Representational Principles of Function Generalization

Pablo León-Villagrá

Doctor of Philosophy

Institute for Language, Cognition and Computation

School of Informatics

University of Edinburgh

2020

Abstract

Generalization is at the core of human intelligence. When the relationship between continuous-valued data is generalized, generalization amounts to function learning. Function learning is important for understanding human cognition, as many everyday tasks and problems involve learning how quantities relate and subsequently using this knowledge to predict novel relationships. While function learning has been studied in psychology since the early 1960s, this thesis argues that questions regarding representational characteristics have not been adequately addressed in previous research.

Previous accounts of function learning have often proposed one-size-fits-all models that excel at capturing how participants learn and extrapolate. In these models, learning amounts to learning the details of the presented patterns. Instead, this thesis presents computational and empirical results arguing that participants often learn abstract features of the data, such as the type of function or the variability of features of the function, instead of the details of the function.

While previous work has emphasized domain-general inductive biases and learning rates, I propose that these biases are more flexible and adaptive than previously suggested. Given contextual information that sequential tasks share the same structure, participants can transfer knowledge from previous training to inform their generalizations.

Furthermore, this thesis argues that function representations can be composed to form more complex hypotheses, and humans are perceptive to, and sometimes generalize according to these compositional features. Previous accounts of function learning had to postulate a fixed set of candidate functions that form a participants' hypothesis space, which ultimately struggled to account for the variety of extrapolations people can produce. In contrast, this thesis's results suggest that a small set of broadly applicable functions, in combination with compositional principles, can produce flexible and productive generalization.

Acknowledgements

First and foremost, I have to thank Chris Lucas for his supervision and guidance throughout this work. From the day I applied for this Ph.D. to its maximum end date, he has always been available for discussion and support. His critical engagement with the shaky default priors in the newest hierarchical model, and patient questions about the *actual hypothesis* motivating my latest offbeat idea had a significant impact on my thinking and has made me a better researcher.

Second, I would like to thank all the people in the Forum that made the Ph.D. less soul-crushing and more fun: Nicolas, for the uplifting bar-plots, the terrace breaks, submission all-nighters, and all the encouragement and feedback writing this thesis. Joe for the sludge lunches, orange Scotts and football-chit-chats. Jenny, for teaching me how to say juice and cranachan and getting me out of the lab to get to know Scotland. Finally, everyone else that gave me the prospect of a MF1 2-hour break, or Teviot evening, especially Akash, Janie, Ludovica, Maria, Todor, and Yevgen.

During this Ph.D. I have been fortunate to work with other researchers and students – I have to thank Eric Schulz, for joining forces on exploring one-shot generalizations and being an example in persistence and diligence. Second, I'd like to thank all the students that I have been lucky to co-supervise in their projects.

I'd also like to thank Tiana for reading all of it and finding every dangling modifier, unclear antecedent, and round circle, and also for support and encouragement in these last stretches of the writing process.

Finally, I'd like to thank my family for providing me the space and time to embark on this journey — I swear I am done studying now!

To my family.

Table of Contents

Publications and Contributions

Parts of the research presented in this thesis have previously appeared in the following publications.

Conference Proceedings

- **Data Availability and Function Extrapolation**

  León-Villagrá, P., Preda, I., and Lucas, C.G. –*Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 2018

  Chapter 3 presents the results from this publication. The chapter expands on the model analysis and discussion.

- **Generalizing Functions in Sparse Domains**

  León-Villagrá, P. and Lucas, C.G.–*Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 2019

  The 3-point results in Chapter 5 were included in this publication. The chapter introduces new 1-point experiments and considerably expands the analysis.

- **Exploring the Representation of Linear Functions**

  León-Villagrá, P., Klar, V.S., Sanborn, A.N., and Lucas, C.G.–*Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 2019

  Chapter 4 presents the results from this publication. The chapter improves

the presentation of figures and discusses the results in the context of the thesis.

