact that the decision maker receives feedback only on chosen alternatives, formalizing it in an elegant way as the exploration-exploitation trade-off.¹

Despite these impressive advances it is still unclear how to make the in-

¹We can choose alternatives that maximize the rewards according to our experiences so far (exploiting) or we can choose new ones that we know little about to learn whether they are better or worse than the known alternatives (exploring). Performing well requires a fine balance between exploration and exploitation, i.e. between acquiring information and reaping rewards (Sutton & Barto, 1998; Tversky & Edwards, 1966).

sights of RL research relevant for the kind of real world problems described at the start of this introduction. The vast majority of RL studies have used toy problems with only a few alternatives, where people can assess the value of actions only from sequences of rewards. In real-world situations with numerous features and contexts, the same situation might never occur twice. RL models fare poorly when faced with such scenarios (Dayan & Niv, 2008; Gershman & Daw, 2017; Gershman & Niv, 2010; Gershman, Pesaran, & Daw, 2009), they need thousands of trials to learn while animals or humans need only few. These models fail to exploit the structure of the task – the relations between features and rewards. Many such features (e.g., a restaurant’s popularity or TripAdvisor rating) are useful predictors of rewards (e.g., the enjoyment of eating a meal at a restaurant). Imbuing the RL models with the ability to explicitly learn the function relating features to rewards would allow them to learn in fewer trials, making fewer errors, and tackle real-world problems more efficiently.

Research on the second process, learning the relations between some observable features or cues and an unobserved outcome or criterion, has a long history in cognitive psychology. In research on multiple-cue learning, originating in early work by Egon Brunswik and his lens model of human judgments (Hammond & Stewart, 2001), researchers relied on a linear framework that defines both judgments and criterion being judged as functions of cues in the environment (e.g., Hammond, Hursch, and Todd 1964 and Hammond 1955, but see for heuristic approach Todd and Gigerenzer 2000 and work that connects it to the lens model, Hogarth and Karelaia 2005b). The criterion being judged was usually continuous and judgment accuracy largely depends on how well people learn the function. In contrast, category learning focused on studying how people learn to map cues to discrete criterion – categories and concepts. In doing so, researchers developed a new class of nonparametric models – exemplar-based models (Medin & Schaffer, 1978; Nosofsky, 1984, 1986), that successfully explain a wide range of empirical phenomena. More recently, researchers brought together these two streams by modeling judgments of continuous criterion with exemplar-based and connectionist models (Busemeyer et al., 1997; DeLosh, Busemeyer, & McDaniel, 1997; Juslin, Olsson, & Olsson, 2003; Kalish, Lewandowsky, & Kruschke, 2004; McDaniel & Busemeyer, 2005; Speekenbrink & Shanks, 2010), coining this field as function learning (but see also early work by Estes, 1960, 1976).

Even though the pioneers of the cognitive revolution already noted that we are actively influencing our observations and concluded it surely plays an important role in category and concept formation (Bruner, Goodnow, & Austin, 1956), studies in these lines of research tend to treat learners as

a passive observers, without any control over their learning environment. There are few notable exceptions. Einhorn and Hogarth (1978) noted that decision makers receive feedback only on chosen courses of actions, and attributed the overconfidence bias to the skewed sample of collected experiences this entails. More recently, Markant and Gureckis (2014) studied the benefits of actively selecting observations for category learning and found that active learners are learning faster (see also Kruschke, 2008; Nelson, 2005; Nelson, McKenzie, Cottrell, & Sejnowski, 2010). People also make decisions to learn the causal structure of the world (Bramley, Lagnado, & Speekenbrink, 2015; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003), again affecting the efficiency of learning. Traditional function and category learning research might thus provide an incomplete picture of how agents acquire real world knowledge of concepts that is unaffected by reward landscapes that skew observations.

3.2 Goals and Scope

In the current work we argue that to make further advances in understanding decision making processes in the wild, we need to examine both function learning and reinforcement learning processes jointly. Our goal is twofold: (1) provide a theoretical framework which can be used to understand how these two processes interact provide functional knowledge and enhance choice performance; and (2) examine people’s behavior in tasks where people can engage in both function learning and decision making, comparing them with theoretical prescriptions. The scope of the framework concerns situations where decision makers repeatedly face a choice between a number of uncertain alternatives (e.g., choosing between restaurants) with the goal of maximizing the total accumulated reward. We focus our attention on situations with alternative specific features that can inform expectations about the value of an alternative (e.g., number of patrons) and to situations where one immediately obtains feedback.

We use a Bayesian optimization framework where we rely on a Bayesian nonparametric approach to model function learning. This similarity-based model gives us not only an estimate of a function (i.e. expected reward for each alternative in a choice set), but also the uncertainty about this estimate. This allows us to model the decision making process using choice rules where both the expected reward and informational value of each alternative can be taken into account. We contrast this model with a Bayesian model that ignores the feature information altogether and learns about mean rewards of the alternatives in an optimal manner.

There are several key predictions stemming from the Bayesian optimization framework. The main one is that the choice allocations of agents that simultaneously learn the function and make decisions will be systematically biased – they will steer away from exploring alternatives for which the learned function predicts low rewards and allocate them to high rewarding regions instead. As soon as you learn that restaurants where you see no patrons are unlikely to result in a rewarding meal you will generalize and rarely visit such restaurants. In contrast, learning strategies that ignore the features such as mean-tracking would explore the choice set more uniformly. You would be more likely to visit the restaurant without patrons if you have not paid attention to feature information. Other predictions include, for example, poorer knowledge of the low rewarding regions of the function, and prior beliefs about the likely function sometimes leading decision makers to lock themselves into a certain region and forming incorrect beliefs about the function.

We evaluated these predictions in three sets of experiments with over 1000 participants. Participants completed a novel feature-based multi-armed bandit (FMAB) task where rewards were a noisy function of two continuous feature values as well as a generalization task that verified their functional knowledge. We compared their exploration behavior to predictions derived from our Bayesian models, as well as to a control group of participants that completed exactly the same task but without access to the feature information (i.e. a standard multi-armed bandit task). The experiments differed in terms of the function governing the rewards, the time horizon, and the nature of the experienced uncertainty.

This article is organized as follows. We begin by formalizing the decision situation we are interested in as a FMAB problem and defining the scope in more detail. Next we describe the Bayesian models that tackle the FMAB problem either through learning the function or tracking the mean rewards. Then we report results of the experiments examining the human behavior in FMAB problems and compare this to predictions from the Bayesian models. Finally, we discuss future directions and applications.

3.3 Feature-based Multi-Armed Bandit Task

The task faced by agents that learn functions simultaneously with deciding on the course of action that will bring them maximum amount of rewards can be neatly formalized within the contextual multi-armed bandit (CMAB) framework (e.g. Auer, 2002; Langford & Zhang, 2008; Li, Chu, Langford, & Schapire, 2010). This framework can capture special cases in which

the outcomes of different alternatives are influenced by a shared context (SMAB), cases where the alternatives share common features with specific feature values (FMAB) or a combination of both (CMAB).²

In the present work we will restrict our attention to the version where the alternatives share common features (FMAB). This version of the problem has received less theoretical attention so far, yet it relates to almost all real-world decision making problems. In each trial $t = 1, \dots, T$ an agent faces a choice between K alternatives, where each alternative k has J continuous valued features, described with a vector $\mathbf{x}_k = (x_{1,k}, \dots, x_{J,k})$. Figure 3.1 shows how the task appeared to our participants in the experiments. They observed 20 alternatives, each characterized by two features – length of horizontal and vertical line. In our restaurant example these features could correspond to restaurant popularity and cleanliness, for instance. Choosing alternative k on trial t will yield a reward R_k^t . Notably, the reward is only revealed for the chosen alternatives. The agent’s task is to maximize the sum of rewards accumulated over time, $\sum_{t=1}^T R^t$. The rewards are derived from a noisy function of its feature values

$$R_k^t = f(\mathbf{x}_k) + \epsilon_k^t,$$

where ϵ_k^t is the alternative and trial specific error term. The function $f(\cdot)$ holds for every alternative, but is initially unknown to the agent, who can only learn about the function by choosing alternatives and observing the corresponding rewards. In our task agents face the same set of K alternatives in each trial t , however, choosing the same alternative is unlikely to give the same reward due to the error term. Even with perfect knowledge of the function, reward of an alternative cannot be completely predicted.³

We postpone the formal treatment of how one should learn and make decisions in this task and first provide some intuition about two qualitatively different approaches. One strategy available to the agent is to learn the initially unknown function f from rewards she receives after choosing an alternative with particular features. Then she could use the learned function to predict the rewards of available alternatives and choose the one that is most promising (Auer, 2002; Shahriari et al., 2016; Srinivas, Krause,

²Most of the literature comes from machine learning. There are other terms in usage – associative bandits, bandits with side information or coavriates, bandits with expert advice or bandits with similarity information. Sometimes there are subtle, but important differences between them.

³Stochasticity makes the task akin to gambling, which is where the term “multi-armed bandit” comes from. One can imagine a set of K slot machines in a casino. One would try to discover the slot machine that has the highest probability of resulting in a win and pull its arm as many times as possible to maximize earnings.

Kakade, & Seeger, 2009). Facing the problem of choosing a restaurant for dinner, one would examine the features of restaurants in the area and use the functional knowledge to choose the one with the best predicted dining outcome. In the restricted version we study here, where the features of the alternatives are constant over time, the FMAB task is in fact identical to a standard stationary MAB task for agents who ignore the feature information. This gives a second, qualitatively different, strategy agents could use to perform well in the task. They could use a trial-and-error strategy and track rewards associated to alternatives only, R_k , learning over time which alternative has the highest expected reward, $E[R_k]$ (Steyvers, Lee, & Wagenmakers, 2009; Sutton & Barto, 1998). If you would visit all the restaurants in the neighborhood enough times, you would not need to inspect the features and instead can solely rely on your acquired restaurant expertise.

A key missing ingredient is the decision strategy – having estimated the rewards of available alternatives, how should an agent choose which alternative to sample next? Choosing only the alternatives the agent deems to be the best at the moment is not a wise strategy. There might be a better alternative out there and the agent should take into account how reliable her knowledge is – she should carefully balance between exploiting her current knowledge, reaping the rewards, and exploring the alternatives in the lookout for better ones. Importantly, in the FMAB task this exploration-exploitation trade-off has a different flavor for the learner that pays attention to feature information – exploring now means learning about the function, not only about mean reward of a particular alternative.

Allowing for two qualitatively different ways to tackle the task was purposeful. Where the function learning approach really excels compared to the mean reward strategy is when the number of alternatives becomes too large to try them all, when choice sets change and novel alternatives enter often, or when exploring is very costly. In other words, whenever the benefit of generalization is substantial one should use the function learning approach instead of tracking the means. We believe the benefits of generalizing knowledge to new situations are large enough to make the function learning approach a default one. Still, some real-world decision situations might be easily dealt with by learning the mean rewards only, so decision makers might be better off not engaging into costly function learning. Hence, people might have both strategies in their repertoire and we were interested in examining inter-individual differences in strategy adoption, as well as what factors might drive the selection. This creates an additional strategy selection problem (Payne et al., 1993; Rieskamp & Otto, 2006) – how do people decide which decision strategy to use (for recent



Figure 3.1 Screenshot of the FMAB task from the experiment. Alternatives were presented as simple red boxes with horizontal and vertical yellow lines of varying lengths representing features, and were kept the same throughout all the trials. Participants faced 20 alternatives to allow for large enough sample of observations to learn the function between the length of lines and rewards. Alternatives in the MAB task were presented as simple red boxes without features, however, the rewards were determined with the same reward function as in the FMAB task. In all experiments bandit tasks looked exactly the same, only the underlying function determining the rewards and time horizon was varied.

developments, see Lieder & Griffiths, 2015; Stojić et al., 2016). However, addressing this problem is beyond the scope of the present article, and our intention here is to determine the extent of differences in strategy adoption.

Besides our previous work on the FMAB task (Stojić et al., 2015), there are other studies using related tasks. Niv et al. (2015) examined how people learn what cues to attend to in a task where participants faced three discrete features and they needed to figure out which feature values were predictive of the rewards. The alternatives were also changing from trial to trial, forcing participants to learn the function to perform well, while in our case participants could perform reasonably well by learning mean rewards only. More work has focused on situations with shared contextual cues (i.e. SMAB special case of CMAB). There is an exciting research by

(Schulz, Konstantinidis, & Speekenbrink, 2015, 2016) tackling the SMAB version with a similar set of models. There are several studies with the standard MAB problem where participants are modeled as if tracking the reward distribution of the whole choice set, which can be seen as a single contextual cue, as in the SMAB problem. For example, Gershman and Niv (2015) used such a model to explain conflicting evidence about treating novel alternatives, while Palminteri, Khamassi, Joffily, and Coricelli (2015) used it to explain punishment avoidance. This is supported by neural evidence; Kolling et al. (2016) find evidence that the dorsal anterior cingulate cortex keeps track of mean reward of the choice set (see also Shenhav, Cohen, & Botvinick, 2016). Finally, there is previous work by Redish, Jensen, Johnson, and Kurth-Nelson (2007) and Gershman, Blei, and Niv (2010) on explaining the long-standing puzzle about extinction, based on the related idea that animals infer latent causes instead of rewards (Gershman, 2016; Reverdy, Srivastava, & Leonard, 2014, see also).

In the following sections we provide a normative treatment of the FMAB problem. We describe two Bayesian models, one that tackles the problem by learning the function f , and one that ignores the feature information altogether and learns expected rewards instead, $E[R]$.

3.4 Function Learning Approach

Numerous every-day situations could be seen as contextual-multi armed bandit problems. Yet most models for tackling such problems have been advanced by the machine learning community. These recent modeling developments have been spurred by practical problems faced by artificial agents on the Internet, such as news or product recommendation, or serving ads to website users (e.g. Li et al., 2010).⁴ An early algorithm developed by Auer (2002), LinUCB, uses a linear model to approximate the function and an optimistic heuristic called Upper Confidence Bound (UCB) to balance exploration and exploitation (see, for an application Chapelle & Li, 2011; Li et al., 2010). It is one of the few algorithms that comes with theoretical guarantees in the CMAB problem – it has a known upper-bound on regret⁵

⁴For example, news websites provide articles with certain features (e.g., article topic or length) and they have users for which they record various features (e.g., previous history or topic most often read). Importantly, the articles are constantly changing, i.e. novel alternatives are entering the choice set, and websites would like to keep the user interested by offering articles that he is likely to explore further.

⁵In theoretical analysis of (contextual) multi-armed bandit problems cumulative regret is usually used a measure for comparing the models. Regret is the difference in rewards

when the function relating observed features and rewards is linear. For more general nonlinear functions there are very few theoretical guarantees (but see Srinivas et al., 2009), however practitioners have been able to go quite far by approximating them with deep neural networks (Mnih et al., 2015).

We opted for a Bayesian treatment – we use a full Bayesian model for learning functions, searching for its maximum (i.e. the point with the highest rewards); an approach called Bayesian optimization (for a recent review, see Shahriari et al., 2016). Algorithm 1 shows a generic procedure of how the function learning model \mathcal{M} and decision strategy π interact in Bayesian optimization. In the FMAB task the goal is to find the alternative \mathbf{x}_k that maximizes the unknown function f and allocate remaining trials to it. This can be expressed as:

$$\mathbf{x}^{opt} = \arg \max_{k \in K} f(\mathbf{x}_k).$$

We have a sequential model-based approach to tackling the FMAB problem; we start with a prior belief over possible reward functions and with no observations at $t = 0$, \mathcal{D}_0 . As new alternatives are chosen and their rewards observed, we update our sample of observations, \mathcal{D} , and update the prior beliefs of our model \mathcal{M} . The Bayesian posterior yields the likely reward function we seek to maximize. With decision strategy π we choose the next alternative to evaluate. An important advantage of using a Bayesian function learning model is that the decision strategy can be sophisticated and rely on the posterior distribution around the estimated function to guide exploration. Intuitively, with strategies that take uncertainty about the function into account, we choose alternatives that look promising in terms of expected rewards in light of the available information about the reward function. This reduces the number of observations needed to get to the best alternative and decreases the chance of getting stuck at local maxima. Next we describe how we instantiate the function learning model \mathcal{M} and decision strategy π .

3.4.1 Gaussian Process Regression

We use Gaussian Process (\mathcal{GP}) regression (Rasmussen & Williams, 2006) as a function learning model, and the UCB decision strategy for balancing exploration and exploitation (Auer, 2002; Kaelbling, 1994), in what is

obtained by an alternative and the reward one could have been obtained if the best alternative was chosen.

Algorithm 1 Bayesian Optimization

Require: Model \mathcal{M}^0 with prior beliefs over possible functions; decision strategy π ; $\mathcal{D}^0 = \{\emptyset\}$

1: **for** $t = 1, \dots, T$ **do**

2: Choose alternative,

$$\mathbf{x}_{k^*}^{t+1} = \arg \max_{k \in K} \pi(\mathbf{x}_k; \mathcal{M}^t)$$

3: Observe reward, $R_{k^*}^{t+1} = f(\mathbf{x}_{k^*}^{t+1}) + \epsilon_k^{t+1}$

4: Update sample, $\mathcal{D}^{t+1} = \{\mathcal{D}^t, (\mathbf{x}_{k^*}^{t+1}, R_{k^*}^{t+1})\}$

5: Update prior beliefs over functions in \mathcal{M}^{t+1}

6: **end for**

commonly called the GP-UCB model (Shahriari et al., 2016; Srinivas et al., 2009). The first component, \mathcal{GP} is a Bayesian nonparametric model that can learn very complex functions, adapting itself to the data at hand (Rasmussen & Williams, 2006).⁶ \mathcal{GP} is fully specified by a mean function, $m(\mathbf{x})$, and covariance function, $k(\mathbf{x}, \mathbf{x}')$ (also called kernel):

$$m(\mathbf{x}) = E[f(\mathbf{x})]$$

$$k(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$$

commonly written as $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. We define the prior distribution over possible reward functions by specifying the mean and covariance function. The mean function is often set to zero, $m(\mathbf{x}) = 0$. As we will see in the experiments, one could imbue \mathcal{GP} with initial bias for say positive linear functions, by setting $m(\mathbf{x}) = c\mathbf{x}$ with some positive slope, $c > 0$. More important is the kernel $k(\cdot)$, which determines expected characteristics of functions like smoothness and linearity. The radial basis kernel (also called Gaussian, or squared exponential kernel) is a popular choice and we use it here as well

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(\frac{1}{2l} |\mathbf{x} - \mathbf{x}'|^2\right).$$

From its definition, one can see that the kernel computes the similarity between two observations, i.e. alternatives. For the radial basis kernel

⁶Technically, \mathcal{GP} is a stochastic process where any finite number of random variables also have a joint Gaussian distribution. Because of its neat marginalization properties due to Gaussian assumptions, usual problems with integration in Bayesian models are avoided and we get a computationally tractable Bayesian nonparametric model.

we see that the covariance is close to maximum for alternatives that have almost the same feature values, and it decreases as their distance in the feature space increases.⁷ Note that radial basis kernel has two parameters – the signal variance parameter, σ_f^2 and the length-scale parameter, l . Increasing the length-scale results in greater similarity between more distant alternatives, effectively inducing wider generalization. The signal variance captures the scaling from feature space to rewards. With the kernel in place we defined a prior distribution over possible reward functions, $p(f)$. In particular, expected mean rewards of our functions are set to zero and the radial basis kernel assumes very smooth functions. Draws from this prior can be seen for a one-dimensional case in Figure 3.2A. No alternatives have been chosen – our sample (or memory) is empty $\mathcal{D}^0 = \{\emptyset\}$, and any reward function within prior uncertainty is possible (the shaded area denotes two standard deviations from the mean).

After selecting an alternative k , it becomes part of the agent’s memory together with its features and observed reward, $\mathcal{D}^{t+1} = \{\mathcal{D}^t, (\mathbf{x}^{t+1}, R^{t+1})\}$. Observed features and associated rewards constrain the space of consistent functions. This is evident in the reduction in uncertainty around observed alternatives in the first panel in Figure 3.2B. Note that uncertainty is reduced for nearby alternatives as well, proportionally to their similarity as judged by the kernel. This is the generalization aspect of the model, and if a new alternative would enter the choice set in the vicinity, \mathcal{GP} would be able to make a good prediction how rewarding it would be. The more observations we have in the memory, the more constrained the space of possible functions and smaller the uncertainty about it, and consequently our predictions improve (see bottom panel in Figure 3.2B). Formally, the new prediction \hat{R} for a single alternative (suppressing k and t), \mathbf{x}^* , is obtained from predictive distribution of the \mathcal{GP} ,

$$p(\hat{R}|\mathbf{x}^*, f(X), X) = N(\hat{R}|A, B)$$

where X is a matrix with features of previously observed alternatives from \mathcal{D} . A is a posterior mean estimate, defined as

$$A = m(\mathbf{x}^*) + \mathbf{k}^{*T}(K + \sigma_n^2 I)^{-1}(R - m(X))$$

and B is the posterior variance estimate

$$B = (K^{**} + \sigma_n^2) - \mathbf{k}^{*T}(K + \sigma_n^2 I)^{-1}\mathbf{k}^*$$

⁷There are numerous choices of kernel function. Matern or Ornstein-Uhlenbeck do not give functions as smooth as the radial basis kernel, while a linear kernel renders the \mathcal{GP} equivalent to Bayesian linear regression. Combinations of kernels are possible as well. For more details see Rasmussen and Williams (2006).

with $\mathbf{k}^* = \mathbf{k}(\mathbf{x}^*, X)$ a vector where each element is the similarity between the alternative we are predicting and the alternative in memory, according to the covariance function, $k(\cdot)$. Since in the FMAB task rewards are noisy functions of features, $R = f(\mathbf{x}) + \epsilon$, $\epsilon \sim N(0, \sigma_n^2)$, we allow for another parameter, the noise variance σ_n^2 , which is added to the kernel for the diagonal terms only. $K = K(X, X')$ is a matrix with pairwise similarities between alternatives in the memory, while $K^{**} = K(\mathbf{x}^*, \mathbf{x}^*)$ is a scalar as we are predicting a single alternative – it gives a perfect similarity and hence only variance around it remains, σ_n^2 .

We chose a Gaussian process model for two reasons. First, it is a powerful way of approximating functions⁸ and we obtain a full posterior distribution in a tractable manner. These are very desirable properties for a rational model – the former means that model can easily learn nonlinear functions, while the latter means that decision strategies can exploit posterior uncertainty to make improve their choices.

Second, our choice was guided by current evidence on human function learning. Exemplar models are among the most successful approaches developed for explaining how people learn relations between features of objects and a continuous or discrete criterion. The Generalized Context Model (Medin & Schaffer, 1978; Nosofsky, 1986, GCM) and ALCOVE (Kruschke, 1992), two prominent models in this family, also rely on kernels to account for possible non-linearities in the mapping from features to criterion values. Similar to a Gaussian process model, exemplar models are also memory-based, they store (all) observations in the memory and when a new observation comes, its value (reward or category) is predicted based on its similarity to the items stored in memory. A second influential class of models are called rule-based models (e.g. Brehmer, 1994; Koh & Meyer, 1991). A prime example is the rule that combines features linearly, but other parametric functions have been considered as well. There has been a long-standing debate in the category and function learning literature which class describes people’s behavior best (Ashby & Maddox, 2011; McDaniel & Bussemeyer, 2005), but it turns out that both can be unified in the Gaussian process framework Lucas, Griffiths, Williams, and Kalish (2015). A rule-based linear model can be instantiated as a \mathcal{GP} with a linear kernel, while a version of the GCM can be expressed as a \mathcal{GP} with a radial basis kernel. This is recent development though and empirical evidence is not yet in whether they are better models than established ones like GCM or

⁸There is an interesting relation between the Gaussian process and neural network models that are used as function approximators (Mnih et al., 2015). Neal (1996) has shown that a neural network with infinitely many hidden units and Gaussian priors on the weights is equivalent to a Gaussian process model.