In Preparation

- **One-shot Learning of Compositional Functions**

León-Villagrá, P., Schulz E., Speekenbrink M,, Gershman S.J., and Lucas, C.G.

Chapter 6 presents the results from this work. The chapter re-examines the results obtained from a Bayesian estimation perspective and discusses the results in the broader context of the thesis.

Chapter 1

Aims of the Thesis

Recent years have seen a dramatic improvement in artificial intelligence most prominently driven by deep neural networks. Fueled by increased computational power, these models have come close to—or even surpassed—human performance for a range of domains, such as object recognition or speech recognition. Conceptually more important advances were achieved in complex dynamic planning and control tasks; artificial learning models have recently shown remarkable performance in complex game playing, beating the best human players in the game of Go (Silver et al., 2016) and poker (Brown and Sandholm, 2018), or achieving human-level performance in Atari games by training only on pixel data (Mnih et al., 2015). Notwithstanding, there remain human cognitive capabilities that are beyond the reach of current, state-of-the-art artificial intelligence. These capabilities include forming long-term plans, inferring deep causal relationships, or generalizing previous knowledge to new domains. More strikingly, humans exhibit these capabilities in severely underconstrained and sparse domains, while most algorithms require large amounts of data.

This sparsity is evident in early development when children first have to acquire basic cognitive and social skills. While early learning is slow (Mervis et al., 1992), at around two years of age, suddenly, learning is characterized by more

rapid improvements. One prominent example is the *shape bias*, where children after the age of two can generalize whole category structures from single category members (Samuelson and Smith, 2005), in so-called one-shot learning. This ability to generalize at an early developmental stage is remarkably similar to the generalization ability of adults. Like two-year-olds, adults confidently infer characteristics or constituents and generalize from, a concept or a physical scene that was presented only once (Vul et al., 2014).

The change of learning, from slow and error-prone to rapid leaps, and the ability to generalize given sparse data, suggests that crucial changes occur during early development. In contrast to human learning, state-of-the-art artificial learners exhibit constant but highly data-reliant learning, and the capability of the model to generalize from the learned data and task is limited (Lake et al., 2017). These differences suggest that human and artificial learners differ in the representation of learned data, their inferential capabilities underlying generalization, or both. Human-like generalization and learning adopts mechanisms such as *learning to learn* (Harlow, 1949), and operates upon *causal* (Ullman et al., 2014), *generative* (Feldman, 1997; Jern and Kemp, 2013), as well as hierarchical and compositional representations (Lake et al., 2015).

In this thesis, I explore how abstract compositional representations can facilitate generalization and suggest ways in which these models can be acquired through development and learning. To provide a tractable field of empirical examination, I focus on cases where a relationship between two continuous quantities must be learned. This type of learning, which amounts to regression from a statistical or machine learning perspective, is called *function learning* or *function estimation* in the psychological literature. While this domain might seem limited and artificial, function learning is a fundamental constituent of human learning, inference, and planning. It is crucial for domains as diverse as motor-learning (hit that three-pointer), scientific reasoning (calculating the orbits of celestial

bodies), or prediction (investing in the stock market, inferring the spread of a disease). As such, human function learning constitutes a domain of study that represents a wide range of human capabilities and is concrete enough for detailed computational analysis.

Function learning has been studied in psychology at least since Carroll (1963). While research in function learning has resulted in the description and modeling of particular human inductive biases and function-dependent learning difficulty, I suggest that many representational characteristics underpinning function learning and extrapolation are under-constrained.

In Chapter 2, I first discuss function learning in the broader space of human generalization. Then, in Chapter 3, I show that while previous research has emphasized learning differences for traditional and more modern experimental paradigms, underlying inductive biases and extrapolations are very similar when accounting for task-mediated memory demands. Through empirical and computational work, I show that in both paradigms hypothesis spaces over functions can be learned as an abstract, high-level encoding of the learned relations.