ALCOVE (but see Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2016; A. G. Wilson, Dann, Lucas, & Xing, 2015). One contribution of the present work is to provide additional evidence on the matter.

3.4.2 Upper Confidence Bound strategy

The second component of the GP-UCB model is a decision strategy, UCB (Auer, 2002; Kaelbling, 1994). It is defined as

$$\arg \max_{k \in K} \hat{R}_k + \alpha \sqrt{v_k}$$

where \hat{R}_k and v_k are the mean and variance of the posterior predictive distribution of rewards for option k , and α is a free parameter that determines how much an agent relies on uncertainty of the function in comparison to the mean rewards. The second term can be interpreted as the confidence bound (as in statistics, if α is set to the 0.95 percentile of a Normally distributed variable), which is where the name comes from. For positive α its highest value is located where the uncertainty in the \mathcal{GP} model is large (exploration) and where the mean predicted rewards are high (exploitation). UCB then selects the alternative with the maximum.

Selecting the Bayes optimal sequence of choices is typically computationally intractable, such solutions are available only for very limited scenarios (for example, Gittins indices are available in the Bernoulli MAB setting, via dynamic programming, Whittle, 1980). This has led to introduction of myopic strategies, such as UCB or Thompson sampling (Chapelle & Li, 2011; Thompson, 1933), that approximate the optimal solution at a fraction of the cost. These strategies are usually compared on a frequentist measure of cumulative regret. For GP-UCB there is a theoretical guarantee that it achieves relatively small regret for certain classes of nonlinear functions (Srinivas et al., 2009). There are better approximations of the Bayes optimal choice allocation. One is Bayes-adaptive Monte Carlo Planning (BAMCP) strategy that explicitly takes the horizon information into account (Guez, Silver, & Dayan, 2012, 2014), unlike the myopic UCB that is blind to it. As the final trials of the task draw to end an agent should switch more to exploiting the knowledge it acquired that far. UCB can look as if it is choosing in this manner, but this is only when uncertainty is reduced over time in the \mathcal{GP} . Computational cost for implementing BAMCP is still quite high – for version of the CMAB problem we presently focus on we do not lose much by using a UCB instead of BAMCP decision strategy, while saving a lot of computation.

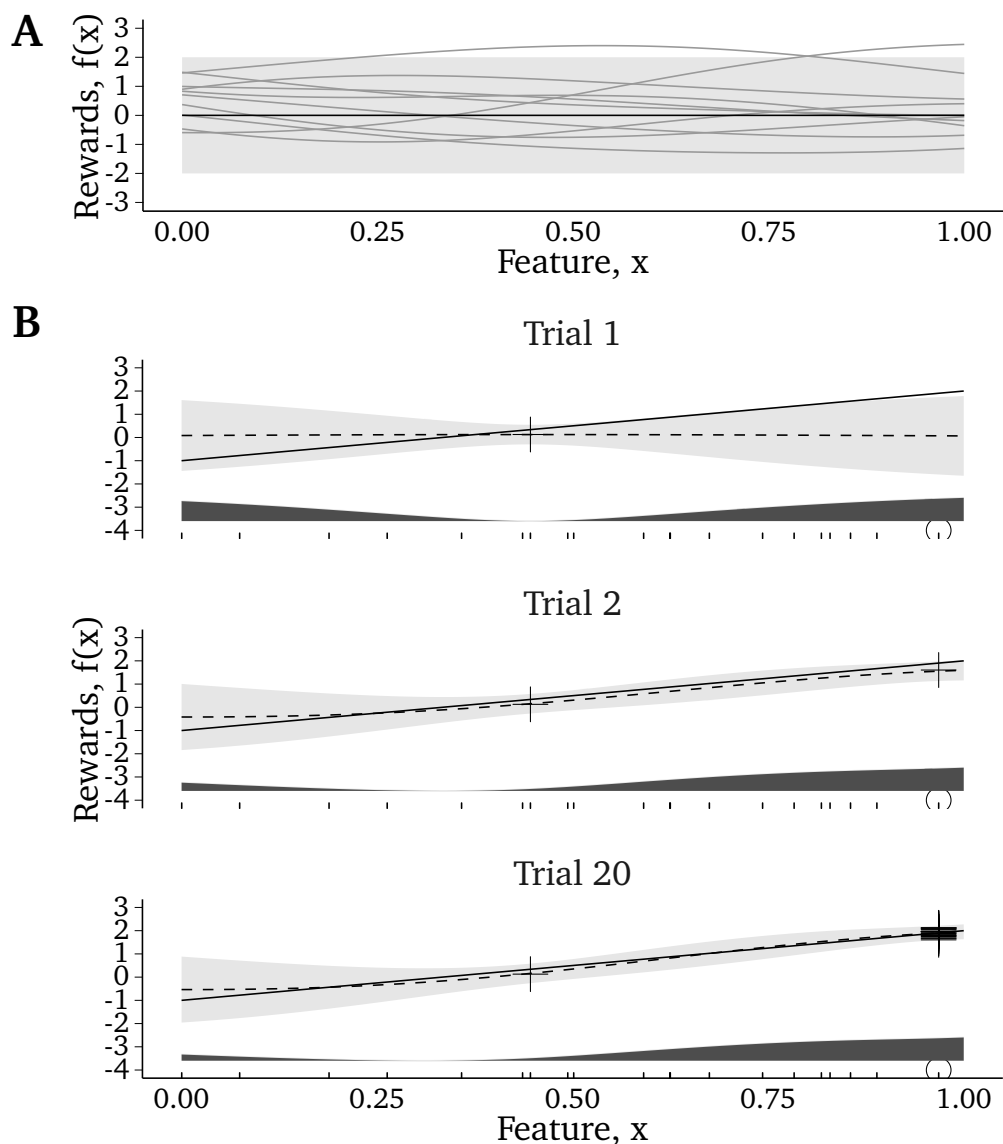


Figure 3.2 Illustrating the GP-UCB model with a single feature function. (A) The \mathcal{GP} prior set by the mean and kernel function reward functions is not constrained and many forms are possible. Gray lines are random draws from the prior, the black line is the mean reward, and the shaded region shows two standard deviations from the mean. (B) Trial 1 shows reduction in uncertainty around the first chosen alternative (“+” symbol, available alternatives are marked on x-axis), but also for neighboring points. The black line is the true reward function, while the dashed line is one estimated by \mathcal{GP} . The dark gray area are values assigned by UCB to points in the feature space – it chooses the rightmost (“o” symbol). After trial two it discovered a high rewarding region and allocates remaining choices in this region. The reward function: $R = -1 + 3x + \epsilon$, $\epsilon \sim N(0, .15)$. The GP-UCB parameters: $\sigma_f^2 = 1$, $\sigma_n^2 = .15$, $l = .5$, and $\alpha = 2$.

Empirical evidence from standard MAB and RL tasks regarding heuristic strategies that take into account uncertainty is mixed, with early evidence pointing toward a simple strategy that relies on mean rewards only (Daw, O’Doherty, Dayan, Seymour, & Dolan, 2006). The most popular strategy in this class is the softmax rule, $P(C = k) \propto \exp(\theta \hat{R}_k)$, (Luce, 1959; Sutton & Barto, 1998), a probabilistic decision strategy that chooses alternatives roughly proportionally to their mean rewards. The probabilistic component can be interpreted as decision noise driving exploration. Recent studies have found evidence for information seeking strategies based on uncertainty (Speekenbrink & Konstantinidis, 2014; R. C. Wilson, Geana, White, Ludvig, & Cohen, 2014). The mixed results may be due to the experimental tasks, where differences in choices between these two different classes are small. In contrast, in the FMAB task uncertainty can play a larger role – sampling a more informative alternative, as indicated by the uncertainty around it, can be highly beneficial as knowledge generalizes to other alternatives. Hence, strategies like UCB might also prove to be a good model of human behavior as well.

In Figure 3.2 we illustrate UCB choices in a single-feature example in more detail. In 3.2A, while no alternative is chosen, the prior does not discriminate between the alternatives – their predicted mean rewards and uncertainties are exactly the same, so UCB will make a choice at random. In 3.2B one observation is in and both predicted means and posterior variances change. If mean rewards and uncertainty is equally weighted, the overall value will roughly follow the upper edge of the posterior (lightly shaded area) – the true valuation of all possible alternatives (rescaled to 0,1 range) is given by the dark gray area. Given that the posterior distribution of the \mathcal{GP} changes trial-by-trial, how UCB balances exploration and exploitation will change dynamically as well. After the uncertainty in the region for which the \mathcal{GP} predicts the highest rewards has shrunk and uncertainty in other parts of the function cannot compensate for those high rewards, UCB will stay in the same region in the remaining trials, switching to pure exploitation.

One of the important predictions of the GP-UCB model, and Bayesian optimization in general, is that when a function is simultaneously learned and optimized, what looks like a low rewarding region is quickly abandoned and exploration moves to alternatives that promise higher rewards. This is a specific pattern we will be looking for when examining people’s behavior in the experiments. One consequence of this pattern, when coupled with local learning kernels like the radial basis, is that uncertainty about the function is relatively high in regions with few observations. The estimated reward function could be even quite wrong in those regions, in-

dicating that even rationally some ignorance is warranted when pursuing rewarding outcomes. As a result, when new alternatives enter the choice set, predicted rewards will be highly uncertain if they have features corresponding to a low rewarding region, and errors will be likely.

3.4.3 Mean Tracking Approach

As stressed previously, one can also tackle the FMAB task by ignoring the feature information and simply learning the mean rewards of the alternatives. We define here a Bayesian Mean Tracking (BMT) model that learns mean rewards, \hat{R} , through Bayesian updating (Gershman, 2015). The decision strategy that uses the estimated mean rewards stays the same – the UCB strategy. The BMT model is obtained with a simple modification of the kernel in the \mathcal{GP} model

$$k(\mathbf{x}, \mathbf{x}') = \begin{cases} 0 & \mathbf{x} \neq \mathbf{x}' \\ \sigma_f^2 & \text{otherwise} \end{cases}$$

This kernel gives positive similarity only for exactly the same alternative, scaled by the free parameter σ_f^2 , no similarity otherwise. We also include the noise variance parameter, σ_n^2 . With such a kernel we have Bayesian updates only for the alternative that is sampled and there is no generalization to other alternatives in the choice set.

This model can be viewed as a Bayesian version of the delta-rule or Rescorla-Wagner model (Rescorla & Wagner, 1972). It learns more efficiently than these models, as it takes into account the uncertainty of each alternative when updating the values. In other words, it has an alternative specific dynamic learning rate, while the Rescorla-Wagner model has constant learning rate for all alternatives. In spite of this boost in learning speed, BMT-UCB is warranted to be slower than GP-UCB. This is because BMT-UCB treats each alternative independently and cannot generalize, it needs to sample alternatives at least once to estimate their mean reward. In contrast, after trying a few alternatives, GP-UCB can already eliminate a subset of similar alternatives that are also likely to yield low reward. The difference in our experiments will be moderate – since we have only a relatively small number of alternatives, and BMT-UCB can also quickly identify the most promising ones. In CMAB problems with either a larger number of alternatives or where new alternatives ones are encountered often, GP-UCB would outperform BMT-UCB by a much larger margin.

The differences between the models will be the most obvious in the beginning of the experiment, where most transfer of learning will occur in

GP-UCB – this will be reflected in quicker avoidance of alternatives with features associated with low rewards in these early trials, while the BMT-UCB model would explore uniformly across the choice set, including these low rewarding alternatives. For this reason we will focus on analyzing exploration patterns of participants in early trials of the FMAB task.

3.5 Summary

Optimal solutions for the CMAB problem are not available. We relied on a Bayesian nonparametric modeling approach coupled with sophisticated heuristics to arrive at a good approximation to the optimal solution. Our theoretical analyses have shown there are two qualitatively different ways of tackling our FMAB task – one based on learning the function between features and observed rewards (GP-UCB), and another based on tracking the mean rewards of alternatives, ignoring the features altogether (BMT-UCB). These are not only close to optimal solutions of the FMAB problem, they stand a good chance of being good models of people’s behavior in the task.

The BMT-UCB model serves as a reference frame against which to compare performance and exploration patterns of the approach we are primarily interested in – the function learning approach as embodied by GP-UCB. We illustrated exploration patterns of agents guided by function learning that we will be looking for in the experiments. Allowing for such qualitatively different ways of solving the task was intentional, as we are interested in strategies people choose to tackle the task. Nevertheless, our prediction is that the majority of participants will use the former – realistic decision scenarios will often rely on generalization of knowledge and function learning is a reasonable default. In the following experiment our goal was to detect whether people’s choices and exploration patterns were guided by feature information and function learning. Since it might be difficult to detect whether people simultaneously learn and optimize the function based on choices in the FMAB task only, we relied on two mechanisms to facilitate the detection: (1) a control group that had a standard MAB task where we can be sure their exploration patterns were not due to function learning, and (2) an additional functional knowledge task after the FMAB task to assess the extent of knowledge about the function.

3.6 Experiment 1A and 1B: Influence of feature information on learning and exploration

In Experiments 1A and 1B we examined whether people simultaneously learn the function and search for its maximum. In addition, we investigated whether there are inter-individual differences in the approach people use to deal with the FMAB task – function learning or mean tracking.

Both studies followed the same between-subject experimental design. One group is randomly allocated to a MAB task where feature values were not visually displayed, while the other group is allocated to a FMAB task where feature values were visible (see Figure 3.1). Importantly, rewards in both conditions were determined by the same function – a positive linear combination of two features. It has been established that people can readily learn such functional forms (Brehmer, 1974; Busemeyer et al., 1997). We intentionally kept the first experiment easy; if people could not handle the simplest versions of the task, there would be little hope for more complex scenarios.

Participants in the FMAB condition completed an additional functional knowledge task, where we examined the extent to which they had learned the function and could use the acquired knowledge to make better choices when facing alternatives they have not seen during the FMAB task. Our rational analysis shows that both models that rely on function learning (GP-UCB) and those that ignore it (BMT-UCB) have similar performance on the stimuli for the bandit task; where they differ is extrapolation to new stimuli. Hence, generalization is a true test of how much participants have relied on function learning in coping with the FMAB task (DeLosh et al., 1997).

We expected that a large majority of participants would use the function learning approach. Following the GP-UCB model, their exploration patterns should be heavily skewed due to feature information – people should allocate disproportionately more choices to alternatives with promising feature values. This should be most evident in early trials, when people learn and explore the most. Moreover, the effect should be enhanced as people usually have a strong prior for positive linear relationships (Brehmer, 1974; Busemeyer et al., 1997). We also expected some inter-individual differences. Given that our FMAB task had only a limited number of alternatives, it could be solved almost equally well by ignoring the feature information and paying attention only to the mean rewards of each alternative, as in the classical MAB task. We expected some people to take this strategy instead.

The two studies are nearly identical. The first was conducted on larger

number of participants recruited via Amazon’s Mechanical Turk online labor market (AMT, <https://www.mturk.com>), while the second study was a lab replication, to verify the quality of Mechanical Turk data.

3.6.1 Method⁹

Participants

In total, 261 participants took part in Experiment 1A and 1B. The number of participants and their socio-demographic characteristics for all studies in the paper are given in Table 3.1, along with several other useful details. Participants were recruited either via Amazon’s Mechanical Turk (AMT, <http://mturk.com>, in Experiment 1A) or from the Universitat Pompeu Fabra subject pool (Experiment 1B). Data of a number of participants did not pass the quality checks and were excluded from the analysis. Our procedures for ensuring data quality are described in more detail in Appendix 3.A. Participants received a fixed minimal payment plus a performance dependent bonus.

Bandit task

The task comprised of 100 trials. On each trial, participants saw the same 20 alternatives and were called to choose one of them. After making a choice k in trial t , they were informed of the reward R_k^t associated with their choice. For each arm $k = 1, \dots, 20$, the reward on trial t was computed according to the following equation:

$$R_k^t = w_1 x_{1,k} + w_2 x_{2,k} + \epsilon_k^t.$$

The two feature values, $x_{1,k}$ and $x_{2,k}$, of each alternative k were drawn from a uniform distribution $U(0.1, 0.9)$, for each participant at the beginning of the task. Weights for all participants were set to $w_1 = 2$ and $w_2 = 1$ in Experiment 1A and to $w_1 = 20$ and $w_2 = 10$ in Experiment 1B. The error term, ϵ_k^t , was drawn randomly and independently for each arm from a normal distribution, $N(0, 0.0625)$ in Experiment 1A, and $N(0, 6.25)$ in Experiment 1B.¹⁰ Note that the feature values stayed the same throughout all 100

⁹Software, exact instructions and stimuli used in all the experiments, are publicly available at the Open Science Framework website: <https://osf.io/fmn45>.

¹⁰We determined the exact size of the error term by examining the results of the pilot study reported in Stojić et al. (2015). In the pilot we used the same weights as in Experiment 1A, but the variance was set to one. As a result learning was hard, and participants exhibited noisy behavior that was difficult to model. We decreased the error variance substantially to boost learning.

Table 3.1 Overview of experiments and characteristics of participants across experimental conditions. Female denotes number of female participants. For age we give the mean and in parentheses the standard deviation and range. For performance contingent (PC) fee and duration we give mean and standard deviation in parentheses. The last row (Total) provides descriptors aggregated over all experiments, for example the total number of participants, or mean age.

Experiment	Condition	Function	N	Female	Age	SU fee	PC fee	Duration (min)	N_e
1A	FMAB-pl	Positive linear	96	47	35.8 (12.1, 18-74)	0.3	0.8 (0.1)	8.9 (4.3)	5
	MAB-pl	Positive linear	90	32	33.5 (11.3, 19-68)	0.3	0.5 (0.1)	5.0 (1.9)	11
1B	FMAB-pl	Positive linear	37	23	21.6 (3.6, 17-36)	4.5	3.7 (0.4)	12.8 (4.6)	0
	MAB-pl	Positive linear	38	26	20.8 (2.5, 17-28)	4.5	2.3 (0.2)	10.0 (5.5)	1
2A	FMAB-ml	Mixed linear	102	44	35.6 (11.6, 18-69)	0.3	0.2 (0.1)	10.2 (5.7)	11
	MAB-ml	Mixed linear	88	35	37.1 (11.8, 18-68)	0.3	0.6 (0.2)	8.1 (9.5)	5
2B	FMAB-q	Quadratic	102	61	37.5 (11.3, 21-71)	0.3	0.4 (0.1)	10.5 (10.6)	8
	MAB-q	Quadratic	84	36	34.3 (10.8, 20-74)	0.3	0.5 (0.2)	6.7 (5.9)	8
3	FMAB-pl	Positive linear	90	41	35.6 (11.6, 19-70)	0.3	0.3 (0.1)	9.8 (5.0)	10
	fFMAB-pl	Positive linear	71	33	35.1 (10.6, 19-74)	0.3	0.4 (0.1)	24.7 (11.9)	0
	fFMAB-pls	Positive linear	56	31	32.7 (11.7, 19-61)	0.3	0.4 (0.1)	17.0 (43.2)	2
	FMAB-q	Quadratic	88	42	34.5 (11.4, 18-69)	0.3	0.4 (0.1)	11.9 (15.8)	7
	fFMAB-q	Quadratic	69	30	36.4 (10.8, 20-62)	0.3	0.4 (0.1)	24.2 (10.8)	4
	fFMAB-qs	Quadratic	57	30	35.5 (11.5, 20-69)	0.3	0.4 (0.1)	20.0 (14.5)	7
Total			1068	511	34.1 (10.1, 17-74)				79

Note. N = Number of participants (without excluded participants); N_e = Number of participants excluded due to quality controls; SU fee = Show-up fee (in US\$); PC fee = Mean performance contingent fee (in US\$); MAB = Multi-armed bandit task; FMAB = Feature-based multi-armed bandit task.



Figure 3.3 Alternatives in the functional knowledge task were designed to examine whether participants learned the function. Features were always visible and rewards were governed by the same function as in the bandit task. Participants with some knowledge about the function should be able to achieve better-than-chance performance.

trials, while payoffs were changing from trial to trial due to the error term. The only difference between the conditions was that the feature values, $x_{1,k}$ and $x_{2,k}$, were visually displayed in the contextual version (FMAB-pl condition) but not in the classic one (MAB-pl condition), as illustrated in Figure 3.1. Participants were randomly assigned to the conditions.

Functional knowledge task

In the final phase, participants faced a functional knowledge (FK) task aimed at examining whether they had acquired knowledge about the function governing the rewards during the bandit task. Only participants in the FMAB condition continued to this choice task, which consisted of 70 trials where in each trial the participants saw three new alternatives. The structure of the task was very similar to the bandit task – the rewards were determined by the same function, and visually they looked the same (Figure 3.3). To encourage participants to use their acquired knowledge, we removed the opportunity to learn further – we did not provide them feedback about the outcome of their choices. The problem was thus reduced to a multi-attribute choice task, as there was no longer an exploration–exploitation trade-off and participants had to generalize their knowledge to new decision situations. Finally, there were five types of items in the FK task, four of them differing in terms of difficulty and one identifying whether participants had learned that one feature had a larger weight. The sampling procedure and item types are described in more detail in Appendix 3.B.

Procedure¹¹

Participants completed the experiment either in a browser via AMT (Experiment 1A) or on desktop computers in the lab at Universitat Pompeu Fabra (Experiment 1B). We have used a custom software written in Javascript, with the help of jsPsych and Psiturk library (Gureckis et al., 2015; Leeuw, 2015).

We first presented participants with consent form. Only those that accepted it could start the experiment. Then participants completed a brief socio-demographic questionnaire and read instructions about the tasks in the experiment. Instructions explained that they would be presented with a decision making task where they would make choices between 20 alternatives for many trials. We explained in detail that for each choice they would receive experimental points that would at the end be converted to money, with an advertised exchange rate. The goal of the game was to win as many experimental points as possible. We also informed them that they would see the same alternatives in every round, but that the rewards associated with each alternative might vary from round to round. Finally, before they started with the bandit task, we asked several questions about the information presented in the instructions in order to check how much attention they paid to the instructions (for more details on attention questions, see Appendix 3.A).

After reading the instructions and completing the questionnaires, participants started the bandit task. On each trial, alternatives were presented in the form of simple square-shaped buttons (see Figure 3.1) and they had as much time as they needed to select an alternative via a mouse click. The number of points won or lost was then displayed immediately below the alternative until they pressed the ENTER key, which would display the next trial. Throughout the task, a counter displayed the total points received thus far, the number of the current trial, and the total number of trials in the phase. Buttons in the MAB condition were empty, while in the FMAB condition feature values were displayed on each button in the form of one horizontal and one vertical line, both starting from the lower left corner of the square (see Figure 3.1). We randomized whether a certain feature was represented as a vertical or an horizontal line across participants. Since features and error terms for each participant were drawn randomly, rewards, alternatives and positions on the screen were effectively randomized as well. After participants in the MAB condition made their 100 choices, we informed them about their total earnings, asked a few optional questions

¹¹Readers can try out one of the early experiments reported in Stojić et al. (2015) at the following URL: <http://experimentnext.com/CMABvsMABexp1>.

about their experience with the experiment and thanked them for their time.

Participants in the FMAB condition continued to the FK task. We announced the task in the instructions at the beginning in this condition, but without specifying details. After finishing the FMAB task, they read the instructions for the FK task. We told them they would face new alternatives in every trial, but would not see any feedback and would no longer see the running total. We also informed them that their payoff would still be affected by their choices. After they completed the task they went through the same procedure as MAB participants.

Information that we recorded for each participant was basic demographic data – age, gender and type of studies, answers to attention questions, stimuli characteristics, choices and response times in both bandit and functional knowledge task.

3.6.2 Results and Discussion¹²

Performance in the bandit task

The Bayesian models set the benchmark performance in the bandit task, shown in Figure 3.4A (see Appendix 3.D for details on parameter values of the models and estimation procedure). We use rank of the chosen alternative as a measure of performance rather than expected reward, as stimuli were drawn randomly for each participant and expected reward would provide a noisier measure. The models achieve large improvements already in the first block of trials. The function learning based model (GP-UCB model) performs better than the model that ignores the feature information and tracks only the mean rewards (BMT-UCB). The GP-UCB model manages to find the best alternative in the choice set very quickly, achieving an overall mean rank of 1.3, while the BMT-UCB model achieves performance very close to the GP-UCB model ($M_{rank} = 2.3$). Such a small difference was expected given the small number of alternatives in our task. The advantage of function learning increases with the number of alternatives and when knowledge has to be generalized to new alternatives entering the choice set. As expected, the differences between the models were the largest at

¹²The code we used for analyzing the data and modeling is publicly available at the same website as the software used for conducting the experiments: <https://osf.io/fmn45>. The raw data from all experiments are publicly available on Figshare data repository: <http://dx.doi.org/10.6084/m9.figshare.3189748>. The code is written in R programming language (R Core Team, 2015), in a way that with minimal effort one could reproduce the results of the analysis and produce all the figures presented in the article.

the beginning of the task and this is where we will focus our exploration analysis.

Bayesian models set a high bar and although we expected participants to improve substantially over time, their learning was unlikely to be as fast as that of Bayesian models. Indeed, Figure 3.4B shows that both MAB and FMAB participants were performing better than chance (rank of 10.5) and improved substantially over time, as validated by one-tailed Wilcoxon signed rank tests. For the MAB condition it shows a significant difference in mean choice rank between the first ($M = 9.30$) and last block ($M = 3.83$), $Z = 4080$, $p < 2.2E - 16$. Similarly, the difference ($M = 8.15$ and $M = 3.96$) is significant for FMAB condition, $Z = 4496$, $p = 1.1E - 15$. This indicates that participants understood and were engaged in the task. However, human performance was not as good as that of Bayesian models. By the end of the bandit task, on average humans managed to identify alternatives that were in the top 20%, but rarely the best. Note that parameters of the Bayesian models were fitted to the stimuli, while participants did not enter the experiment with their “parameters” all set for the task they would encounter, thus, the difference is somewhat exaggerated.

Except for the first block, there is seemingly little difference in choice performance between the conditions. Even though BMT models achieve similar performance to GP models, people as a rule learn slower than statistical models and we expected that feature information would substantially benefit learners in the FMAB condition, so the small differences between conditions surprised us initially. Even though we envisaged that some participants in the FMAB condition would ignore the feature information and opt for a mean reward tracking strategy, the proportion turned out to be quite large. We used the performance on the FK task to distinguish between these two types. Figure 3.4C shows the results of performing a K-means clustering on mean choice ranks from the FK task. We find a large group of people that seem to have very little knowledge about the function, making choices almost randomly (“Mean trackers” group, $N = 43$, with mean rank performance of $M = 1.94$, $SD = 0.28$), and another large group that possesses accurate function knowledge, choosing the alternative with highest function value (i.e. largest reward, based on the reward function used in the bandit task) most of the time (“Function learners” group, $N = 53$, $M = 1.17$, $SD = 0.17$). Figure 3.4D illustrates choice performance of FMAB participants, once these inter-individual differences in tendency to rely on function learning are taken into account. We see that performance of Function learners is much better than that of Mean trackers, with Mean tracker performance resembling closely that of MAB participants (see Figure 3.4B). Function learners achieve an overall mean rank of $M = 4.59$ ($SD = 1.95$),

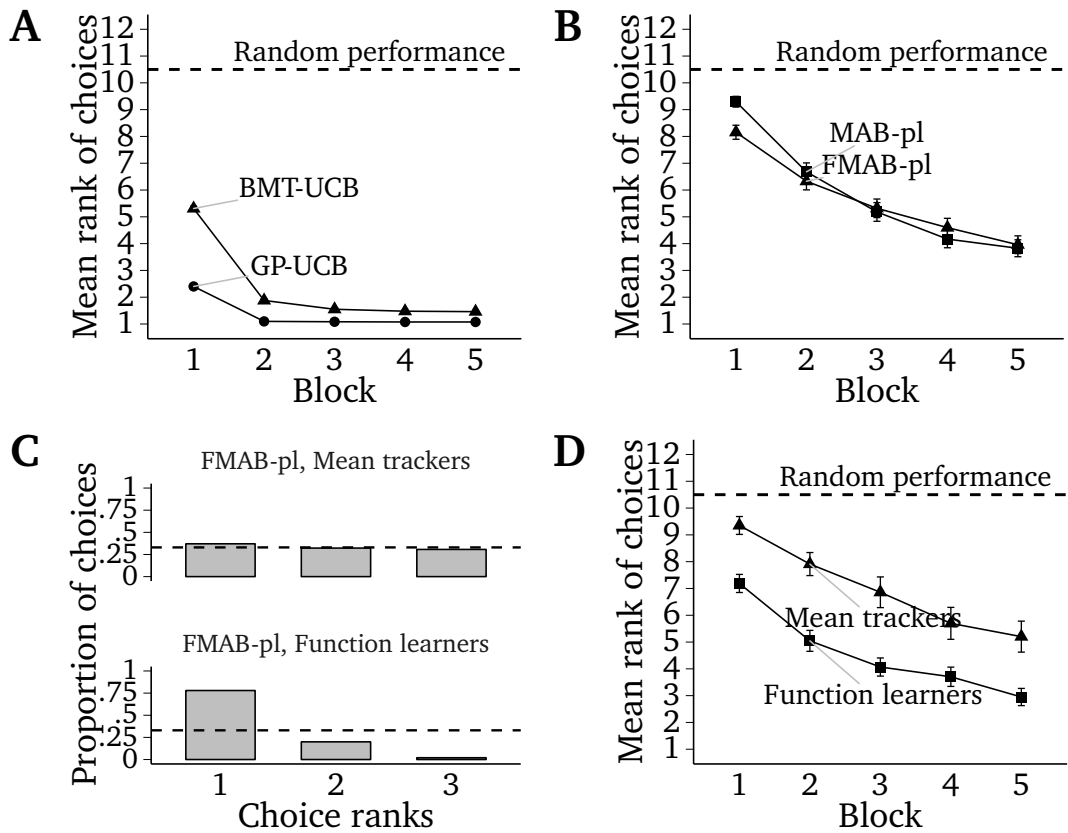


Figure 3.4 Choice performance of Bayesian models and participants in Experiment 1A. (A) The performance of Bayesian models in the FMAB task – mean accuracy of models’ choices (the lower the rank the better) increases across trials (grouped in five blocks of 20 trials). Function learning based model (GP-UCB) achieves the best performance, but Bayesian mean tracker model is very close (BMT-UCB). (B) Participants learn to make good choices in both MAB and FMAB task – participants’ mean accuracy increases over time, but does not reach the level of Bayesian models. There is also seemingly little effect of feature information. (C) However, according to the performance in the FK task there are strong inter-individual differences in how well the participants have learned the function in the FMAB condition. We find a cluster of people that do seem to learn the function (Function learners) and a cluster that does not exhibit almost any knowledge about the function (Mean trackers). (D) Once these inter-individual differences are taken into account, we see that Function learners perform much better than the MAB group, while Mean trackers perform similar to the MAB group.

while Mean trackers reach $M = 7.00$ ($SD = 2.84$), a significant difference as indicated by a two-tailed Wilcoxon rank-sum test, $Z = 1700$, $p = 3.6E - 05$.

A potential caveat with the clustering analysis is that the Mean tracker

group might also include unmotivated participants that performed poorly in both tasks. Indeed, there appear to be such participants, although not many – six participants in the Mean tracker and one in the Function learner group have mean rank above 10 in fifth block. This is not a major issue, however: as shown in Figure 3.4D, even with these participants the Mean tracker group still improves greatly over time, and without them, the difference between Mean trackers ($M = 6.32$, $SD = 2.41$) and Function learners ($M = 4.50$, $SD = 1.87$) is still very large, $Z = 1376$, $p = .0005$ (two-tailed Wilcoxon rank-sum test).

Exploration

To tackle our main question more directly – whether people’s behavior exhibits the patterns of simultaneously learning the function and searching for its maximum – we examined allocation of participants’ choices with respect to feature values (Figure 3.5). We focus on the first 10 trials where people learn and explore the most and our models show the largest differences. We predicted that people’s exploration patterns in the FMAB condition would be skewed toward alternatives whose feature values place them in high rewarding regions. This is predicted by our GP-UCB model, as illustrated in Figure 3.5A. The BMT-UCB model has no way of generalizing knowledge from one alternative to the other, so in the beginning it tends to allocate choices uniformly over all alternatives. There is a slightly larger proportion of choices in the high reward region (upper right area), owing to the BMT-UCB model learning optimally. It adjusts the learning rate with the posterior estimates of rewards, and uses the posterior in balancing exploration and exploitation. The GP-UCB exploits the feature information to learn the function between feature values and rewards, and as a result generalizes from one alternative to all others in the choice set. It starts with no biases and an expectation of zero rewards. With a single alternative tried out, it learns to avoid those with similar features if the reward was negative. With two to three alternatives it already identifies a high rewarding region in the case of simple functions like in this experiment. This leads to very small proportions of choices allocated to low reward regions in the feature space, and results in significant differences in performances of the models in the beginning, as illustrated in Figure 3.4A. On average, the GP-UCB model tried out 3.6 alternatives, while the BMT-UCB tried out 8.9 alternatives. If we were to look at the last trials, the allocations between the models differ very little, both BMT and GP models identify the good alternatives by that time.

Figure 3.5B depicts participants’ allocation of choices in feature space

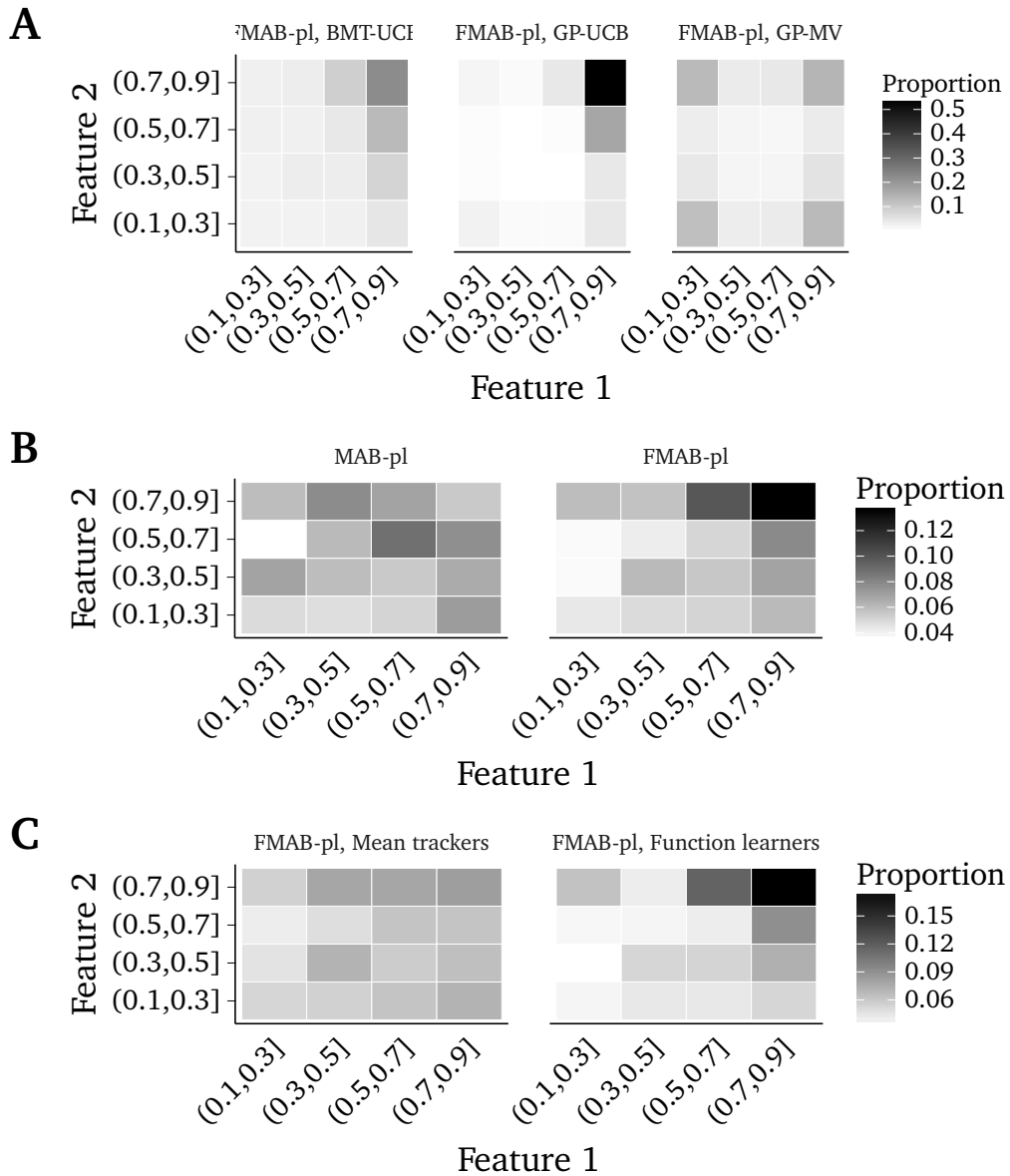


Figure 3.5 Exploration patterns of Bayesian models and participants in the bandit tasks in Experiment 1A. (A) Proportion of all choices in the first 10 trials allocated by models to alternatives with feature values falling into one of the four bins indicated on x-axis for Feature 1 ($w_1 = 2$) and y-axis for Feature 2 ($w_2 = 1$). The GP-UCB identifies the high rewarding region more quickly than the BMT-UCB. GP-MV makes choices to learn the function as well as possible. (B) Participants in the FMAB condition started allocating a larger proportion of choices to high rewarding region quickly, while in the MAB condition they are exploring very uniformly. (C) Cluster of FMAB participants that has poor knowledge of the function (mean trackers) explores the same as MAB group, while cluster with good knowledge (function learners) explores similar to the GP-UCB model.

in the first 10 trials, separately for the MAB and FMAB condition. Choices of MAB participants are close to uniformly distributed over the alternatives. In comparison with the BMT-UCB, people learn more slowly and did not discover the good alternatives in the first 10 trials. By the end of the bandit task, however, they do narrow down to good alternatives and their allocations are highly concentrated in the high rewarding region (see Figure 3.C.1 in Appendix 3.C). Choice allocation of the FMAB group differs strongly from that of the MAB group in the first trials of the task. They start avoiding the alternatives with features associated with low rewards (lower left region) and choose more frequently those in high rewards (upper right region), as determined by the reward function. We validated the observed differences with a permutation-based statistical test, which showed a significant difference between the two distributions, $D_H = .131$, $95\%CI = [.114, .175]$, $p < 1E-07$.¹³ As the permutation test is silent on direction of the difference, we computed the average number of choices (out of 10) allocated to alternatives with both feature values between 0.5 and 0.9 and conducted a one-tailed Wilcoxon rank-sum test. The test shows that the FMAB-pl group allocated on average 3.67 ($SD = 2.43$) choices to the upper right region, significantly more than MAB-pl group (2.91, $SD = 1.92$), $Z = 5032$, $p = .025$. Finally, according to our modeling analysis, the MAB-pl group should have tried more alternatives than the FMAB-pl group. This is not the case, MAB-pl group sampled on average 16.54 ($SD = 4.07$) alternatives (out of 20), very similar to FMAB-pl group (16.12, $SD = 4.01$), $Z = 4601$, $p = .217$. The observed difference is akin to the difference in exploration between the GP-UCB and the BMT-UCB model, although the distributions are not nearly as skewed as for Bayesian models (note the different scales for each figure).

Figure 3.5C breaks down the FMAB condition on allocations of Mean trackers and Function learners. The exploration pattern of Mean trackers strongly resembles that of the MAB group. If we repeat the tests reported above, comparing Function learners with the MAB group instead, we get larger differences in expected directions. Moreover, a test of difference in number of alternatives tried is now significant as well – the MAB-pl group sampled on average 16.54 ($SD = 4.07$) alternatives (out of 20), more than Function learners (15.25, $SD = 4.31$), $Z = 2800$, $p = .0389$.

¹³We compute the Hellinger distance (D_H) between two discrete distributions and then randomly permuted the labels a million times to get an empirical null distribution of no difference between them. The disadvantage of this procedure is that it also breaks spatial patterns, treating each bin in the distribution as being independent of the others. The confidence interval of the distance between the distributions is computed with bootstrap procedure.

Observed choice patterns do not correspond to behavior one would expect of agents that are solely concerned with learning the function. Allocating choices to learn the function as well as possible (i.e. active learning) is illustrated by the GP-MV model with Maximum Variance (MV) choice rule (Figure 3.5A). This rule ignores the estimated mean rewards and chooses the alternatives with largest uncertainty, notably in the corners of the feature space. It effectively shows the pattern of choices people might make if they were only after learning the function, and not trying to maximize the rewards in the same time. We can clearly see that neither the FMAB group nor Function learners allocate choices symmetrically to all corners or edges. Instead, their allocation seems to be guided by both learning the function and exploiting it.

Generalization performance in the FK task

Figure 3.6 shows the performance of BMT-UCB and GP-UCB model in the FK task. The BMT-UCB cannot generalize to new alternatives and can make choices only randomly. The GP-UCB model is however very good at extrapolating to new decision situations, but indicatively, it is not predicting perfectly – in about 10% of trials it makes an incorrect choice. Mean trackers and function learners exhibit similar mean rank performance (Figure 3.4C).

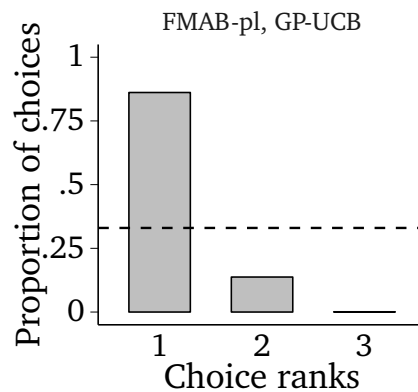


Figure 3.6 The GP-UCB model is able to generalize and performs very well on the FK task on stimuli from Experiment 1A, after learning during the bandit task. While the BMT-UCB model performs only slightly worse than GP-UCB in the bandit task, it cannot generalize what it has learned to the FK task and its performance here is at chance level.

Similarities are even more striking when looking at FK task performance broken down into item types (Figure 3.C.2). Both GP-UCB and Function

learners err in difficult and weight comparison items. Even though in GP-UCB predictions we kept using the parameters estimated for the bandit task, including the positive uncertainty parameter α in the UCB choice rule, its errors are not due to exploration tendencies. Even if we set this parameter to zero, we still get essentially the same results. Imperfect generalization is a result of imperfect knowledge of the function – the GP-UCB model still has a large uncertainty in regions it did not experience, and can misrepresent the function in those regions. The choice sets in the FK task include alternatives from such regions and the model then ends up making mistakes. This result is also suggestive about the type of function learning people do – namely, that people’s functional knowledge is based on local, rather than global knowledge (Bussemeyer et al., 1997).

Results of Experiment 1B

The results on AMT sample in Experiment 1A were replicated on the more homogeneous, student sample in a lab experiment in Experiment 1B (see Figure 3.C.3 for results). The results are qualitatively the same, but trends and differences are greater, most likely reflecting greater motivation of lab participants. Lab participants were on average more attentive than the AMT sample, as judged by attention questionnaire answers, 3.5 correct answers out of 4 in comparison to an average of 3 for the AMT sample. This seems reflected in more participants in the FMAB condition engaging in function learning, and a larger proportion of participants is classified as belonging to Function learner group (73% in comparison to 55% in Experiment 1A). Overall, differences in raw choice performance between the FMAB and MAB condition are greater than in the AMT sample (panel A in Figure 3.C.3), and the difference in exploration patterns is correspondingly stronger ($D_H = .231$, $95\%CI = [.203, .302]$) than in the AMT sample ($D_H = .131$, $95\%CI = [.114, .175]$). In Experiment 1B, the FMAB group tried significantly fewer alternatives, 13.70 ($SD = 5.03$) than the MAB group, 16.58 ($SD = 3.49$), $Z = 913$, $p = .012$ (one-tailed Wilcoxon rank-sum test), which was not the case in the AMT sample.

3.7 Experiment 2A and AB: Are People Function Learners and the Hidden Dangers of Function Learning

It has been consistently shown that people learning functional mappings tend to place strong priors on positive linear relationships between the cue values and the to-be-predicted criterion value (Brehmer, 1974; Busemeyer et al., 1997; DeLosh et al., 1997; Kalish et al., 2004; Lucas et al., 2015; McDaniel & Busemeyer, 2005). Hence, it is likely that participants' priors corresponded exactly to the structure they encountered in the bandit task of Experiment 1. One could argue then that there was little function learning necessary, and the advantage of function learners might largely stem from them being able to use the prior knowledge that they brought into the experiment.

In Experiment 2A, we address these possible limitations by using a mixed linear reward function which is less likely to correspond to people's priors. By studying how people learn other functional forms we also validate our framework further. In this reward function one feature is positively correlated with rewards and one negatively, which would cause those Function learners with positive linear priors to allocate choices to lower value alternatives initially. Observing a shift in choice allocation from regions in the feature space indicated by the prior toward regions indicated by the mixed linear function would provide clearer evidence that people indeed engage in function learning. Simultaneously, this would be evidence of feature information working against the sophisticated learner that tries to learn the function. Priors can obviously speed up learning substantially, but if the structure of the environment does not correspond to the priors, feature information can impede the function learners in comparison to learners that ignore it.

If people use their knowledge of the function or priors to guide exploration, we should be able to identify cases in which their priors may lead them astray. Experiment 2B was designed to investigate such a scenario. To this end, we introduced a U-shaped quadratic mapping from feature-values to utility. We studied whether people acquire incorrect beliefs about relationships between features and rewards due to interactions between function learning and decision processes. The Bayesian optimization framework allows for this possibility – due to simultaneous function learning and function optimization, what are thought to be less rewarding regions of the function space are known less well. Depending on the prior and the properties of the actual reward function, one could end ex-

periencing only alternatives from a part of the function space. Such systematic bias in the sample can significantly affect the accuracy of beliefs about the structure of the world – the true function underlying the rewards. For the quadratic function, we predicted that most people would end the task believing incorrectly that the rewards are a positive linear function of the features, and a smaller portion believing the function is negative linear; the tip being balanced in favor of positive one due to peoples’ priors. This is related to the “hot-stove” effect and adaptive sampling ideas in featureless multi-armed bandit problems (Denrell, 2007; Denrell & Le Mens, 2007; Denrell & March, 2001). With the quadratic function and specially designed FK task items we aimed to detect such locked-in effects where people end up with incorrect beliefs about the world and reinforcing them with subsequent choices.

3.7.1 Method

Participants and experimental design

In total, 190 people (79 females), in the age range from 18 to 69 ($M = 36.4$, $SD = 11.7$) participated in Experiment 2A. In Experiment 2B, 186 people (97 females), in the age range from 20 to 74 ($M = 35.9$, $SD = 11.0$) participated. We recruited participants through AMT and we checked the data quality following the procedures described in Appendix 3.A. As in Experiment 1, participants received a fixed show-up fee plus a performance-dependent bonus.