In Chapter 4, I explore the structure of these spaces in more detail. By adopting a novel experimental paradigm, I show that variability in training forms graded, and often multimodal hypothesis spaces.

Finally, in Chapters 5, 6 and 7, I evaluate how these hypotheses can be used. I suggest that abstract hypotheses over functions provide reusable models that can be transferred to perform extrapolation, even if data is sparse. Then, I present evidence that these hypotheses can be composed into more complex generative models, and that humans are perceptive to and generalize accordingly.

Chapter 2

Introduction

At the core of intelligence is the ability to learn and adapt. Every day, we face novel situations, entities, or objects, and every day, we need to solve problems to achieve our goals and, ultimately, survive. In all but the most artificial situations, learning amounts to generalization: memorization can only be adaptive if the *exact same* situation reoccurs. Generalization is the process of applying prior knowledge of similar events, objects, or relationships to novel situations.

This thesis discusses generalization in function learning tasks, where participants have to learn the relationship between two continuous-valued variables. Specifically, I discuss how people choose between different hypothesized functions and how they reuse and combine previously learned knowledge to form extrapolations. Since these questions parallel work in other research areas, such as categorization or rule-learning, and the results of this work are relevant for broader questions about knowledge representation and generalization, I will start by summarizing previous results in generalization research.

First, I will discuss how implicit, similarity-based, and explicit, rule-based hypotheses can account for human generalizations before discussing the structure and representation of the hypotheses themselves. Then, I will motivate function learning as an essential field of study for developing an understanding of hu-

man generalization and summarize the main results of function learning research. Finally, I discuss open questions that motivate the work in this thesis: abstract knowledge in human function extrapolations and the structure and representation of the hypothesis space underlying function learning.

2.1   The Basis for Generalization

On which basis are experiences judged as similar, and what types of knowledge can be generalized? Traditionally, two opposing answers have been put forward: generalization is performed based on a perceived similarity between the task at hand and previous knowledge, or generalization is the application of a learned rule. In the next paragraphs, I will first summarize the main ideas behind similarity-based approaches to generalization. Then, I will briefly present research on rule-based and structure-based generalization.

Pioneering work by Shepard (1987) provided strong evidence for psychological distance as the basis for category generalization. Imagine a child learning that the fluffy, sleepy creature in the house is a cat. When the child subsequently encounters other fluffy creatures, she will generalize their categories based on the similarity to the previously met cat. Formally, the child has to learn a function that implicitly maps from perceptual features to categories; for example, see Figure 2.1. Shepard's law postulates that both the cat and new animals are represented in a psychological metric space. In this space, both known and new animals are represented as points, and their similarity is a function of the distance between these points. Shepard argued that this approximately-exponential function was universal and showed that it applied to a broad range of generalization domains, and sensory modalities, both for humans and a wide range of other animals.