The core structure and visual design of the experiment was the same as in Experiment 1. We varied whether feature information was visible or not (FMAB vs. MAB task), In experiment 2A the function determining the rewards was a mixed linear while in experiment 2B it was a quadratic (identified by suffix “ml” and “q” in the condition name, respectively). Adding the two versions of Experiment 2 together, this yielded a 2×2 between-subject design and four experimental conditions: FMAB-ml, MAB-ml, FMAB-q and MAB-q.

Stimuli and Procedure

The bandit task in this experiment differed in the underlying function that generated the rewards. In the conditions of Experiment 2A we used a mixed linear function, where one weight was positive, and the other negative, $w_1 = 40$ and $w_2 = -30$, and the error term was drawn from $N(0, 6.25)$. Feature values were sampled from a uniform distribution, $U(0.1, 0.9)$. In

the conditions of Experiment 2B, we used a nonlinear function – rewards were determined by a U-shaped quadratic function:

$$R_k^t = 1 + 60(x_{1,k} - 0.02)^2 + 60(x_{2,k} - 0.02)^2 + 30x_{1,k}x_{2,k} + \epsilon_k^t.$$

The error term was the same as in the mixed linear function conditions, $N(0, 6.25)$, while the feature values were sampled from a uniform interval $U(-0.4, 0.4)$, to get the U-shaped form.

The FK task items for mixed linear conditions were constructed same as in Experiment 1, with modified feature value ranges to account for the differences in functional form. In the quadratic conditions we had different item types, designed to detect participants’ beliefs about the form of the function – whether they thought that it was positive linear, negative linear or U-shaped. More details about the design of the items in the task can be found in Appendix 3.B.

We followed exactly the same procedure as described in Experiment 1. Since the nature of the objective function governing the stimuli was never displayed, we were able to use the same instructions. We also recorded the same type of information for each participant.

3.7.2 Results and Discussion

Experiment 2A: Priors alone cannot explain the results

Figure 3.7 illustrates allocation of choices by the Bayesian models in the FMAB-ml condition of Experiment 2A. The version of the models we have used thus far does not incorporate any bias that would correspond to the positive linear expectations often found in experiments with humans. The leftmost and middle panel in the figure show that both the mean tracking model (BMT-UCB) and function learning based model (GP-UCB) have little difficulty identifying the good alternatives. Gaussian processes are very flexible however, and we do not necessarily need to use a zero mean function – we can easily make it a positive linear function of the features, which would correspond to the priors that can be inferred from human experiments. The rightmost panel illustrates choice allocation by the GPI-UCB model where we added a strong bias toward positive linear relationships. We see that in the indicative early trials it does allocate many choices to the upper right quadrant. This particular model is very fast in unlearning such a prior; the UCB choice rule, due to its usage of uncertainty, quickly starts trying out alternatives from other corners of the feature space and finds better alternatives. These patterns, however, vary with the exact pa-

parameter values and choice rules used. Clearly, the BMT-UCB model cannot have such systemic bias, as it ignores the features altogether.

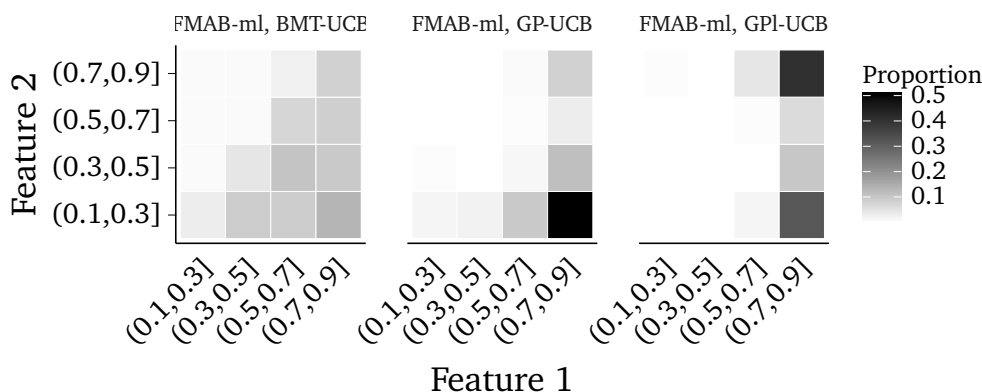


Figure 3.7 Exploration patterns of Bayesian models in Experiment 2A. The models were fitted to the same stimuli seen by the participants and then for each participant we simulated choices 33 times. Here we show the proportion of choices in the first 10 trials allocated to alternatives with feature values falling into one of the four bins indicated on the x-axis for Feature 1 ($w_1 = 40$) and the y-axis for Feature 2 ($w_2 = -30$). Leftmost, BMT-UCB cannot have any bias based on features, it quickly identifies the high rewarding alternatives. In the middle, GP-UCB has no initial bias (assumes zero mean reward) and as expected, it is even faster in detecting the good alternatives. Rightmost, if we add positive linear mean function to the GP model, akin to assumed people’s prior, we see divided allocation between upper and lower right corners. Not shown here, all three models concentrate their choices in the last 10 trials appropriately to higher rewarding, lower right region.

Participants’ choice performance is better than chance (rank of 10.5) and improves substantially over time (Figure 3.C.4A), similar to Experiment 1, and people’s performance is not as good as that of the Bayesian models. For the MAB-ml condition it shows a significant difference in mean choice rank between the first ($M = 8.25$, $SD = 2.47$) and last block ($M = 2.94$, $SD = 2.51$), $Z = 3734$, $p = 5.5E - 15$. Similarly, the difference between the first block performance ($M = 7.61$, $SD = 2.49$) and the last block ($M = 2.97$, $SD = 2.55$) is significant for FMAB-ml condition, $Z = 5132$, $p < 2.2E - 16$.

Figure 3.8 illustrates participants’ choice allocations in the first 10 and last 10 trials of the bandit task where rewards are governed by a mixed linear function. Firstly, participants in the FMAB condition indeed show clear signs of expecting a positive linear function. Many participants allocated a large proportion of their choices in the first 10 trials to alternatives with high feature values (Figure 3.8A). This further strengthens the belief that

this allocation was guided by feature information. In contrast, the MAB participants successfully avoided choosing alternatives from this region, as predicted by the BMT-UCB model. The Hellinger distance between the two distributions is .10, $95\%CI = [.094, .153]$, significantly different from 0 based on a permutation test, $p = .0002$. Direction was further validated with a one-tailed Wilcoxon rank-sum test. FMAB participants allocated on average 3.59 ($SD = 2.75$) choices (out of 10) to alternatives with both feature values between 0.5 and 0.9, significantly more than MAB participants (2.56, $SD = 2.03$), $Z = 5387$, $p = .008$. Secondly, FMAB participants with positive linear priors updated their beliefs and by the end of the task shifted their choices to the high rewarding alternatives defined by the true function in the environment (Figure 3.8B). This offers evidence that people do learn the function and do not simply use their prior knowledge. However, allocation of choices changes more gradually than predicted by the optimal GP-UCB model. This is likely due to a lower learning rate, or potentially using choice rules that do not use uncertainty, such as softmax (Luce, 1959; Sutton & Barto, 1998). Finally, according to the predictions, the MAB group sampled more alternatives on average, 14.86 ($SD = 5.05$), than the FMAB group, 13.53 ($SD = 5.09$), as supported by a one-tailed Wilcoxon rank-sum test, $Z = 5210.5$, $p = .027$.

The behavior of FMAB-ml participants is more nuanced however, as revealed by dividing the participants into clusters as in Experiment 1 (Figure 3.C.5). We find a small cluster ($N = 23$) with very good knowledge of the reward function (“Fast learners”), as indicated by their FK task results ($M = 1.23$, $SD = .17$), who showed very little evidence of an initial bias towards a positive linear function. A much larger cluster ($N = 79$) seems to have poor knowledge of the true underlying function ($M = 2.02$, $SD = .19$). However, they do not seem to have adopted the mean tracking strategy. This is the group of participants that exhibited a strong positive linear bias, and compared to the choice allocation of MAB participants, they are clearly guided by the feature information (this is why we refer to them as “Slow learners”). Figure 3.C.5B shows that they discover the good region by the end of the task, as if they engaged in function learning and updated their knowledge of the function; however, according to the FK task, they do not seem to have learned the true function adequately. Perhaps their prior beliefs were extremely strong and they choose accordingly in the FK task. Since our test items were designed to detect how well one knows the mixed linear function, it does a poor job of detecting other types of knowledge of the function. Another possibility is that this group used their prior at the beginning, but from then onwards used a mean tracking strategy to detect high rewarding alternatives – a mixture of the strategies we have consid-

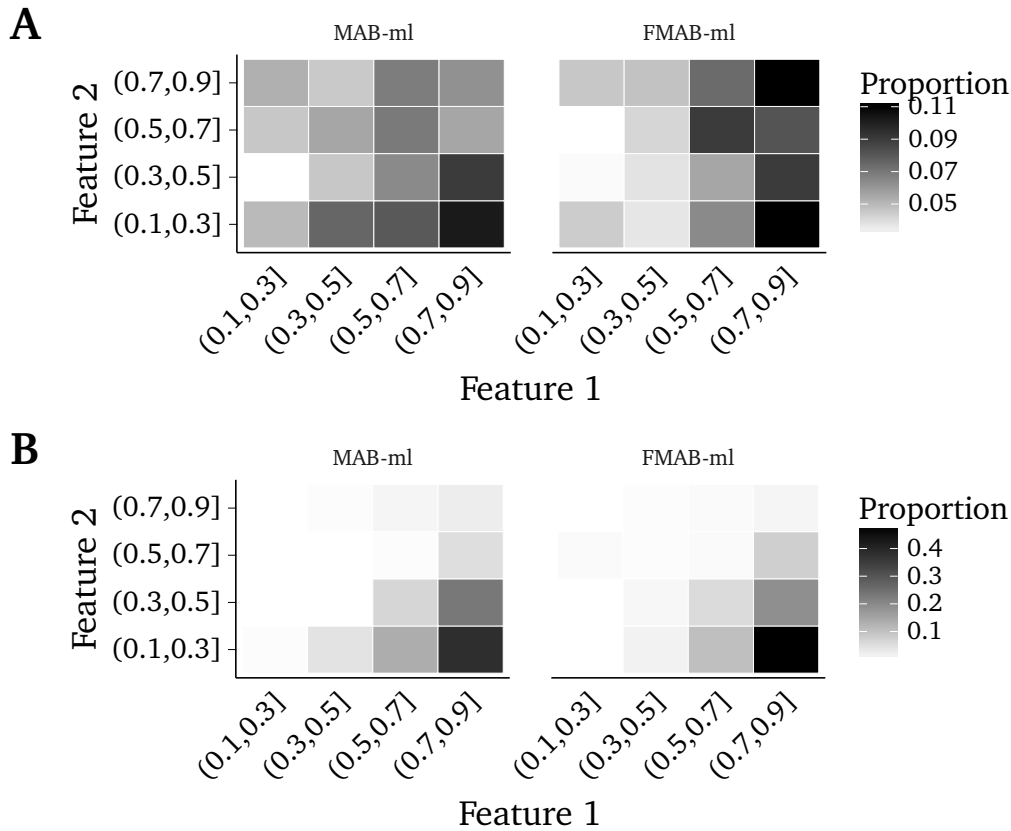


Figure 3.8 Exploration patterns of participants in Experiment 2A in the first 10 trials of the bandit tasks and the last 10 trials. (A) Proportion of all choices in first 10 trials allocated to alternatives with feature values falling into one of the four bins indicated on x-axis for Feature 1 ($w_1 = 40$) and y-axis for Feature 2 ($w_2 = -30$). Participants in the FMAB condition allocate many choices to the upper right quadrant, consistent with our assumption that people have expected positive linear relationships a priori. In contrast, participants in the MAB condition are quicker in identifying the high rewarding alternatives. (B) Analogous to the previous panel, but now showing allocation of choices for the last 10 trials in the bandit task. Participants in the FMAB condition now shifted their choices to the alternatives with feature values that lead to higher rewards, positioned in the lower right corner, where feature 2 has low values.

ered thus far. This points out that although the FK task is a useful tool for studying inter-individual differences it cannot easily discriminate what strategy people rely on – mean tracking or function learning. Nevertheless, it seems clear that there is large difference in the extent to which people are affected by a mismatch between the priors and the actual function governing the encountered environment.

The difference in choice allocations between MAB and FMAB participants in the first 10 trials is also evidence that feature information might not always be beneficial. Misaligned priors can guide people astray and deprive them of the benefits of function learning. In this case there is little difference in choice performance between the conditions. FMAB participants achieve mean rank across all 100 trials of 4.58 ($SD = 2.32$), while MAB participants reach 4.98 ($SD = 2.35$), and we cannot reject the null hypothesis of no difference ($Z = 3918$, $p = .1315$, two-tailed Wilcoxon rank-sum test).

Experiment 2B: Locked-in effect

Even though the quadratic reward function was more complex from the function learning perspective, participants' choices improved over time with a similar rate as in the other experiments (see Figure 3.C.4B). Over all 100 trials, there is no significant difference in choice performance, FMAB-q participants achieve mean rank of 5.28 ($SD = 2.73$), while MAB-q participants reach 5.57 ($SD = 2.89$), $Z = 4063$, $p = .5453$, according to two-tailed Wilcoxon rank-sum test. However, in terms of exploration the differences are substantial. First, there is a difference in terms of number of alternatives tried. The MAB-q group sampled more alternatives on average, 17.08 ($SD = 3.46$), than the FMAB-q group, 15.29 ($SD = 4.83$), as shown by one-tailed Wilcoxon rank-sum test, $Z = 5130.5$, $p = .0093$. Second, the Hellinger distance between choice allocations of MAB-q and FMAB-q groups in the first 10 trials amounts to $D_H = .098$ (95%CI = [.087, 0.146]), significant according to the permutation test, $p = .0025$. The FMAB-q group allocated substantially more choices (out of 10) to alternatives with high feature values ($M = 3.07$, $SD = 2.49$) with both feature values between 0.5 and 0.9, than MAB participants ($M = 2.36$, $SD = 1.78$), $Z = 4897.5$, $p = .044$ (one-tailed Wilcoxon rank-sum test).

The purpose of the quadratic reward function conditions was to illustrate how strong interactions between function learning and decision making processes can be, sometimes leading people to lock themselves into incorrect beliefs about the world. The reward function in the MAB-q and FMAB-q conditions is bowl-shaped, with high rewards toward the corners of the feature space where feature values are either very high or very low. Largest rewards were located where feature values are the lowest, and slightly smaller for alternatives with high values for both features. The quadratic function had steep slopes, such that small differences in feature values towards the middle of the feature space would decrease rewards sharply toward zero. With such a reward function, we predicted that peo-

ple that start sampling in one part of the feature space are likely to stay there, exploring only the local neighborhood. This could happen simply through a random first choice in a certain part of the feature space, or through prior beliefs about functional relationships. In either case, the result would be a tendency to remain in one corner of the feature space, potentially believing that the function is either positive or negative linear.

We designed special item types in the FK task for the quadratic reward function condition, with the aim of detecting whether people believe the functional relationship is positive linear, negative linear, or whether they realized it is a U-shaped quadratic function. For instance, in the “Min Local” item type, there is a dominating and middle alternative that have large feature values, while the dominated alternative is in the middle of the interval. A person with negative linear beliefs would choose mostly the least rewarding dominated alternative, while someone with a positive linear belief would mostly choose exactly the opposite, the dominating alternative. Similarly, patterns of choices should differ for other item types depending on the beliefs. We computed distances between mean choice on each item type and choices a person with highly certain positive, negative or quadratic function beliefs would make, and classified people as having those beliefs to which the distance was the smallest. Item types and classification is explained in more detail in Appendix 3.B, while an overview of choices in the FK task, broken down into item types and beliefs, can be found in Figure 3.C.6B. As expected, positive linear beliefs are more widespread ($N = 69$) than negative ones ($N = 32$). There is only one person classified as believing the reward function is quadratic.¹⁴ Choice patterns are also somewhat clearer for the positive belief group. Although people did not learn the true underlying function, the fact that they exhibit these complex choice patterns demonstrates that they nevertheless engaged in function learning during the bandit task.

Figure 3.9 shows the allocation of choices over all 100 trials for the FMAB-q condition, broken down according to the type of belief participants had, as indicated by a classification based on the FK task performance. The left panel shows that participants who believe that the reward function is positive linear (“Positive”, $N = 69$) allocate a large proportion of choices to alternatives with high feature values. The right panel shows an analogous result for participants with negative reward function beliefs (“Negative”, $N = 32$), who choose predominantly those alterna-

¹⁴We do not show choices for the quadratic participant due to a single datum. Our FK task can only serve as a rough measure of people’s beliefs, the number of people that detected the quadratic relationship is probably larger.

tives with low feature values. Some people did try out alternatives from other corners, where other high rewarding alternatives were positioned, but seemingly not enough to develop an accurate representation of the reward function. These choice patterns are consistent with our prediction of a strong interaction between beliefs about the functional relationship and how the alternatives that people choose are positioned in feature space. We validated these observations with statistical tests on number of choices allocated to upper right and lower left parts of the feature space between Positive and Negative believers. The Positive group allocated substantially more choices (out of 100) to alternatives with high feature values ($M = 44.3$, $SD = 37.5$) with both feature values between 0.5 and 0.9, than Negative believers ($M = 18.9$, $SD = 18.0$), $Z = 1462.5$, $p = .0044$ (one-tailed Wilcoxon rank-sum test). The pattern is opposite of course for allocation to alternatives with feature values between 0.1 and 0.5. Negative believers allocated more choices in this region ($M = 51.7$, $SD = 30.2$), than Positive believers ($M = 24.7$, $SD = 29.8$), $Z = 1630$, $p = 6.1E - 5$ (one-tailed Wilcoxon rank-sum test).

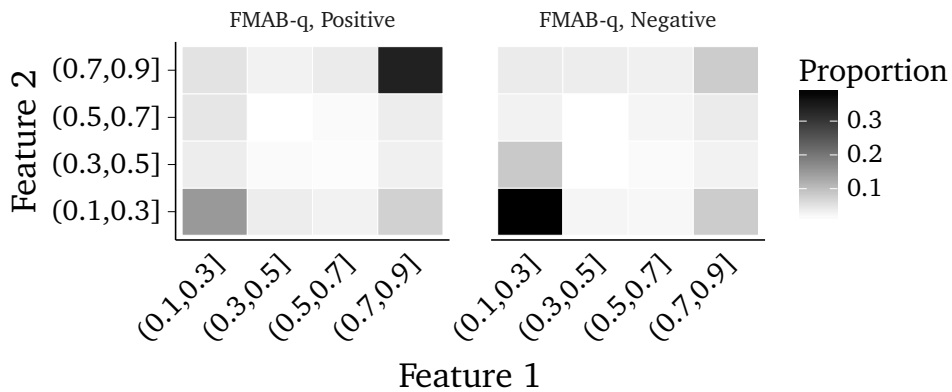


Figure 3.9 Allocation of choices in the feature space by participants in the FMAB condition in Experiment 2B broken down according to the type of beliefs participants as determined from FK task performance. Participants classified as having positive linear function beliefs (“Positive”) spent most of the 100 trials choosing alternatives with relatively high feature values. Participants with negative linear function beliefs (“Negative”) spent most time in the lower left quadrant, sampling alternatives with relatively low feature values. Most of the time they did not venture to the opposite corner, hence having little opportunity to realize that the function is actually nonlinear.

It is important to note that the locked-in phenomenon is not necessarily the result of irrational behavior, but a consequence of simultaneous

function learning and function optimization; what are thought to be less rewarding regions of the function space are known less well because exploration of these regions is thought too costly in terms of missing out on good rewards. If the function is steep in those parts of the function, as was the case in our quadratic function, one might severely mis-estimate the reward function, and miss out on substantial rewards. The Bayesian models develop similar blind spots, albeit less pronounced. In Figure 3.C.6A we illustrated performance of the \mathcal{GP} model with two very different choice rules – Maximum Variance (MV) and UCB. The MV chooses alternatives to reduce the uncertainty around the estimated function as much as possible, i.e. it tries to learn the function as best as it can. The figure shows it succeeds in this endeavor, having almost perfect scores in the FK task. Using our belief classification scheme, all simulated participants would be classified as having quadratic beliefs. The price this rule pays for such knowledge is a low amount of cumulative rewards collected.¹⁵ The MV performance provides a stark contrast to the UCB choice rule. The UCB rule tries to balance exploration and exploitation by taking into account both the mean and variance of predicted rewards. Once it arrives in a high rewarding region from an initial point partly determined by random choice, it stays there. This happens because of the steep slope of the quadratic function, such that uncertainty in other corners can not offset the high expected rewards in the currently favored one. This is validated by analyzing its choices in the FK task. We classified its choices into the three belief types and found it dominantly adopts negative linear beliefs ($N = 63$), then positive ($N = 32$) and in only few cases it learns the true nature of the function ($N = 6$). Overall, it is slightly better in detecting the global maximum, but in terms of knowledge about the world it remains as knowledgeable as humans.¹⁶

¹⁵As it happens, our quadratic function had largest rewards in the corners of the feature space, which coincidentally is where most of the uncertainty is as well for the GP model. By spending lot of trials in these regions MV rule actually gained large amount of rewards as well. This is usually not the case however; in positive and mixed linear environment its choice performance was close to random (see mean ranks of the models in Appendix 3.D).

¹⁶We can tip the balance toward positive linear beliefs by implementing a small bias into the mean function of the GP model, with positive slopes on features. With such modification, the GPI-UCB model dominantly ends up having positive linear beliefs.

3.8 Experiment 3: Different Flavors of Exploration and Factors Affecting Exploration

When people explore in the FMAB task, do they explore new alternatives at random, or are they more sophisticated about it, choosing those alternatives that are more informative about learning the function? The first goal of this experiment was to examine this question in detail.

This question can be mapped to type of decision strategy – UCB exploration is skewed toward the more informative alternatives, that is, those that would lead to more information about the underlying reward function,¹⁷ in contrast to simpler rules like Softmax (SM, Luce, 1959; Sutton & Barto, 1998) that are a probabilistic function of mean rewards only. In the classical MAB problems most studies found support for simple decision noise, embodied in the SM rule (e.g. Daw et al. 2006; but see Speekenbrink and Konstantinidis 2015). These findings are disappointing, considering that rationally, uncertainty should play a prominent role in exploration (as in Gittins indices, Whittle, 1980). However, in the FMAB problem information seeking is more valuable as knowledge obtained about the function generalizes to other alternatives as well. Hence, we expected that people are likely to be driven by uncertainty when exploring in the FMAB problems.

In examining this question, studies thus far relied mostly on modeling evidence (e.g., Daw et al., 2006, but see R. C. Wilson et al., 2014). Our approach was to obtain more direct behavioral evidence. One factor in the experiment determined whether participants do or do not have knowledge of the function before the FMAB task starts (fFMAB conditions and FMAB conditions, respectively). In the beginning of the experiment, we introduced a function learning task, inspired by the literature on multiple cue probability learning (Brehmer, 1974; Busemeyer et al., 1997; Hammond, 1955; Speekenbrink & Shanks, 2010). The participants encountered one alternative at a time, and had to predict its true value. In the bandit part of the study, we added alternative-specific intercept terms to the reward functions. In other words, the function learned did not completely determine the rewards and once these participants got to the FMAB task, they still had

¹⁷In the machine learning community, an approach where agents take an active role in deciding which samples to obtain is called “active learning” (e.g. Cohn, Atlas, & Ladner, 1994); see also a recent study by Markant, Settles, and Gureckis (2015) for an application in category learning. However, this term is usually used in the context where an agent would be interested in maximally informative alternatives, while in reinforcement learning the problem differs as an agent is also interested in obtaining high rewards.

an incentive to explore, but importantly, not in order to learn the function, but to learn the alternative-specific intercepts instead.¹⁸ In the FMAB task they would largely know the function already – their exploration should be constrained to a high rewarding region and be driven mostly by mean estimated rewards and decision noise. In contrast, exploration of the participants that complete the FMAB task without going through the function learning task should be more guided by uncertainty about the function. As a result, FMAB participants will in general explore more – try more alternatives, and more specifically allocate more choices to alternatives from lower rewarding regions in order to learn the function better.

Our second goal was to examine some of the factors affecting how much people explore in the FMAB problem, and consequently how well they learn the function. If we would like to improve people’s choice performance, it would be beneficial to understand how they react to factors that theoretically should affect the level of exploration. To this end, we examined two questions: how does the amount of uncertainty about the function and horizon length influence behavior. In the first, the larger the experienced uncertainty the more exploration there should be. For example, nonlinear functions are more difficult and require more samples to learn (Brehmer, 1974). People should be more uncertain about the function when facing such relatively difficult situations. Hence, another factor we varied in the experiment was the type of function, participants faced using either a positive linear or a quadratic reward function (FMAB-pl, fFMAB-pl, and FMAB-q, fFMAB-q conditions). Our prediction here was that the participants learning about the quadratic function would still be largely guided by uncertainty, even after the function learning task. Thus, the difference in exploration patterns in the beginning trials of the bandit task between the FMAB-q and fFMAB-q condition will be much smaller than between the FMAB-pl and fFMAB-pl.

The other factor we examined is horizon length. The UCB choice rule is boundedly rational in the sense that it does not use information about the horizon, that is, the number of trials left till the end of the bandit task. A Bayes optimal policy, such as based on Gittins indices in the Bernoulli MAB problem (Whittle, 1980), would dynamically decrease the tendency to explore and exploit more as the end of the game draws near. In most

¹⁸So far we have assumed that the function completely determines the rewards, even if probabilistically; the experienced uncertainty is then observational noise. In Experiment 3 part of the error term is also due to unobserved systematic factors (i.e. the intercepts). In real-life problems decision-makers tend to face both forms of uncertainty. Note that if all the uncertainty was due to the random intercepts the problem would almost reduce to an ordered search problem (Analytis, Kothiyal, & Katsikopoulos, 2014; Weitzman, 1979).

cases, computing optimal policies is an intractable problem, and in practice heuristic rules like UCB are commonly used instead (but see Guez et al., 2012, 2014, for approximations that do take the horizon into account). In all our experiments participants could see the number of trials left in the bandit task and could use that information in balancing exploration and exploitation. A recent study by R. C. Wilson et al. (2014) showed that humans are sensitive to information about the horizon and change their exploration in the direction prescribed by the optimal models. This was done in the classic MAB task and we were interested whether people exhibit such a close to optimal reaction to horizon information in the more complex FMAB setting. This made the final factor we manipulated in the experiment, either a 30-trial (fFMAB-pls and fFMAB-qs) or 100-trial FMAB task (fFMAB-pl and fFMAB-q). Our main prediction was that in comparison to the first 30 trials of the 100-trial-long horizon conditions, participants in the 30-trial-long conditions will try out fewer alternatives and concentrate their search more in the regions they believe are highly rewarding.

3.8.1 Method

Participants and experimental design

In total, 431 participants (207 females), in the age range from 18 to 74 ($M = 34.9$, $SD = 11.2$), took part in Experiment 3 (see Table 3.1 for more details). Participants were recruited through AMT and data quality was checked following the procedures described in Appendix 3.A. Participants received a fixed minimal payment plus a performance dependent bonus.

In this experiment, participants in all conditions completed the FMAB task. The main factor that we varied was whether participants first completed a function learning (FL) task (FL task vs. no FL task, identified by prefix “f” or no prefix in the condition name, respectively). We varied the type of reward function determining the rewards (positive linear vs. quadratic, identified by suffix “pl” and “q” in the condition name, respectively), and we partially varied the number of trials in the FMAB task (30 trials vs. 100 trials, implemented only for conditions with FL task, identified by additional suffix “s” or no suffix in the condition name, respectively), yielding the total of six between-subject conditions: FMAB-pl, fFMAB-pl, fFMAB-qs, FMAB-q, fFMAB-q, fFMAB-qs.

Stimuli


In four of the conditions participants first completed the FL task where they could learn the function more directly, by predicting the reward of a single alternative instead of choosing between several of them. Participants completed 100 trials where in each trial we presented them with an alternative, looking exactly the same as in the bandit tasks (see Figure 3.10). We asked them to make a prediction about the amount of reward they would get from the presented item and after they made a prediction we gave them the information about the actual reward. The rewards were governed by the functions described in previous experiment, with a few small changes. In conditions with the positive linear function, fFMAB-pl and fFMAB-pls, the parameters of the function were the same as in Experiment 1B, while in conditions with the quadratic function, fFMAB-q and fFMAB-qs, the parameters of the function were the same as in the quadratic conditions in Experiment 2. However, in this experiment we added an alternative-specific intercept terms to both functions drawn from a Normal distribution, $\mu_k \sim N(0, 9)$. The variance of the usual error term was reduced to 4: $\epsilon_k^t \sim N(0, 4)$, to compensate for the additional noise due to the random intercept term. The intercept term was included to provide an incentive for exploration in the contextual bandit task, even if they know the observable part of the function perfectly. In each trial participants encountered a new alternative, with feature values, intercept and error values drawn randomly from their respective distributions. Since they saw the alternatives only once, they could not know how much of the experienced error could be attributed to the intercept. However, once they would come to bandit task and sample the same alternative multiple times they could quickly realize that part of the variance is systematic.

Stimuli in the FMAB tasks were constructed in the same manner as in previous experiments, with the difference that the rewards were determined with functions described above in the FL task, with an additional intercept term. Note that in the bandit task alternatives stay the same in every trial, so that the alternative-specific intercept term was drawn once at the beginning of the task and remained the same throughout.

We constructed items in the functional knowledge task items in the same way as in previous experiments, but using the functions specific to this experiment. Alternative-specific intercepts were left out when generating stimuli for the FK task. More details about the design of the items is available in Appendix 3.B.

Total number of rounds: 100
Current round: 1

Running total: -7.7



Estimate:

7.3

Enter your estimate of the object value and then CLICK on a square to submit your estimate.
You will receive information about its true value after that.
Press C key to continue to the next round.

Figure 3.10 In the function learning (FL) task participants predicted the amount of reward a single item would yield and got information about it after they made a prediction. We incentivized them to make as good predictions as possible, and consequently learn the function, by making their payoff dependent on their accuracy.

Procedure

The procedure in this experiment was slightly different for conditions that had the FL task before the FMAB and FK tasks. We instructed participants in these conditions, after they accepted the consent form and completed the sociodemographic questionnaire, that they would be presented with a single object and their task was to predict the value of the object. We explained that their payoff in the experiment would depend on their accuracy in this task: ten experimental points minus the absolute difference between their prediction and actual value. We illustrated the formula in several examples. The goal of the task was to win as many experimental points as possible, which would later be converted to money according to the advertised exchange rate. We told them that they would complete two other tasks afterwards, without specifying the details. Instructions for other tasks were kept the same as in previous experiments. In this experiment, we additionally collected information about participants' predicted values and time they took to make a prediction in each trial.

3.8.2 Results and Discussion

As in the previous experiments, participants' choice performance is better than chance (rank of 10.5) from the outset and improves substantially over time (Figure 3.C.7). In this experiment we were interested in dif-

ferences in choice allocation in the first 10 trials between conditions with and without function learning pre-training. Our prediction was that people will use uncertainty when learning the function in the FMAB task and we should observe a difference in exploration patterns between people who know the function well and those who only began learning it. Figure 3.11A shows that the difference between conditions with the positive linear reward function is sizable. Participants that already learned about the function through the FL task allocate more choices to alternatives with feature values associated with high rewards (upper right region).¹⁹ Hence, people in the FMAB-pl condition allocate more early choices to alternatives with low feature values in an attempt to learn the function better. We validated the observed differences with a permutation-based statistical test, which showed a significant difference between the two distributions, $D_H = .185$, $95\%CI = [.164, .232]$, $p < 1E - 07$. Moreover, a one-tailed Wilcoxon rank-sum test shows that the fFMAB-pl group allocated on average 6.07 ($SD = 3.24$) choices (out of 10) to alternatives with both feature values between 0.5 and 0.9, significantly more than FMAB-pl group (4.34, $SD = 2.54$), $Z = 4278$, $p = .0001$. Finally, FMAB-pl group should have tried more alternatives altogether than the fFMAB-pl group. This is also the case: FMAB-pl group tried on average 14.06 ($SD = 5.00$) alternatives (out of 20), significantly more than fFMAB-pl group (9.97, $SD = 5.87$), $Z = 4448$, $p = 9.3E - 06$.

Interestingly, in the fFMAB-pl condition there is a small spike in choosing the alternatives with lowest feature values (lower left corner). An optimal test if one entertains a hypothesis that the function is linear is to try alternatives from all four corners. Potentially, participants wanted to verify that it is still the same function as in the FL task, despite being instructed so.

A quadratic function is more difficult to learn than a positive linear function (Brehmer, 1974; Busemeyer et al., 1997). As a consequence, participants in the fFMAB-q condition should have larger uncertainty about the function after the FL task than those in the fFMAB-pl condition. The fFMAB-q group had more difficulties learning the function in the FL task, their mean MAD between predictions and observed rewards in the last 20 trials was 6.41 ($SD = 1.54$), significantly larger than for the fFMAB-

¹⁹fFMAB-pl participants perform reasonably well in the FL task. Most of the learning occurs in the first 20 trials where mean absolute deviation between their predictions and observed values was 6.38 ($SD = 1.80$) and this decreases further to 5.54 ($SD = 1.43$) in the last 20 trials ($Z = 1847$, $p = .0002$, one-tailed Wilcoxon signed-rank test). The mean correlation also increased from 0.44 to 0.54. Given that the total standard deviation of the error term in the task was equal to 5, this is a good performance.

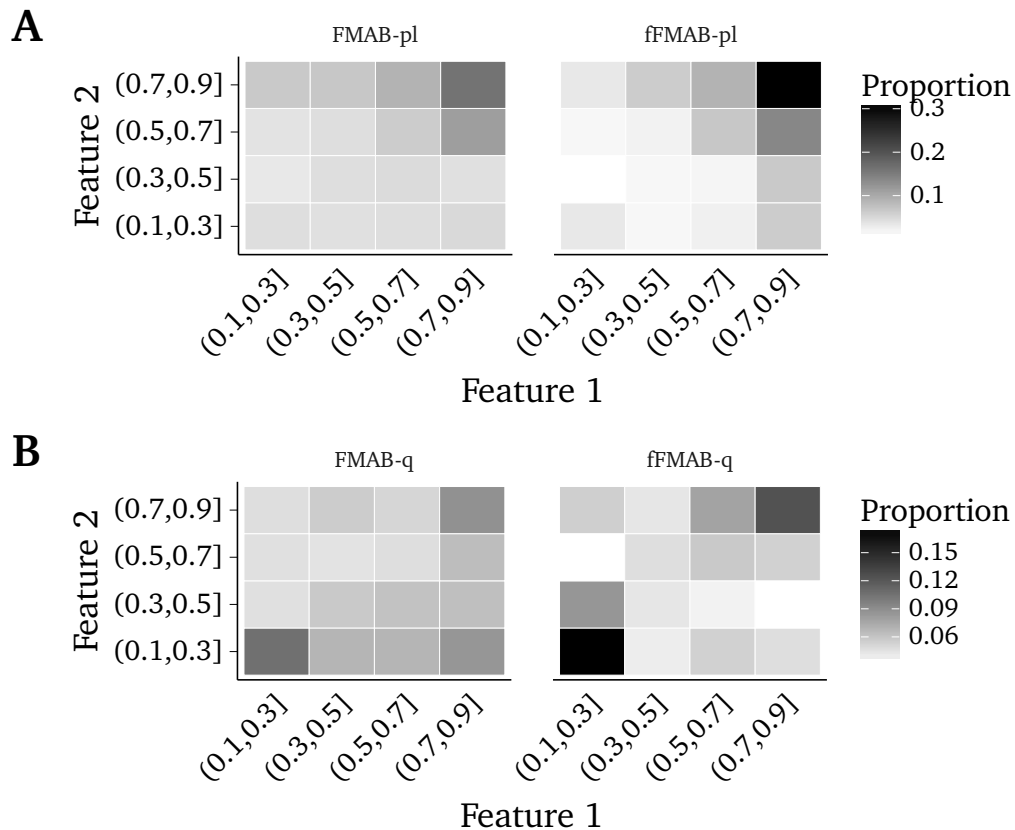


Figure 3.11 Differences in the exploration patterns in Experiment 3 between participants that completed a function learning task before the FMAB task (fFMAB conditions) and those that went directly to the FMAB task (FMAB conditions). (A) Proportion of all choices in the first 10 trials allocated by participants in FMAB-pl and fFMAB-pl conditions to alternatives with feature values falling into one of the four bins indicated on x-axis for Feature 1 and y-axis for Feature 2. Participants with function learning pretraining (fFMAB-pl) allocate many more choices to the alternatives in the high rewarding region (upper right corner). Their exploration should be mostly guided by decision noise, while FMAB-pl participants should be still driven with uncertainty. The difference is evidence that uncertainty plays an important role in function learning. (B) Analogous to the previous panel, but now showing difference in choice allocations between participants in conditions with quadratic reward function (FMAB-q and fFMAB-q). The difference is smaller now, showing evidence for uncertainty about function affecting people’s tendency to explore.

pl group, $Z = 3276$, $p = .0003$ (two-tailed Wilcoxon rank-sum test). Given the equal horizon in the FMAB task, participants facing the quadratic reward function should therefore be more exploratory. Hence, the differ-

ence in exploration in the first 10 trials of the FMAB task between FMAB-q and fFMAB-q should be smaller than the difference between FMAB-pl and fFMAB-pl (see Figure 3.11B). Although there is a difference in Hellinger distances between distributions of choices, .151 and .185, respectively, the confidence interval of the difference, .034, is quite broad, $CI = [-.016, .08]$ (two-tailed 95-percentile bootstrap), indicating that the observed difference is not significant. This goes to some extent against the results from our previous analysis, that people would use uncertainty in their exploration. However, this result also depends on few other assumptions – differences in how well the functions are learned, as well as exact functional forms. A clearer test would involve using the same function but with smaller number of trials in the FL task, and this investigation is left for future studies.

Even though UCB choice rule is sophisticated – it takes into account the uncertainty about the function as well as its mean, it does not take into account the information about the horizon. Do we have evidence that people behave more optimally than what our suboptimal UCB-based framework would suggest? In the final analysis we examine the differences in exploration patterns between pretraining conditions where participants continued to our standard 100-trial FMAB task and conditions where they completed a shorter 30-trial long FMAB task. Keep in mind that participants always had the information about the remaining number of trials available on the screen.

Figure 3.12 illustrates choice allocations in feature space for the relevant conditions.²⁰ Shorter horizons in fFMAB-pls and fFMAB-qs conditions should have led to more exploitation of functional knowledge, but the figure shows this is not the case. In the conditions with a positive linear reward function (Figure 3.12A), the distance between choice allocations is significant according to our permutation test, $D_H = .072$, $CI = [.064, .106]$, $p = .0008$. However, the difference goes in the opposite direction from our prediction: the group with the shorter horizon (fFMAB-pls) is more exploratory than the group with the longer horizon. There was also no significant difference in choice allocations to high rewarding alternatives with feature values from 0.5 to 0.9, $Z = 1872$, $p = 0.668$ (one-tailed Wilcoxon rank-sum test). For the quadratic reward function conditions we find that the difference is not significant, $D_H = .101$, $CI = [.091, .163]$, $p = .0537$. Moreover, differences in exploration tendency should be reflected in the number of alternatives tried in the first 30 trials. Consistent with our re-

²⁰In this analysis we look at all 30 trials instead of first 10 trials as elsewhere in the article. Since our prediction was that participants in the short horizon conditions will start exploiting much more quickly than in the long horizon conditions, it is reasonable to examine the final trials of the short conditions as well.

sults above, in the linear conditions the number of alternatives tried in the longer horizon (8.61, $SD = 5.19$) is not significantly greater than that in the condition with the shorter horizon (9.87, $SD = 4.55$), $Z = 1595$, $p = 0.964$ (one-tailed Wilcoxon rank-sum test). We get the same result for the quadratic conditions: participants do not try significantly more alternatives in the longer horizon (10.68, $SD = 5.80$) than in the shorter horizon (11.00, $SD = 5.70$), $Z = 1847$, $p = 0.664$ (one-tailed Wilcoxon rank-sum test). If anything, in shorter horizons participants try slightly more alternatives on average. Finally, we can examine the differences in mean choice ranks. In short horizons there might not be enough time to explore the intercepts and relying too much on functional knowledge from the FL task would lead to some loss in choice performance. We can see from Figure 3.C.7 that the choice performance of the fFMAB-pls group mirrors that of the fFMAB-pl group, with a mean rank of 5.41 ($SD = 2.85$) and 4.97 ($SD = 3.07$) in the first 30 trials, respectively. The difference is not significant, $Z = 1753$, $p = 0.309$ (two-tailed Wilcoxon rank-sum test). In the quadratic conditions, the fFMAB-qs group performed worse than the fFMAB-q in the last two blocks, but overall same result holds. Although mean rank achieved by the fFMAB-qs group was slightly worse ($M = 6.74$, $SD = 3.14$), than that of the fFMAB-q group ($M = 5.86$, $SD = 2.85$), the difference is not significant, $Z = 1626$, $p = 0.129$ (two-tailed Wilcoxon rank-sum test). In summary, even though some studies find evidence that people appropriately change their exploration strategy in response to horizon in classical MAB tasks (R. C. Wilson et al., 2014), in our FMAB tasks we find very little evidence that people take into account such information. This suggests that heuristic choice strategies like UCB provide a good enough description of participants' behavior.

3.9 General Discussion

Human learning of functional relations and learning to choose rewarding actions have been studied in isolation thus far. We argued that in many situations these two learning processes interact and that it is thus important to study how people simultaneously learn reward functions and make decisions. This natural extension of previous research on reinforcement learning and function learning is relevant for both theoretical and practical reasons. Theoretically, time and tried reinforcement learning models, such as Temporal Difference learning models (Sutton, 1988; Sutton & Barto, 1998), cannot explain how humans and animals learn in situations characterized by realistic high-dimensional stimuli. And practically, it is difficult

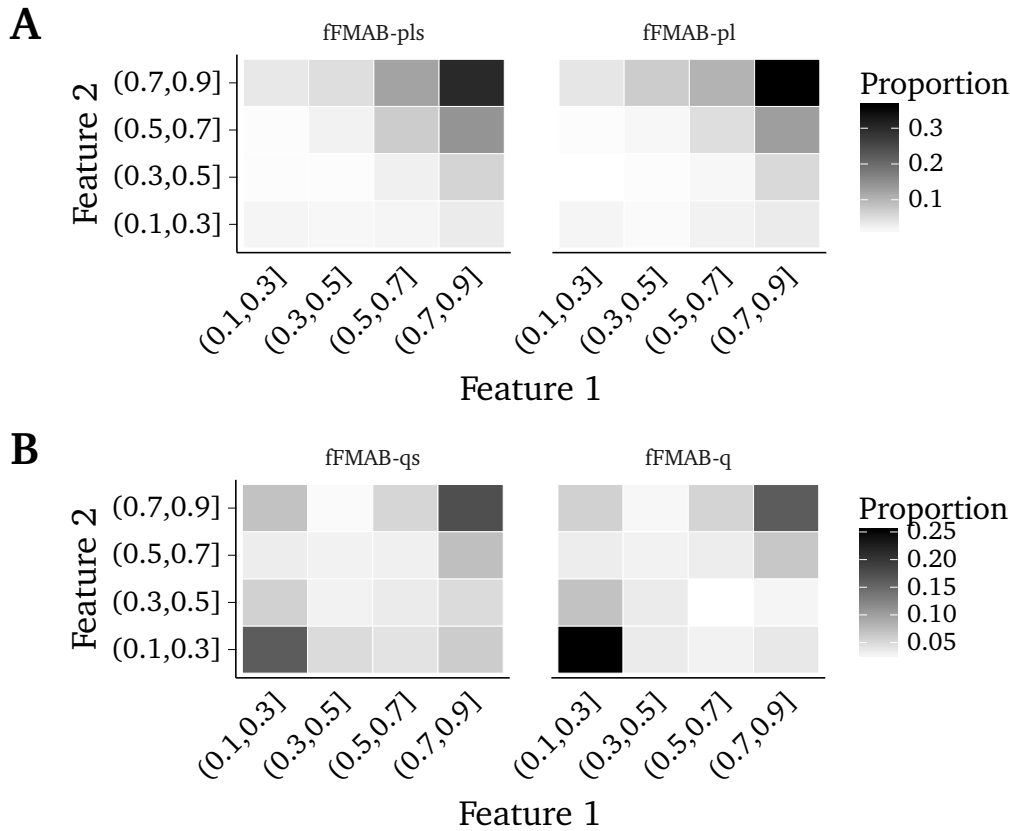


Figure 3.12 Differences in exploration patterns in Experiment 3 between participants that completed a function learning task (fFMAB conditions) and either a short horizon FMAB task (30 trials) or a long horizon FMAB task (100 trials). (A) Proportion of all choices allocated by participants in fFMAB-pls (short horizon) and fFMAB-pl (long horizon) conditions to alternatives with feature values falling into one of the four bins indicated on x-axis for Feature 1 and y-axis for Feature 2. Participants that did a short horizon version explore the choice set in first 30 trials in a similar way as a long horizon group, suggesting that information about the horizon did not affect their choice strategy. (B) Analogous to the previous panel, but now showing difference in choice allocations between participants in conditions with quadratic reward function (fFMAB-qs and fFMAB-q). As for the positive linear function, the exploration patterns of both groups are very similar.

to generalize findings from these two separate strands of literature to real-life knowledge of concepts and decision making.

We formalized the problem of simultaneous learning of reward functions and decision making as a contextual multi-armed bandit problem (Auer, 2002). In particular, we focused on a version with alternative-specific features only – the feature-based multi-armed bandit task. We

proposed Bayesian optimization (Shahriari et al., 2016) as a framework for understanding the interaction between function learning and decision making processes. We showed that our FMAB task can be solved in two ways – by learning the function, or ignoring the features and learning mean rewards instead. The key prediction is that, in comparison to mean trackers, function learners should benefit substantially in the early stages of the task, as they can generalize their initial knowledge to other alternatives in the choice set and avoid trying poor alternatives. The goal of maximizing rewards will lead to biased sampling of relatively high rewarding alternatives, which can result in biased functional knowledge. Finally, we predicted that uncertainty will play an important role in function learners’ exploration of the alternatives.

In a series of experiments we found broad support for a number of novel predictions. We find evidence that: (1) people’s exploration is guided by features and the functions they learn, (2) there are strong inter-individual differences, some people take the function learning approach, others ignore the features and learn only mean rewards, (3) paying attention to context can hurt if prior expectations about the reward functions do not correspond to the actual functions encountered in the environment, which can result in a locked-in effect and maintenance of incorrect beliefs, (4) people explore alternatives specifically to learn the function better, suggesting that they wisely use uncertainty about the function when deciding how to explore the available alternatives, and (5) time horizon does not affect people’s choices, suggesting that heuristic decision strategies that incorporate uncertainty in a myopic way might describe the behavior sufficiently well.

3.9.1 Implications

We take the opportunity to illustrate the fertility of the research program that we have proposed in this work. We highlight several exciting lines of future research, as well as implications for several psychological phenomena.

An obvious line of inquiry is examining transfer of knowledge *within* a function. The advantage of decision strategies powered by function learning is their ability to generalize. Hence, they shine when novel alternatives enter the choice set (e.g., a new restaurant opens in the neighborhood). Novel alternatives here are of the same kind, determined by the same function (hence within). Generalization per se is not of primary interest here, we know already that people are able to do that from function learning literature (Busemeyer et al., 1997; DeLosh et al., 1997; Kalish et al., 2004;

McDaniel & Busemeyer, 2005; Speekenbrink & Shanks, 2010). Examining people's behavior toward novel alternatives would be a test of the framework and that people indeed do learn the functions while choosing. In some sense our FK task was exactly that. These items were designed with a different goal in mind, but more importantly, situations where novel alternatives enter the bandit task instead would be more interesting for model evaluation (for related research using the standard MAB paradigm, see Le Mens, Kareev, & Avrahami, 2016). Features of novel items can be designed so that the predicted probability of choosing them differs starkly between function learners and mean trackers. A mean tracker which also tracks the general mean of a set of alternatives, as used by Gershman and Niv (2015) to explain neophilia and neophobia, would also give differing predictions. The Bayesian optimization framework also predicts larger uncertainty in the feature space where rewards are relatively low. Hence, another test of the framework would involve designing novel alternatives dynamically based on people's experience in the bandit task and compare predictions of such alternatives with those coming from part of the feature space with high rewards.

A more general question is how people transfer their knowledge *between* one context (e.g., Italian restaurants) and another (e.g., Japanese restaurants). That is, how do people generalize from one function to another? Benefits of generalization can extend beyond extrapolating within the same decision situation. Rather than learning the function from scratch, an agent facing a new decision situation could benefit from transferring her experience with similar situations, effectively generalizing from one function to another. Such more general transfer of learning cannot be captured by the current framework and this is an exciting venue for further theoretical work. One way to formalize this idea is to implement an analogical similarity measure in the space of functions (see, for example, Gentner, 1983; Gentner & Markman, 1997). Psychologically, the implication of this proposition is that there is no true novelty – there is no truly new decision situation, we can always reuse our experience in similar situations.

Following this it is easy to see that a lot can be said about preference learning. This research proposes that the basic unit of preferences are functions, i.e. knowledge of the functions drive our preferences. How well the reward functions are learned depends on one's collection of experiences. These experiences are biased by the reward landscapes, priors, and the initial alternatives tried. Such factors could easily result in heterogeneous preferences in a population of decision makers. Consider our quadratic function from Experiment 2B. The reward function of certain products could have such a nonlinear form. As shown in the experiment, due to pri-

ors or random samples at the beginning, one might end up knowing only one part of the space and having very poor estimates for the other part (see, for related explanations based on standard MAB paradigm, Denrell, 2005; Denrell & Le Mens, 2007; Fiedler & Juslin, 2006; Le Mens & Denrell, 2011). This could account for polarization of preferences, at least partly – for example, one group of consumers favoring soft cheeses, and the other hard cheeses. Another prediction of the framework would be that breadth of choice has a beneficial influence – the more choice there is the larger the chances that our sample of experiences is larger and the reward function estimated better. Large product categories would then help people navigate to the best alternatives, potentially leading to more unequal market shares of the products. Our framework could give a novel perspective on the currently mixed evidence for choice-overload effects (see for a recent meta-analysis Scheibehenne, Greifeneder, & Todd, 2010). Our dynamic, function driven take on preferences would be somewhat controversial. For example, it goes against a firm tenet in economics, whereby preferences are fixed (Mas-Colell, Whinston, & Green, 1995). Since choices rely on an inference process, where people infer rewards from features and context, our proposition also confronts some findings that maintain that “preferences need no inferences” (Zajonc, 1980).

In optimal solutions to standard MAB problems, such as Gittins indices (Whittle, 1980), uncertainty plays a key role in exploration of the choice set and identifying the optimal choice as fast possible. Disappointingly, research thus far has found little support for it (e.g., Daw et al. 2006, but see Speekenbrink and Konstantinidis 2015). Even though in contextual bandit tasks the role of uncertainty in exploration is far greater due to the generalization effect, this does not necessarily mean that exploration should be completely driven by uncertainty. There are also clear-cut benefits in exploring at random. Such exploration is guaranteed to lead to unbiased samples of experience, as it decouples the choice process from knowledge about the world. Further, randomly acquired samples stand a better chance of detecting changes in the underlying reward functions and, as a result, they reduce the chance of a locked-in effect. R. C. Wilson et al. (2014) recently provided convincing evidence that people use both types of exploration, depending on the task demands. However, there is a lot of scope for further research in this direction, especially in contextual bandit problems.

In line with previous work on learning and decision making Stojić, Olsson, and Analytis (2016), we have found strong inter-individual differences in how people tackle the FMAB task – they seem to either take a function learning or mean tracking approach. Clearly, we need a better understanding of the factors behind these choices. Such a research line could help us

to improve people’s choices. For example, if the decision problem involves a lot of generalization, people equipped with function knowledge would be able to perform well even when confronted with new decision situations. Finding ways to incite people to engage more into function learning might have large practical impact. At the moment it is unclear whether these two approaches are really independent or people use both of them jointly. In the latter, the tendency to use function learning could be potentially formalized in a model and captured with a parameter, similar to the tendency to use model-based vs. model-free strategies as proposed by Daw, Gershman, Seymour, Dayan, and Dolan (2011). This would be an important direction of further theoretical development. There is also scope for investigating how to improve people’s function learning, even if they are already using such an approach. For example, by using smarter decision strategies that take the uncertainty into account or information about the horizon. We started this line of research in Experiment 3 by manipulating the horizon, while examining other factors is left for future studies.

3.9.2 Limitations

In the present research we focused on the FMAB problem, a constrained version of the general CMAB problem. We could easily extend our paradigm to include shared context cues, for example by varying the color of the buttons from trial to trial (e.g., how hungry you are or what is the weather like). Would people be able to cope with learning the function in such complex CMAB scenarios? Or would they give up and fall back to a mean tracking strategy? Schulz, Konstantinidis, and Speekenbrink (2016) tackled the SMAB problem, a version of the CMAB with shared context only. Using similar \mathcal{GP} models, they found evidence that people also learn the reward functions, even several of them simultaneously. There is hope then that people would be able to cope with tasks including both alternative-specific features and shared context. Such a task with greater external validity was unfortunately beyond our scope and we leave it for future research.

Even though the FMAB task (and CMAB more generally) captures realistic decision situations fairly well, it is missing some important characteristics of more general RL problems. In our FMAB task, agents receive feedback on their decisions immediately. This is a relevant simplification as it precludes long-term consequences of actions, that is, the credit assignment problem is eliminated (Barto, Sutton, & Anderson, 1983; Sutton & Barto, 1998). Delayed feedback is an important aspect of realistic decision

situations; consider choosing a career path or a major at university – many intermittent choices will be made before the outcome is known. This is a difficult problem, especially when combined with function learning, requiring further theoretical work. There are several directions which could be pursued. One is to use more appropriate decision strategies, such as the recently developed BAMCP (Guez et al., 2012, 2014), that can deal with delayed feedback. Another would be to draw on recent work on adapting Kalman filters to temporal difference learning (Geist & Pietquin, 2010; Gershman, 2015).²¹

So far the empirical results are inconclusive on whether the GP framework should be understood only as a rational benchmark, poised at Marr's computational level (Marr, 1982), or whether it can be seen as an algorithmic process model of how people tackle the task (Lucas et al., 2015; Schulz, Tenenbaum, et al., 2016). One could postulate delta-learning on the feature level (Gluck & Bower, 1988; Niv et al., 2015) for learning linear functions, or an exemplar-based model for dealing with nonlinear functions (Nosofsky, 1984). An important drawback of these models is that they give only point estimates, not the posterior distributions needed for uncertainty-based exploration. Incorporating uncertainty in such models is an important venue for future theoretical development.

In our experiments so far we have assumed that the alternatives are characterized by two informative features. In real life problems, the potential number of useful features is much larger and people have to discover which are the most useful ones (Klayman, 1988; Niv et al., 2015, e.g.). A few good cues could already improve the performance of decision makers by a large margin, as opposed to random exploration. In fact, in some cases simple models relying on one or a few cues could even outperform more complex models integrating all of the available information (Davis-Stober, 2011; Hogarth & Karelaia, 2005a; Todd & Gigerenzer, 2000). There is sufficient evidence that people learn to integrate the informative features and ignore those that do not lead to improvements in choice or estimation (Rieskamp, 2008). Our experiments have not dealt yet with problems of cue selection, but this can be done readily in the future. GP models can nicely deal with feature selection issues using automatic relevance determination (ARD) techniques.

²¹In the Artificial Intelligence community this more general version of the problem has been tackled with some success. Researchers at Google Deepmind developed an algorithm combining deep neural networks (LeCun, Bengio, & Hinton, 2015) and Q-learning (C. J. C. H. Watkins & Dayan, 1992) that achieves human-level performance on Atari video games, receiving only raw video feed and game scores (Mnih et al., 2015).

3.9.3 Concluding Remarks

Most real life problems require a combination of our cognitive capacities to be dealt with appropriately. Yet, experimental paradigms and theories often ignore interactions between cognitive processes and focus on single processes instead. Function learning and reinforcement learning are a prime example: although in real life they are used in tandem, they have been studied independently so far and the rich space of interactions between them remains largely unexplored. In this paper we advanced a new experimental paradigm that allows us to study their interactions and developed a new conceptual framework to understand how they operate. In the future, our methodology will allow us to tackle a number of psychological problems, such as preference learning and the transfer of learning between situations, that have not yet been addressed adequately.

Appendix

3.A Ensuring data quality

Most of the participants in our studies were recruited through Amazon's Mechanical Turk (AMT, <http://mturk.com>). Many classical tasks from experimental psychology were successfully replicated with participants drawn from the AMT pool (Crump, McDonnell, & Gureckis, 2013), who conducted the experiments on their browsers. Using AMT comes with many advantages – from low cost and collection speed to more heterogeneous participants that are more representative of the population at large (Paolacci & Chandler, 2014; Paolacci, Chandler, & Ipeirotis, 2010; Stewart et al., 2015). The downside is that motivation of such unsupervised participants is lower and they seem to be less attentive. However, this issue is alleviated by including “catch trials” that can be used to identify less attentive subjects and exclude them from the analysis (Paolacci et al., 2010). All the procedures for ensuring the data quality described below were used for the study with lab participants as well (Experiment 1B). Overall, 79 out of 1068 participants (or 7.4%) was excluded.

Our “catch trials” consisted of four simple attention question that participants had to answer after they finished with reading the instructions. The questions checked whether they can recall basic information from the instructions. Questions in all studies were: “What is the shape of the option buttons?”, “From how many options you can choose from?”, “What is the fixed payment in US dollars you will receive regardless of your performance in the experiment?” and “How many experimental points (EP) will be exchanged for a dollar?”. Participants were allowed to continue regardless of their answers. We analyzed how the number of correctly answered attention questions relates to performance in the experiment on the data from the pilot study reported in Stojić et al. (2015). Performance in FMAB and MAB task improves as a function of correctly answered questions, going from mean rank of 9.48 for those with zero correct answers to 7.87 for the group that answered all four questions correctly. The same pattern

holds for test choices. This indicates that questions indeed capture participants' motivation or attention. Overall, performance is relatively similar for participants that answered two or more questions correctly, while it is significantly worse for those that answered fewer questions correctly. For this reason we decided to exclude all the participants with zero or one attention question correctly answered from the analysis in studies reported here. We also compared the distributions of attention question answers between the AMT participants in Experiment 1A and lab participants in Experiment 1B. Although AMT participants do answer fewer questions correctly than lab ones (Wilcoxon rank sum test, $Z = 5408$, $p < .0001$, $M = 3.02$ ($SD = 0.93$) and $M = 3.51$ ($SD = 0.66$), respectively), in absolute terms mean accuracy is still high, which gives us confidence in the data acquired over AMT.

In addition to these attention questions, we used AMT's Qualification system to screen out members that have a poor history of providing good quality work (Chandler, Mueller, & Paolacci, 2014). Participants were required to be based in the United States and have an approval rate of 95% or above. This means that in at least 95% of cases they were paid for the work they had done—a rough measure of the quality of the work done on AMT.

Another criterion we used for ensuring the quality was to exclude the participants that chose the same alternative throughout the CMAB or MAB task. Given the large number of alternatives in our tasks, this is unlikely to be a decision strategy, and more likely to be an indication of either lack of motivation or lack of understanding the task. Such participants usually had extremely low experiment duration as well. There were in total 10 such participants across all experiments.

3.B Details on stimuli in the functional knowledge task

3.B.1 Items in the linear environments

In Experiment 1A and 1B we used positive linear function for constructing stimuli in the bandit task. In the FK task participants faced 70 trials where in each trial they had to choose between three alternatives they have not seen before. Hence, to perform well in the task they should have some knowledge about the functional relationship between the feature values and rewards. Choice triplets always consisted of a dominating, a middle, and a dominated alternative. Feature values were drawn randomly from

specifically designed uniform distribution intervals. Since participants were not receiving feedback in this task, the error term contribution was not added to the reward calculation. Finally, only participants in the contextual conditions completed this task.

There were five types of items in Experiment 1A and 1B. Two of them were easy and difficult *interpolation* triplets (15 easy and 25 difficult items), where feature values were never drawn from outside of $U(0.1, 0.9)$ interval, which participants experienced in the bandit task. Two other types were easy and difficult *extrapolation* triplets (10 easy and 10 difficult items), where feature values were exclusively drawn from regions that participants did not see in the bandit task, $U(0, 0.1)$ and $U(0.9, 1)$. Denoting dominating, middle and dominated alternative as x , y and z , respectively, sampling procedure was the following. We always drew first the feature values for the dominating alternative. For easy interpolation case, $x_1 \sim U(0.4, 0.8)$ and $x_2 \sim U(0.4, 0.8)$ and for difficult interpolation case, $x_1 \sim U(0.4, 0.8)$ and $x_2 \sim U(0.3, 0.9)$. Once these feature values were known, exact interval for sampling feature values of the middle alternative was set. For example, for easy interpolation case $y_1 \sim U(0.3, x_1 - 0.05)$ and $y_2 \sim U(0.3, x_2 - 0.05)$. Distinction between easy and difficult items was that lower boundary of the interval was also dependent on feature values of x and was made smaller, effectively making feature values of the middle alternative more similar to the dominating alternative, and their rewards being very close to each other. Construction of extrapolation trials followed the same logic, while drawing only from $U(0, 0.1)$ and $U(0.9, 1)$. Special fifth type of items, weight comparison type (10 items), was intended to detect whether participants learned that one feature had a larger weight than the other. Here a trial consisted of one alternative that had a large value on a feature with higher weight and a small value on the other feature, one alternative with the opposite pattern, and one alternative that was clearly dominated. Hence, feature values for dominating alternative were sampled from $U(0.7, 0.8)$ and $U(0.2, 0.3)$, for the middle one from $U(0.2, 0.3)$ and $U(0.7, 0.8)$ and for dominated alternative from $U(0.2, 0.3)$ and $U(0.2, 0.3)$.

Task items for mixed linear FMAB condition in Experiment 2 are very similar to those in Experiment 1A and 1B. Analysis of the FK task results from previous experiments showed that there is no difference in performance between interpolation and extrapolation type pf items. For this reason we did not distinguish between interpolation and extrapolation anymore and sampled feature values from the whole interval $U(0, 1)$. Hence, in this condition FK task consisted of three item types, 20 easy, 30 difficult and 20 weight comparison items. The sampling procedure was analogous to the one described for previous experiments. For easy and difficult items

we first sample feature values for dominating alternatives ($x_1 \sim U(0.5, 1)$ and $x_2 \sim U(0, 0.5)$), and then the intervals for the middle alternative are modified depending on the sampled value for the dominating alternative, and similarly for the dominated alternative, depending on the feature values of the middle alternative. Same as before, for difficult items distances between the intervals from which we sampled feature values were very small. For weight comparison items, feature values for dominating alternative were sampled from $U(0.7, 0.8)$ and $U(0.25, 0.35)$, for the middle one from $U(0.7, 0.8)$ and $U(0.7, 0.8)$ and for dominated alternative from $U(0.25, 0.35)$ and $U(0.4, 0.5)$

In three experimental conditions in Experiment 3 (FMAB-pl, fFMAB-pl and fFMAB-pls) we used the positive linear function from Experiment 1B. However, in the design of the FK task items, we discarded the distinction between interpolation and extrapolation and we sampled feature values from the whole interval $U(0, 1)$.

3.B.2 Items in the nonlinear environments

In Experiment 2 and 3 we had experimental conditions where rewards were determined by U-shaped quadratic function (FMAB-q, MAB-q, FMAB-q, fFMAB-q, fFMAB-qs). One of the hypotheses in these conditions was that participants might not discover the true nature of the function – they might not sample observations from certain part of the feature space and end up believing that the function is either positive or negative linear. Hence, we designed items that would allow us to detect whether participants believe that the function is positive linear, negative linear or quadratic.

There were six item types in total. First two types, “Max” and “Max Decoy” were aimed at detecting whether participants know where the true global maximum is (15 items each). In “Max” the dominating alternative was a global maximum (very low feature values), while the middle was a local maximum (very large feature values) and the dominated alternative was closer to the dominating one in terms of feature values. Participants with negative linear or nonlinear knowledge would mostly choose the highest ranking alternative, while those with positive linear beliefs would mostly choose the second ranking alternative. In “Max Decoy” the dominated alternative was closer to the middle alternative (local maximum) in feature values, potentially making it more attractive. Thus, this triplet can be considered to be a “decoy” test.

With next two types, “Min Local” and “Min Global”, we aimed to detect whether participants realized the minimum is in the middle of the interval

and the relationship is nonlinear (10 items each). In “Min Local” dominating and middle alternative have large feature values, while dominated alternative is in the middle of the interval. “Min Global” type is similar, but now dominating and middle alternative have small feature values, closer to global maximum. If people realized it is a quadratic relationship, they would rightly choose dominating alternatives with small or large feature values in “Min Global” and “Min Local” and avoid ones with features from the middle of the interval. We can detect positive or negative linear believers as well. With negative linear beliefs, participant would choose mostly the dominated alternative in “Min Local” and dominating in “Min Global”, while with positive linear beliefs pattern of choices would be exactly the opposite.

Final two types, “Slope Global” and “Slope Local”, had similar purposes as Min types (10 items each). In “Slope Global” dominating alternative had small feature values (global maximum), middle had large values, while dominated alternative was sampled from the middle of the interval. In “Slope Local” feature values of dominating and middle alternative were switched. Participant with quadratic function in mind would tend to choose dominating arms in both “Slope Global” and “Slope Local”. Participant with negative linear beliefs would choose the dominating and the dominated alternative in “Slope Global”, and the middle and dominated in “Slope Local”. For positive linear believers patterns would exchange – they would choose the dominating and the dominated alternative in “Slope Local”, and the middle and dominated in “Slope Global”.

Overall, a person with perfect knowledge of the function would choose the dominating alternative in each item type, but as explained above, there are specific patterns of answers that are indicative of linear or negative linear function beliefs. To facilitate the analysis we classified participants into one of the tree types of beliefs – “Positive”, “Negative” and “Quadratic”. We computed distances between mean choice on each item type and choices a person with perfect knowledge of positive, negative or quadratic function would make, and classified people as having those beliefs to which the distance was the smallest. Finally, we down-weighted the importance of “Max” and “Max decoy” item type choices in the similarity index (10% of the import of other item types), as choice sets were drawn randomly for each participant, in some choice sets the highest rewarding alternative was one with large feature values. Hence, even if people knew that the general shape of the function was quadratic, they would not necessarily detect global vs. local maximum due to properties of their choice set.

3.C Additional Results

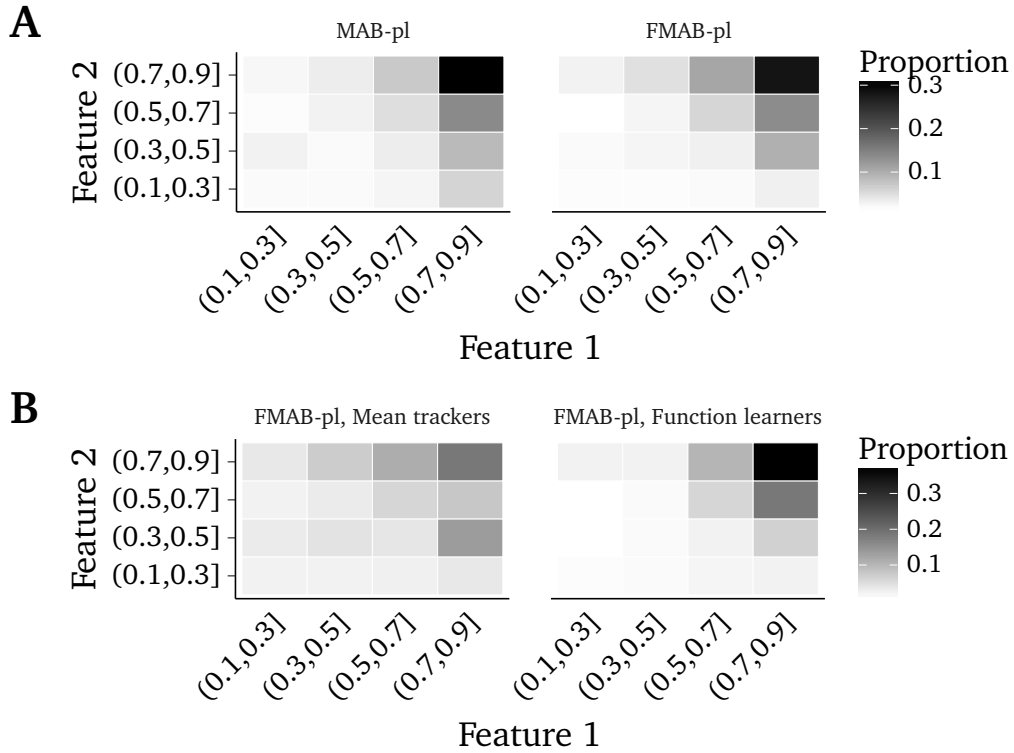


Figure 3.C.1 Exploration patterns of participants in all 100 trials of the bandit tasks in Experiment 1A. (A) By the end of the task allocations of choices of participants in both MAB and FMAB conditions converge – they concentrate their choices in high rewarding region of the feature space, as indicated by feature values displayed on x-axis for Feature 1 ($w_1 = 2$) and y-axis for Feature 2 ($w_2 = 1$). (B) Similarly, allocations of choices of two clusters of FMAB participants also become very similar by the end of the bandit task.

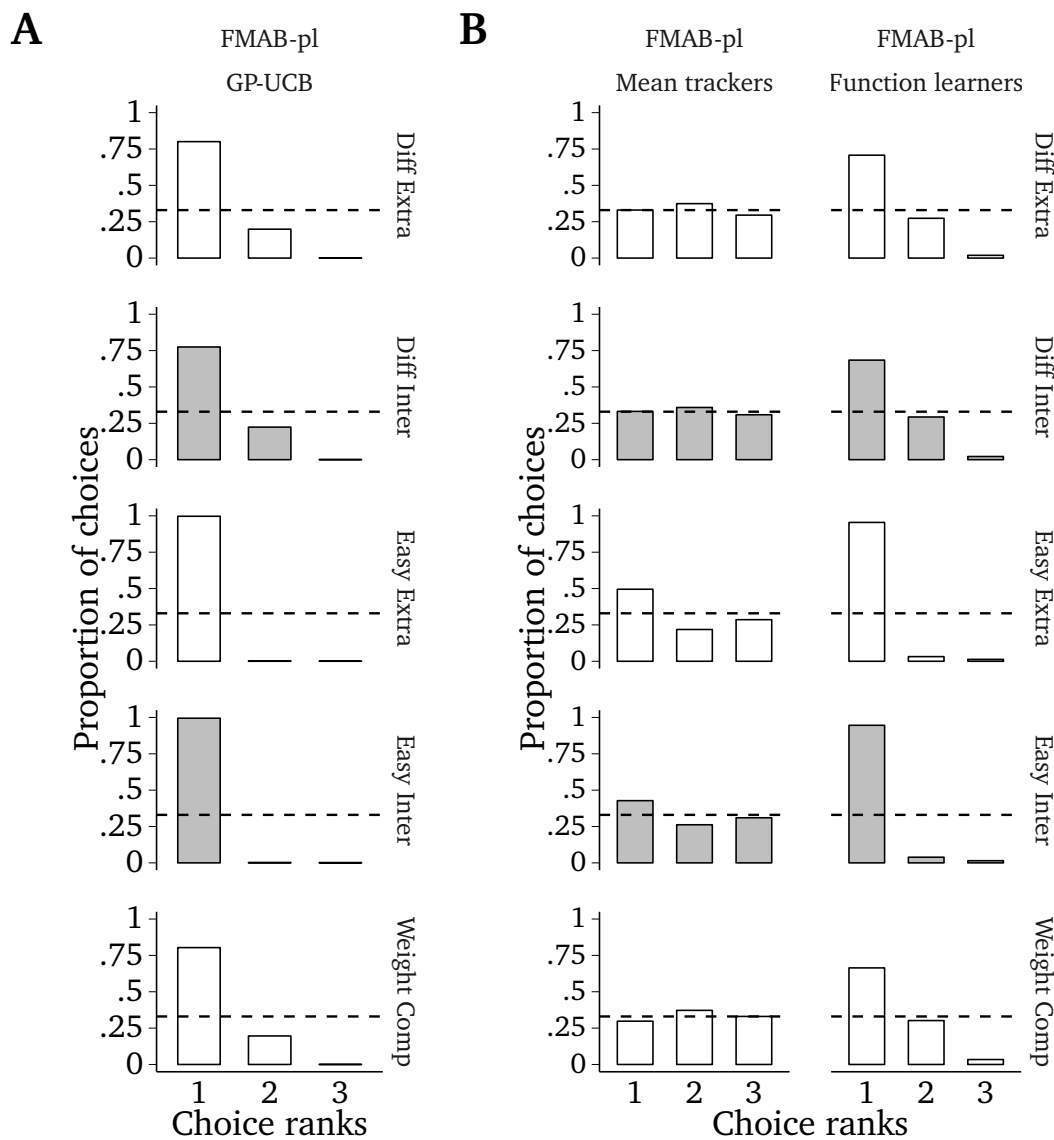


Figure 3.C.2 Performance of GP-UCB model and participants in Experiment 1A on FK task, broken across item types. (A) GP-UCB is able to generalize and performs very well, although for difficult and weight comparison items it errs sometimes. Not shown here, performance of the BMT-UCB model would equal to random choice, in contrast. (B) Performance of two clusters of participants from FMAB-pl condition – mean trackers and function learners, is very similar to BMT-UCB and GP-UCB models, respectively. On average, mean trackers make choices that are close to random, while function learners exhibit the same pattern as GP-UCB, very good performance with some errors on difficult weight comparison items.

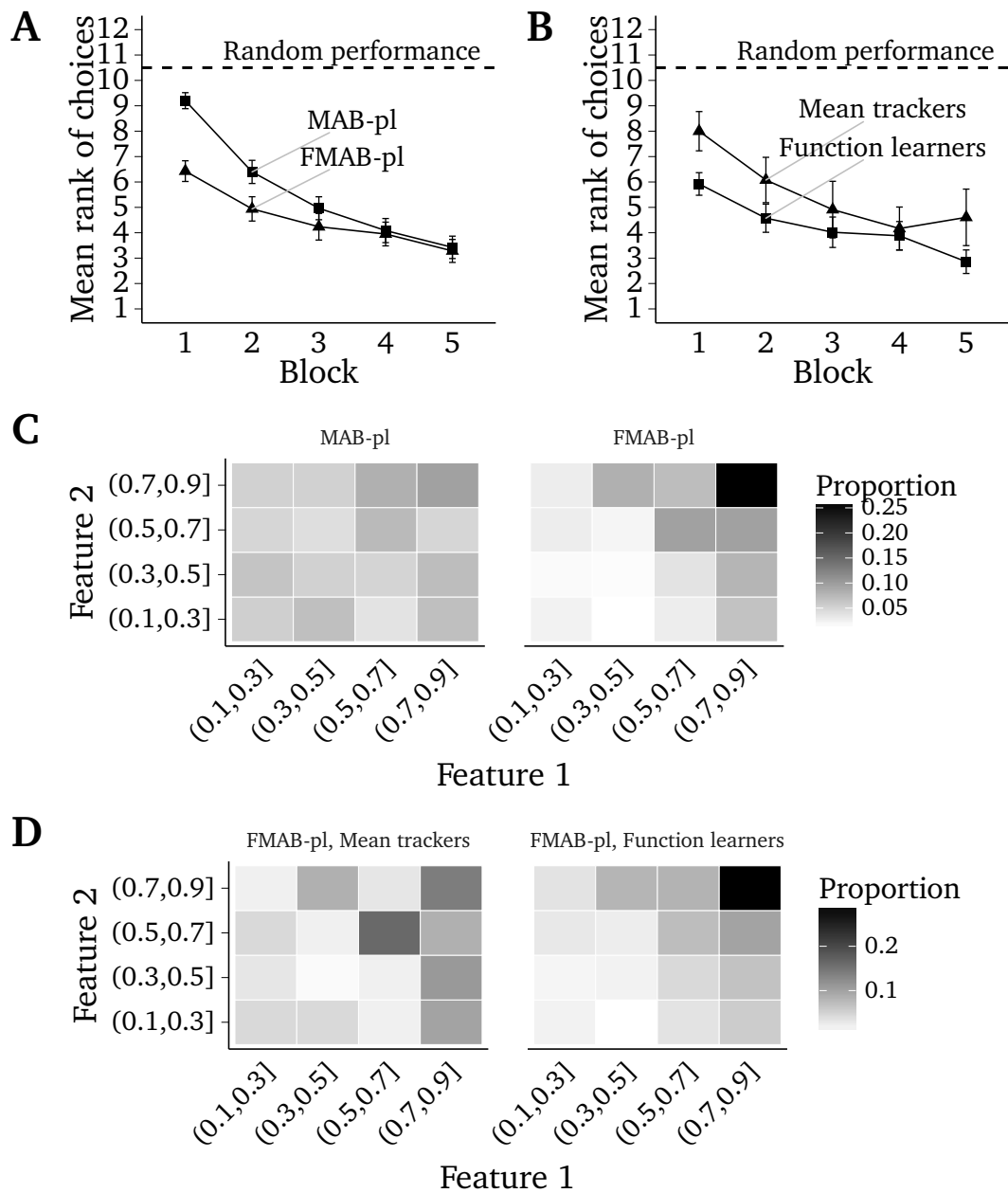


Figure 3.C.3 Behavioral results of participants in the bandit task in Experiment 1B. (A) Mean accuracy of choices (the lower the better) increased across trials (grouped in five blocks of 20 trials). (B) Similar to Experiment 1A, participants that learn the function (according to the FK task) are doing much better than MAB participants and FMAB participants that ignored the feature information. (C) Exploration patterns of participants in the first 10 trials of the bandit tasks reveal that participants in the FMAB condition start allocating the choices to the high rewarding region much faster than MAB participants. (D) Cluster that does not exhibit functional knowledge explores similar to MAB participants.

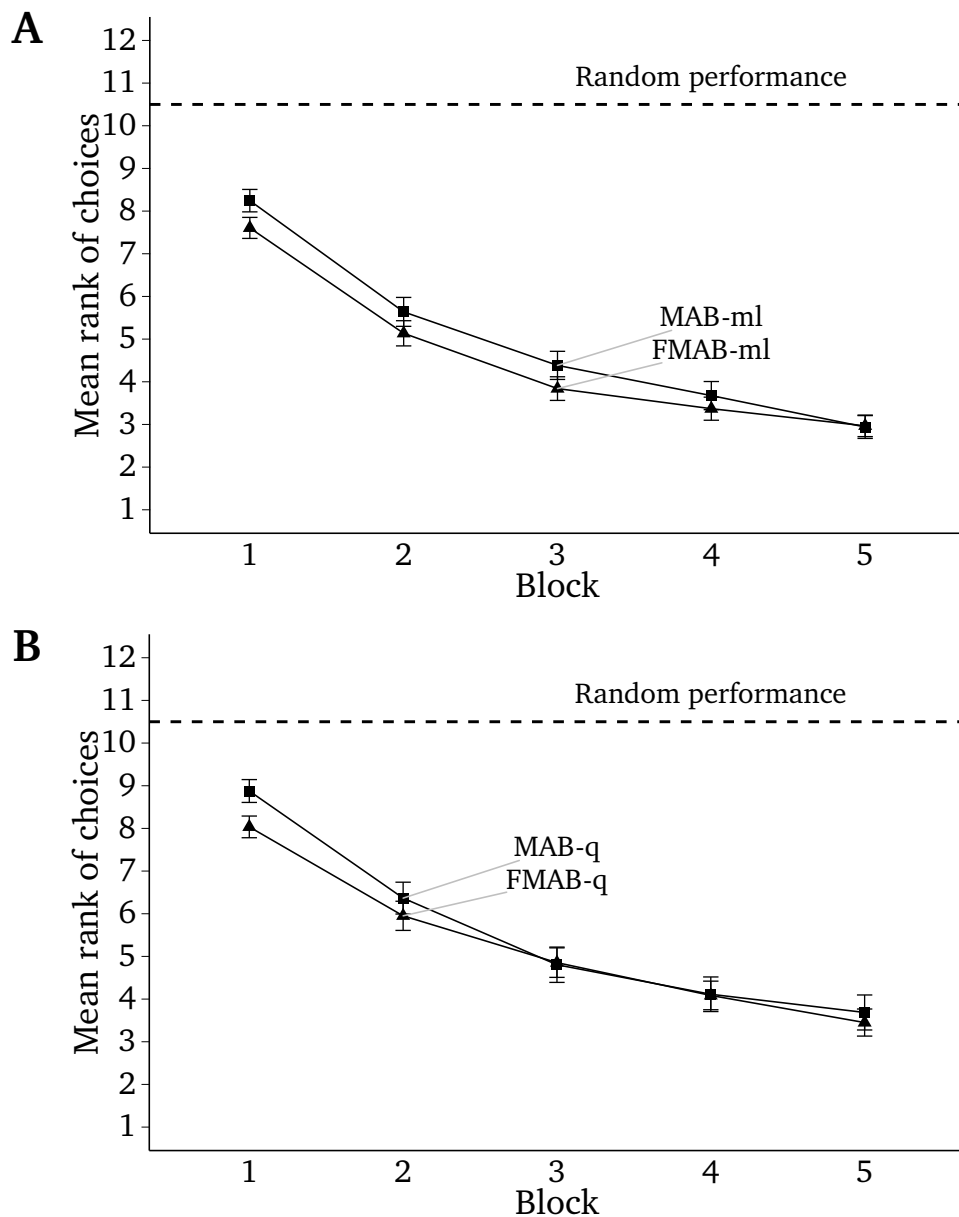


Figure 3.C.4 Choice performance of participants in Experiment 2 expressed as mean choice ranks (the lower the rank the better) as a function of trials in the bandit task (grouped in five blocks of 20 trials). (A) Participants in the condition with mixed linear reward function learn to make good choices in both MAB-ml and FMAB-ml task – their mean rank increases substantially over time, and FMAB participants are performing slightly, but consistently better than MAB participants. (B) Analogous to the previous panel, but now showing performance for conditions with quadratic reward function, with essentially the same results.

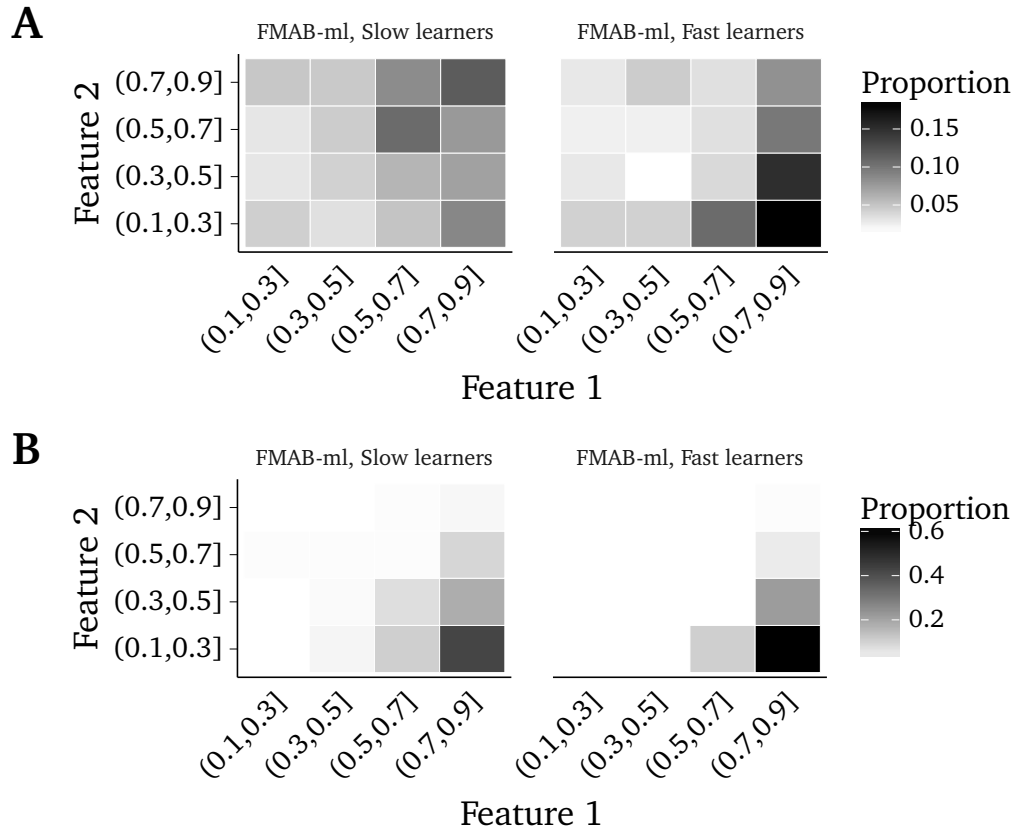


Figure 3.C.5 Exploration patterns of two clusters of participants in the FMAB-ml condition in Experiment 2 in the first 10 trials of the bandit tasks and the last 10 trials. Clusters were determined with K-means method on mean choice ranks from the FK task for each participant. (A) Proportion of all choices in first 10 trials allocated by clusters of participants to alternatives with feature values falling into one of the four bins indicated on x-axis for Feature 1 ($w_1 = 40$) and y-axis for Feature 2 ($w_2 = -30$). One group – “Slow learners”, that has poor FK task performance, allocates particularly large portion of choices to the upper right corner, as if guided by a prior on expecting positive linear relationships. The other group, “Fast learners” has very good knowledge about the function according to the FK task, but quickly unlearns their prior, if the group had it at all. (B) Analogous to the previous panel, but now showing allocation of choices for the last 10 trials in the FMAB task. Both clusters shifted their choices to the alternatives with feature values that lead to higher rewards, positioned in the lower right corner, where second feature has low values.

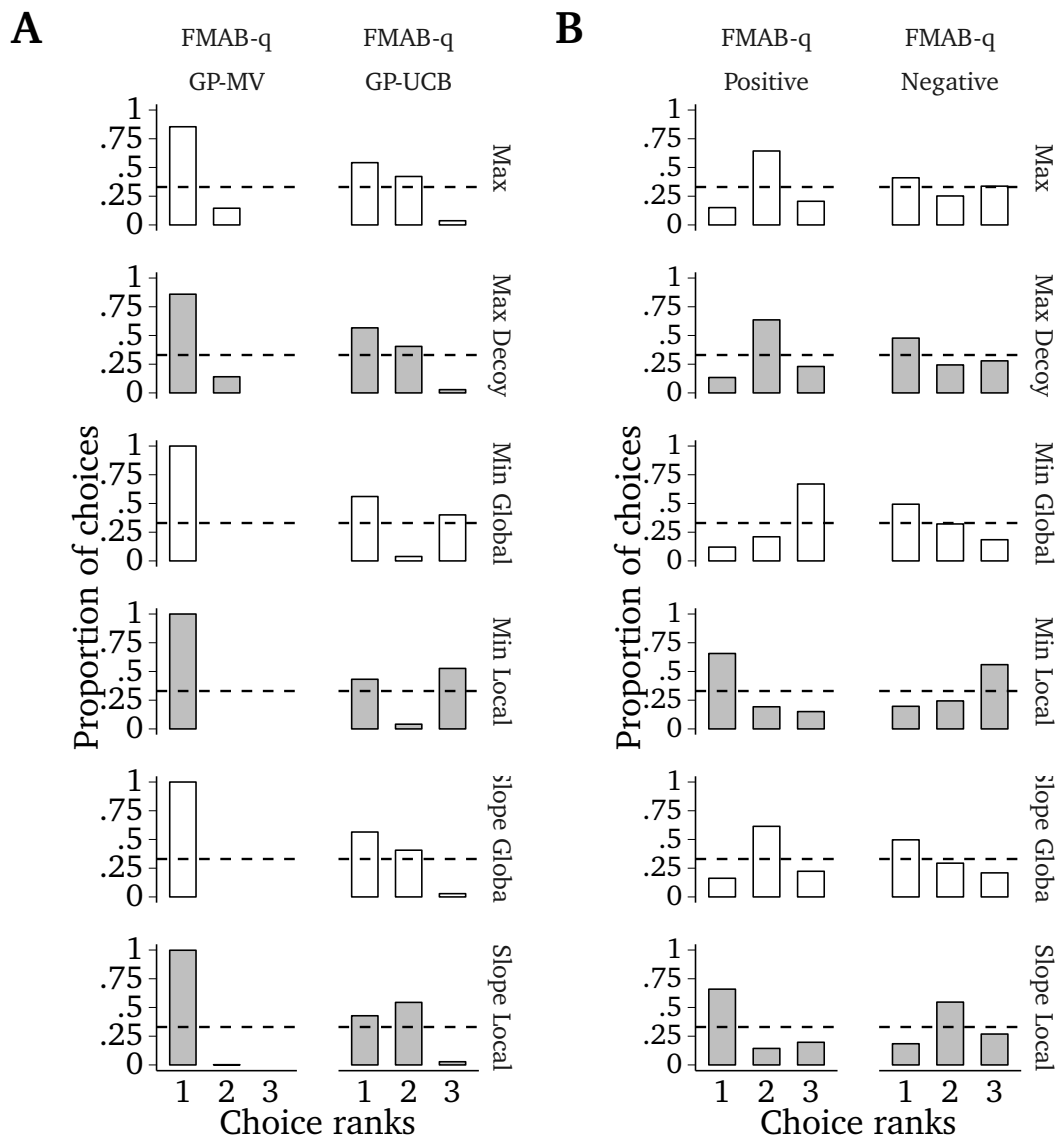


Figure 3.C.6 Performance of Bayesian models and participants in FMAB-q condition in Experiment 2 on FK task. (A) In the left panel, FK task performance of the GP-MV model that chooses the alternatives with highest uncertainty. It learns the function and performs extremely well on the task, choosing the dominating alternative most of the time on all item types. The GP-UCB model performs poorly, as it focuses on high rewarding part of the function and does not learn the other parts well, ending up with either positive or negative functional knowledge. (B) Performance in the FK task decomposes on two groups of participants – one with positive linear and one with negative linear beliefs, according to their pattern of choices in the task. Positive group is bigger ($N = 69$) than the negative ($N = 32$), most likely due to priors, and has clearer results. Quadratic group is not shown due to small sample size ($N = 1$). Pattern of allocations on all six items is according to predictions (see Appendix 3.B).

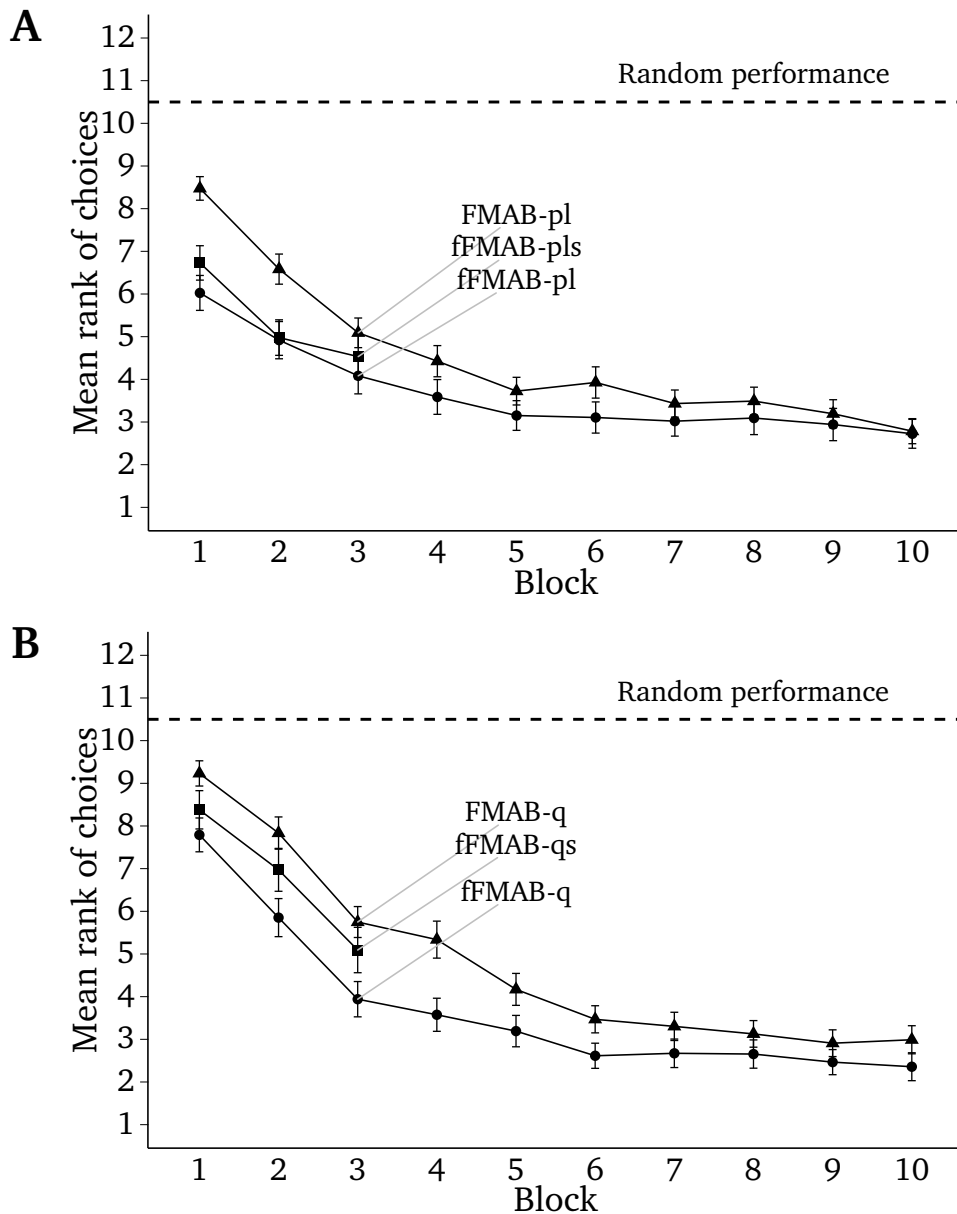


Figure 3.C.7 Choice performance of participants in Experiment 3 expressed as mean choice ranks (the lower the rank the better) as a function of trials in the bandit task (grouped in 10 blocks of 10 trials). (A) Participants in the condition with positive linear reward function learn to make good choices in all three conditions – their mean rank increases substantially over time. Notably, conditions with function learning pretraining perform significantly better, knowing the function better pays off. However, there (B) Analogous to the previous panel, but now showing performance for conditions with quadratic reward function, with qualitatively the same results. As expected, quadratic function is more difficult to learn and advantage of function learning pretraining is smaller.

3.D Bayesian Models Performance and Parameter Overview

To arrive at benchmark performance for participants in each experiment we took stimuli that was generated for each participant and fitted our Bayesian models to the stimuli. The parameters of the models were estimated by using the mean rank of chosen alternatives in the bandit task as model performance measure. Since the model choices for each set of parameters are stochastic, we simulated the model for each set 20 times and took a mean of the resulting mean rank of chosen alternatives. Optimization was done using the Nelder–Mead simplex algorithm implemented in the `optim` function in R (R Core Team, 2015). We used a multi-start procedure where we first generated 30 sets of parameters, uniformly dispersed in the parameter space, and chose two best sets as initial points in the Nelder–Mead simplex algorithm.

Table 3.D.1 Overview of performance and estimated parameters of the models on the same stimuli that participants have had in the experiments. Performance of the model is expressed as mean rank of the chosen alternative throughout the bandit task. For performance and each parameter we display the mean value across all subject-specific stimuli, together with standard deviation in parenthesis.

Experiment	Condition	Model	Rank	σ_r	σ_f	l	α
Exp 1A	FMAB-pl	BMT-UCB	1.87 (0.26)	0.3 (0.23)	1.26 (0.28)	–	2.21 (0.39)
		GP-MV	9.81 (0.73)	0.59 (0.45)	0.87 (0.62)	1.72 (3.59)	–
		GP-UCB	1.28 (0.15)	0.26 (0.2)	1.4 (0.43)	1.34 (0.44)	1.9 (0.65)
Exp 1B	FMAB-pl	BMT-UCB	2.21 (0.6)	2.54 (1.5)	8.05 (2.02)	–	5.35 (2.66)
		GP-MV	9.87 (0.67)	0.6 (0.39)	0.95 (0.53)	0.94 (1.18)	–
		GP-UCB	1.38 (0.24)	0.09 (0.03)	1.74 (0.44)	1.69 (0.22)	1.98 (0.97)
Exp 2A	FMAB-nl	BMT-UCB	3.59 (0.82)	0.3 (0.28)	1.73 (0.81)	–	2.28 (1)
		GP-MV	9.63 (0.88)	0.69 (0.44)	0.88 (0.68)	2.98 (17.74)	–
		GP-UCB	1.57 (0.48)	0.24 (0.18)	1.88 (0.37)	0.98 (0.5)	2.04 (0.8)
Exp 2B	FMAB-q	BMT-UCB	3.94 (1.7)	0.2 (0.22)	2.28 (1.18)	–	2.98 (1.3)
		GP-MV	2.79 (1.12)	1.16 (0.76)	0.24 (0.3)	6.2 (17.85)	–
		GP-UCB	2.01 (0.43)	0.2 (0.18)	1.58 (0.38)	1.47 (0.48)	2.17 (0.96)
Exp 3	FCMAB-pl	BMT-UCB	2.35 (0.63)	2.1 (0.53)	6.92 (1.73)	–	4.72 (1.05)
		GP-UCB	1.4 (0.53)	1.26 (0.7)	6.65 (2.9)	3.04 (2.5)	6.99 (2.31)
		BMT-UCB	2.5 (0.91)	1.94 (0.9)	7.87 (1.03)	–	4.25 (1.77)
		GP-UCB	1.26 (0.77)	1.58 (0.3)	4.88 (0.99)	4.77 (1.34)	5.44 (2.89)
	FCMAB-q	BMT-UCB	1.98 (0.64)	1.38 (0.75)	4.92 (2.24)	–	5.48 (1.03)
		GP-UCB	1.37 (0.45)	2.06 (1)	6.61 (2.27)	4.96 (3.68)	7.34 (1.99)
		BMT-UCB	2.36 (0.71)	1.75 (0.67)	4.06 (2.43)	–	6.86 (3.07)
		GP-UCB	1.12 (0.36)	1.41 (0.56)	6.06 (1.43)	3.33 (2.1)	5.83 (1.84)
	fFCMAB-pls	BMT-UCB	3.5 (0.74)	1.67 (0.78)	6.83 (1.4)	–	3.99 (1.1)
		GP-UCB	1.04 (0.27)	1.62 (0.64)	5.25 (0.95)	4.59 (1.86)	5.12 (1.63)
		BMT-UCB	2.92 (0.43)	1.16 (0.57)	4.71 (2.28)	–	4.95 (2.41)
		GP-UCB	1.18 (0.36)	1.32 (0.53)	6.44 (1.65)	2.98 (2.09)	5.7 (1.5)

Note: FMAB = Feature-based multi-armed bandit task, suffix denotes the function determining the reward, with pl = positive linear function, nl = mixed linear function, q = quadratic function; BMT = Bayesian Mean Tracker model for learning the average rewards; GP = Gaussian Process function learning model; MV = Maximum Variance choice rule; UCB = Upper Confidence Bound choice rule.

Bibliography

- Analytis, P. P., Kothiyal, A., & Katsikopoulos, K. V. (2014). Multi-attribute utility models as cognitive search engines. *Judgment and Decision Making*, 9, 403–419.
- Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56, 149–78. doi: 10.1146/annurev.psych.56.091103.070217
- Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, 1224, 147–161. doi: 10.1111/j.1749-6632.2010.05874.x
- Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3, 397–422.
- Auguie, B. (2012). *gridExtra: functions in Grid graphics*. Retrieved from <http://cran.r-project.org/package=gridExtra>
- Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16, 215–233. doi: 10.1002/bdm.443
- Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13, 834–846. doi: 10.1109/TSMC.1983.6313077
- Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, 3, 439–449.
- Bergert, F. B., & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 107–129. doi: 10.1037/0278-7393.33.1.107
- Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 708–731. doi: 10.1037/xlm0000061

- Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, 11, 1–27. doi: 10.1016/0030-5073(74)90002-6
- Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica*, 87, 137–154.
- Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 611–625. doi: 10.1037/0278-7393.29.4.611
- Bröder, A. (2012). The quest for take the best - Insights and outlooks from experimental research. In P. M. Todd, G. Gigerenzer, & the ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 216–240). New York, NY, US: Oxford University Press.
- Bröder, A., & Schiffer, S. (2006). Adaptive flexibility and maladaptive routines in selecting fast and frugal decision strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 904–918. doi: 10.1037/0278-7393.32.4.904
- Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. Wiley.
- Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. R. Shanks (Eds.), *Knowledge, concepts and categories. studies in cognition*. (pp. 408–437). Cambridge, MA, US: MIT Press.
- Busemeyer, J. R., & Wang, Y.-M. (2000). Model Comparisons and Model Selections Based on Generalization Criterion Methodology. *Journal of Mathematical Psychology*, 44, 171–189. doi: 10.1006/jmps.1999.1282
- Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46, 112–130. doi: 10.3758/s13428-013-0365-7
- Chapelle, O., & Li, L. (2011). An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems*, 2249–2257.
- Christensen-Szalanski, J. J. (1978). Problem solving strategies: A selection mechanism, some implications, and some data. *Organizational Behavior and Human*, 22, 307–323. doi: 10.1016/0030-5073(78)90019-3
- Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, 15, 201–221. doi: 10.1007/BF00993277

- Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One*, *8*, e57410. doi: 10.1371/journal.pone.0057410
- Davis-Stober, C. P. (2011). A geometric analysis of when fixed weighting schemes will outperform ordinary least squares. *Psychometrika*, *76*, 650–669. doi: 10.1007/s11336-011-9229-1
- Davis-Stober, C. P., Dana, J., & Budescu, D. V. (2010). Why recognition is rational: Optimality results on single-variable decision rules. *Judgment and Decision Making*, *5*, 216–229.
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204–1215. doi: 10.1016/j.neuron.2011.02.027
- Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879. doi: 10.1038/nature04766
- Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*, 185–196. doi: 10.1016/j.conb.2008.08.003
- DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: the sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968–986.
- Denrell, J. (2005). Why most people disapprove of me: experience sampling in impression formation. *Psychological Review*, *112*, 951–978. doi: 10.1037/0033-295X.112.4.951
- Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review*, *114*, 177–187. doi: 10.1037/0033-295X.114.1.177
- Denrell, J., & Le Mens, G. (2007). Interdependent sampling and social influence. *Psychological Review*, *114*, 398–422. doi: 10.1037/0033-295X.114.2.398
- Denrell, J., & March, J. G. (2001). Adaptation as Information Restriction: The Hot Stove Effect. *Organization Science*, *12*, 523–538. doi: 10.1287/orsc.12.5.523.10092
- Einhorn, H. J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin*, *73*, 221–230.
- Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*, 395–416. doi: 10.1037/0033-295X.85.5.395
- Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*, 912–31. doi: 10.1037/0033-295X.112.4.912
- Estes, W. K. (1960). Learning theory and the new "mental chemistry".

- Psychological Review*, 67, 207–223. doi: 10.1037/h0041624
- Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, 83, 37–64. doi: 10.1037/0033-295X.83.1.37
- Fiedler, K., & Juslin, P. (2006). *Information sampling and adaptive cognition* (K. Fiedler & P. Juslin, Eds.). New York, NY, US: Cambridge University Press.
- Geist, M., & Pietquin, O. (2010). Kalman temporal differences. *Journal of Artificial Intelligence Research*, 39, 483–532. doi: 10.1613/jair.3077
- Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7, 155–170.
- Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52, 45–56. doi: 10.1037/0003-066X.52.1.45
- Gershman, S. J. (2015). A Unifying Probabilistic View of Associative Learning. *PLoS Computational Biology*, 11, 1–20. doi: 10.1371/journal.pcbi.1004567
- Gershman, S. J. (2016). Context-dependent learning and causal structure. *Psychonomic Bulletin & Review*, 1–25. doi: 10.3758/s13423-016-1110-x
- Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, 117, 197–209. doi: 10.1037/a0017808
- Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*, 68, 5.1–5.28. doi: 10.1146/annurev-psych-122414-033625
- Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, 20, 251–256. doi: 10.1016/j.conb.2010.02.008
- Gershman, S. J., & Niv, Y. (2015). Novelty and Inductive Generalization in Human Reinforcement Learning. *Topics in Cognitive Science*, 1–25. doi: 10.1111/tops.12138
- Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human Reinforcement Learning Subdivides Structured Action Spaces by Learning Effector-Specific Values. *Journal of Neuroscience*, 29, 13524–13531. doi: 10.1523/JNEUROSCI.2469-09.2009
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650–669.
- Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York, NY, US: Oxford University Press.

- Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247.
- Guez, A., Silver, D., & Dayan, P. (2012). Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search. *Advances in Neural Information Processing Systems*, 1–9.
- Guez, A., Silver, D., & Dayan, P. (2014). *Better Optimism By Bayes: Adaptive Planning with Rich Models*. (arXiv)
- Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., . . . Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 1–14. doi: 10.3758/s13428-015-0642-8
- Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, *62*, 255–262.
- Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review*, *71*, 438–456.
- Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.
- Hoffmann, J., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: how memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General*, *143*, 2242–2261. doi: 10.1037/a0037989
- Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: how cognitive load can improve judgments. *Psychological Science*, *24*, 869–879. doi: 10.1177/0956797612463581
- Hogarth, R. M., & Karelaia, N. (2005a). Ignoring information in binary choice with continuous variables: When is less more? *Journal of Mathematical Psychology*, *49*, 115–124. doi: 10.1016/j.jmp.2005.01.001
- Hogarth, R. M., & Karelaia, N. (2005b). Simple Models for Multiattribute Choice with Many Alternatives: When It Does and Does Not Pay to Face Trade-offs with Binary Attributes. *Management Science*, *51*, 1860–1872. doi: 10.1287/mnsc.1050.0448
- Hogarth, R. M., & Karelaia, N. (2006a). Regions of Rationality: Maps for Bounded Agents. *Decision Analysis*, *3*, 124–144. doi: 10.1287/deca.1060.0063
- Hogarth, R. M., & Karelaia, N. (2006b). "Take-the-best" and other simple strategies: Why and when they work "well" with binary cues. *Theory and Decision*, *61*, 205–249. doi: 10.1007/s11238-006-9000-8
- Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of

- judgment: matching rules and environments. *Psychological Review*, 114, 733–758. doi: 10.1037/0033-295X.114.3.733
- Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of information processing in the basal ganglia* (Vol. 13, pp. 249–270).
- Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 924–941. doi: 10.1037/0278-7393.29.5.924
- Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, 132, 133–156. doi: 10.1037/0096-3445.132.1.133
- Kaelbling, L. P. (1994). Associative reinforcement learning: A generate and test algorithm. *Machine Learning*, 15, 299–319.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY, US: Farrar, Straus and Giroux.
- Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, 111, 1072–1099. doi: 10.1037/0033-295X.111.4.1072
- Karlsson, L., Juslin, P., & Olsson, H. (2007). Adaptive changes between cue abstraction and exemplar memory in a multiple-cue judgment. *Psychonomic Bulletin & Review*, 14, 1140–1146.
- Katsikopoulos, K. V., Schooler, L. J., & Hertwig, R. (2010). The robust beauty of ordinary information. *Psychological Review*, 117, 1259–1266. doi: 10.1037/a0020418
- Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 317–330. doi: 10.1037//0278-7393.14.2.317
- Koh, K., & Meyer, D. (1991). Function Learning: Induction of continuous Stimulus-Response Relation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 811–836. doi: 10.1037/0278-7393.17.5.811
- Kolling, N., Wittmann, M. K., Behrens, T. E. J., Boorman, E. D., Mars, R. B., & Rushworth, M. F. S. (2016). Value, search, persistence and model updating in anterior cingulate cortex. *Nature Neuroscience*, 19, 1280–1285. doi: 10.1038/nn.4382
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22–44. doi: 10.1037/0033-295X.99.1.22

- Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*, 210–226. doi: 10.3758/LB.36.3.210
- Langford, J., & Zhang, T. (2008). The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 20* (Vol. 20, pp. 817–824). Curran Associates, Inc.
- Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: on the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, *118*, 379–92. doi: 10.1037/a0023010
- Le Mens, G., Kareev, Y., & Avrahami, J. (2016). The Evaluative Advantage of Novel Alternatives: An Information-Sampling Account. *Psychological Science*, *27*, 161–168. doi: 10.1177/0956797615615581
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. doi: 10.1038/nature14539
- Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*, 1–12. doi: 10.3758/s13428-014-0458-y
- Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on world wide web* (pp. 661–670). ACM Press.
- Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 1–6). Austin, TX, US: Cognitive Science Society.
- Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*, 1193–1215. doi: 10.3758/s13423-015-0808-5
- Luce, R. D. (1959). *Individual Choice Behavior*. New York, NY, US: Wiley.
- Luce, R. D. (2000). Fast, frugal, and surprisingly accurate heuristics. *Behavioral and Brain Sciences*, *23*, 757–758.
- Markant, D. B., & Gureckis, T. M. (2014). Is It Better to Select or to Receive? Learning via Active and Passive Hypothesis Testing. *Journal of Experimental Psychology: General*, *143*, 94–122. doi: 10.1037/a0032108
- Markant, D. B., Settles, B., & Gureckis, T. M. (2015). Self-Directed Learning Favors Local, Rather Than Global, Uncertainty. *Cognitive science*, 1–21. doi: 10.1111/cogs.12220
- Marr, D. (1982). *Vision: A computational investigation into the human*

- representation and processing of visual information*. San Francisco, CA, US: Freeman.
- Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, *52*, 29–71.
- Martignon, L., & Laskey, K. B. (1999). Bayesian Benchmarks for Fast and Frugal Heuristics. In G. Gigerenzer, P. M. Todd, & A. Group (Eds.), *Simple heuristics that make us smart* (pp. 169–189). New York, NY, US: Oxford University Press.
- Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. New York, NY, US: Oxford University Press.
- McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*, 24–42. doi: 10.3758/BF03196347
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533. doi: 10.1038/nature14236
- Morton, A. (2000). Heuristics all the way up? *Behavioral and Brain Sciences*, *23*, 758–759.
- Neal, R. M. (1996). *Bayesian learning for neural networks*. Springer Verlag.
- Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979–999. doi: 10.1037/0033-295X.112.4.979
- Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: information acquisition optimizes probability gain. *Psychological Science*, *21*, 960–969. doi: 10.1177/0956797610372637
- Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*, 139–154. doi: 10.1016/j.jmp.2008.12.005
- Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, *35*, 8145–8157. doi: 10.1523/JNEUROSCI.2978-14.2015
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114. doi: 10.1037/0278-7393.10.1.104
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: Gen-*

- eral, 115, 39–61. doi: 10.1037/0096-3445.115.1.39
- Nosofsky, R. M., & Bergert, F. B. (2007). Limitations of Exemplar Models of Multi-Attribute Probabilistic Inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 999–1019. doi: 10.1037/0278-7393.33.6.999
- Olsson, A.-C., Enkvist, T., & Juslin, P. (2006). Go with the flow: How to master a nonlinear multiple-cue judgment task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 1371–1384. doi: 10.1037/0278-7393.32.6.1371
- Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, 65, 207–240. doi: 10.1016/j.cogpsych.2012.03.003
- Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6, 8096. doi: 10.1038/ncomms9096
- Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, 23, 184–188. doi: 10.1177/0963721414531598
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5, 411–419.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York, NY, US: Cambridge University Press.
- Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, 162, 8–13. doi: 10.1016/j.jneumeth.2006.11.017
- R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.r-project.org>
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.
- Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, 114, 784–805. doi: 10.1037/0033-295X.114.3.784
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York, NY, US: Appleton-Century-Crofts.

- Reverdy, P. B., Srivastava, V., & Leonard, N. E. (2014). Modeling human decision making in generalized gaussian multiarmed bandits. *Proceedings of the IEEE*, *102*, 544–571. doi: 10.1109/JPROC.2014.2307024
- RevolutionAnalytics. (2014). *doMC: Foreach parallel adaptor for the multi-core package*. Retrieved from <http://cran.r-project.org/package=doMC>
- Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1355–1370. doi: 10.1037/0278-7393.32.6.1355
- Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1446–65. doi: 10.1037/a0013646
- Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: how opportunity costs affect strategy selection. *Acta Psychologica*, *127*, 258–76. doi: 10.1016/j.actpsy.2007.05.004
- Rieskamp, J., & Otto, P. E. (2006). SSL: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236. doi: 10.1037/0096-3445.135.2.207
- Russell, S. J., & Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence*, *49*, 361–395. doi: 10.1016/0004-3702(91)90015-C
- Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload. *Journal of Consumer Research*, *37*, 409–425. doi: 10.1086/651235
- Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: a Bayesian hierarchical approach. *Psychological Review*, *120*, 39–64. doi: 10.1037/a0030777
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*, 1593–1599. doi: 10.1126/science.275.5306.1593
- Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015). Learning and decisions in contextual multi-armed bandit tasks. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2122–2127).
- Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2016). *Putting bandits into context: How function learning supports decision making*. doi: 10.1101/081091
- Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2016). *Probing the compositionality of intuitive functions*.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Opti-

- mization. *Proceedings of the IEEE*, 104, 148–175. doi: 10.1109/JPROC.2015.2494218
- Sharpsteen, C., & Bracken, C. (2015). *tikzDevice: R Graphics Output in LaTeX Format*. Retrieved from <http://cran.r-project.org/package=tikzDevice>
- Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience*, 19, 1286–1291. doi: 10.1038/nn.4384
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2), 127–190. doi: 10.1037/0033-295X.84.2.127
- Speekenbrink, M., & Konstantinidis, E. (2014). Uncertainty and exploration in a restless bandit task. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1–6).
- Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, 7, 351–367. doi: 10.1111/tops.12145
- Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, 139, 266–298. doi: 10.1037/a0018620
- Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2009). *Gaussian process optimization in the bandit setting: No regret and experimental design*.
- Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, 10, 479–491.
- Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53, 168–179. doi: 10.1016/j.jmp.2008.11.002
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27, 453–489. doi: 10.1016/S0364-0213(03)00010-7
- Stojic, H., Analytis, P. P., Dayan, P., & Speekenbrink, M. (n.d.). *Trials-with-fewer-errors: Feature-based learning and exploration*. (Unpublished working paper)
- Stojić, H., Olsson, H., & Speekenbrink, M. (2016). *Not everything looks like a nail: Learning to select appropriate decision strategies in multiple environments*. (PsyArXiv) doi: 10.17605/OSF.IO/FMA3P
- Stojić, H., Analytis, P. P., Dayan, P., & Speekenbrink, M. (2016a). *Trials-*

- with-fewer-errors: Feature-based learning and exploration - Project files.* Open Science Framework. Retrieved from <https://osf.io/fmn45/>
- Stojić, H., Analytis, P. P., Dayan, P., & Speekenbrink, M. (2016b). Trials-with-fewer-errors: Feature-based learning and exploration - Raw data from experiments. *figshare*. doi: 10.6084/m9.figshare.3189748
- Stojić, H., Analytis, P. P., & Speekenbrink, M. (2015). Human behavior in contextual multi-armed bandit problems. In D. Noelle et al. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 2290–2295). Austin, TX, US: Cognitive Science Society.
- Stojić, H., Olsson, H., & Analytis, P. P. (2016). Explaining inter-individual variability in strategy selection: A cue weight learning approach. In D. Reitter & F. E. Ritter (Eds.), *Proceedings of the 14th International Conference on Cognitive Modeling* (pp. 144–150). University Park, PA: Penn State.
- Stojić, H., Olsson, H., & Speekenbrink, M. (2016a). *Not everything looks like a nail: Learning to select appropriate decision strategies in multiple environments - Project files.* Open Science Framework. Retrieved from osf.io/3q5if
- Stojić, H., Olsson, H., & Speekenbrink, M. (2016b). Not everything looks like a nail: Learning to select appropriate decision strategies in multiple environments - Raw data from experiments. *figshare*. doi: 10.6084/m9.figshare.1585822
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3, 9–44. doi: 10.1007/BF00115009
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA, US: MIT Press.
- Thompson, W. R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25, 285–294. doi: 10.2307/2332286
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273–286.
- Todd, P. M., & Gigerenzer, G. (2000). Précis of Simple heuristics that make us smart. *Behavioral and Brain Sciences*, 23, 727–780.
- Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, 71, 680–683. doi: 10.1037/h0023123
- von Helversen, B., & Rieskamp, J. (2008). The mapping model: a cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General*, 137, 73–96. doi: 10.1037/0096-3445.137.1.73
- Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11, 192–196. doi: 10.3758/

BF03206482

- Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, 8, 279–292.
- Watkins, M. J. (1984). Models as toothbrushes. *Behavioral and Brain Sciences*, 7, 86.
- Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, 641–654.
- Whittle, P. (1980). Multi-Armed Bandits and the Gittins Index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42, 143–149.
- Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21, 1–20.
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.
- Wickham, H., & Francois, R. (2015). *dplyr: A Grammar of Data Manipulation*. Retrieved from <http://cran.r-project.org/package=dplyr>
- Wilson, A. G., Dann, C., Lucas, C. G., & Xing, E. P. (2015). The Human Kernel. In *Advances in neural information processing systems* (pp. 2854–2862).
- Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143, 2074–2081. doi: 10.1037/a0038199
- Wuertz, D. (2013). *fOptions: Basics of Option Valuation*. Retrieved from <http://cran.r-project.org/package=fOptions>
- Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35, 151–175. doi: 10.1037//0003-066X.35.2.151