Generalization, as a function of psychological distance, is a core constituent
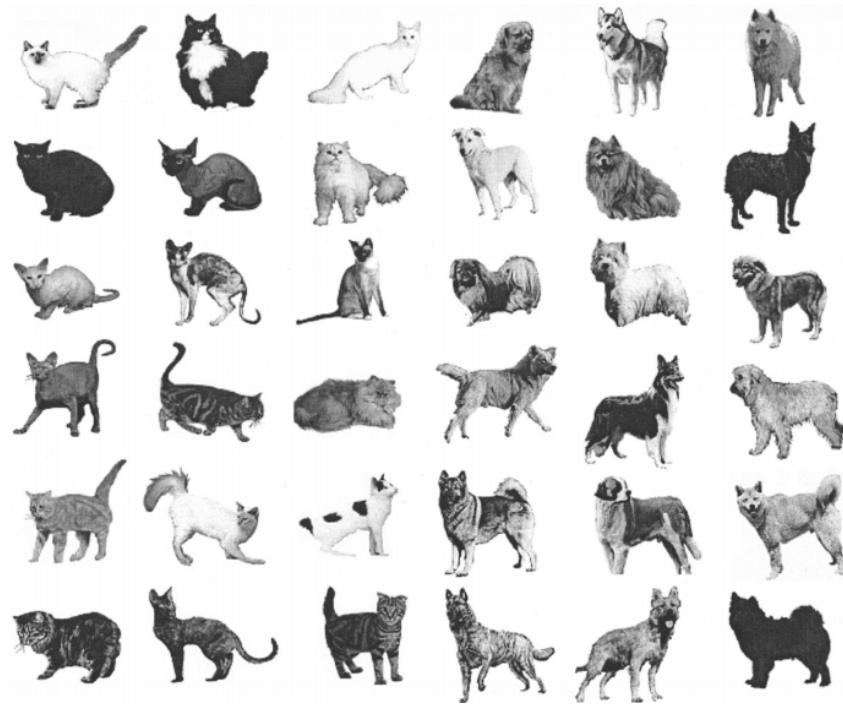
Figure 2.1: Figure reproduced from French et al. (2004). Natural categories can vary widely in their category members and for some, membership might not be clear. For example, experiments by French et al. (2004) showed that 3-4 month-old infants produced asymmetric classification errors for cat and dog categories: they included dogs within cat categories, but did not misclassify cats as dogs.

of many categorization models. Most prominently, exemplar models of catego-
rization assume an underlying representation space and distances as the basis
for generalization gradients (Erickson and Kruschke, 1998). Subsequent research
has expanded Shepard's rule of generalization, applying it to arbitrary repre-
sentations, allowing generalizations from multiple instances, and characterizing
its mathematical foundations more formally (Ashby and Alfonso-Reese, 1995;
Tenenbaum and Griffiths, 2001; Jäkel et al., 2008).

Instead of implicitly inducing a category from its representation in psycho-
logical space, early research in cognitive science focused on situations in which
highly structured, logical combinations of features determined category member-
ship (Bruner et al., 1956). Repeatedly seeing cats being *fluffy and sleepy* would
allow the child to learn the rule *fluffy + sleepy = cat.* While this approach is
unsatisfactory for the vast majority of categories — try listing necessary and suf-
ficient conditions for any non-artificial category — people can learn and apply
these rules. They often shift between similarity-based and rule-based inference
depending on the task, and people generally exhibit individual differences (von
Helversen and Rieskamp, 2009; Pachur and Olsson, 2012; Little and McDaniel,
2015). For an example of a category for which laypeople might have implicit
knowledge, but with experience might learn rule-like patterns, see Figure 2.2.

From a probabilistic perspective, learning about categories, either in terms
of psychological distances or distributions over category-specific rules, amounts
to updating prior beliefs over possible categories or rules. For example, seeing
a hairless Sphynx cat will result in an updated belief in which *fluffiness* is less
pronounced as a feature, or in which the distance between cats and dogs are
less disparate. Features of the training, such as the number and diversity of the
examples, or the overlap between categories, inform the resulting generalizations
(Hahn et al., 2005; French et al., 2004; Osherson et al., 1990; Hayes et al., 2019).
In Chapter 4, I will present a first experiment uncovering whether people can

Figure 2.2: Given the left cloud pattern, which of the three patterns of the right belong to the same classification? One pattern is commonly associated with approaching storms, the other three with stormy weather. The left pattern and the first and last of the three candidate patterns are nimbostratus clouds that usually produce continuous rain or snow. The second candidate pattern depicts altostratus clouds, which often precede approaching storms. For many categories, laypeople may lack explicit knowledge about defining features of the category. However, even without this knowledge people can perceive similarities and generalize flexibly.

perceive and represent the diversity of training in a function learning setup.

Following Tenenbaum and Griffiths (2001), we can express these prior beliefs as a *hypothesis space*, $\mathcal{H}$, either over the geometry of categories in psychological space (for Shepard consequential regions) or, for rule-based category learning, as hypotheses over rules and their constituents (Goodman et al., 2008). For example, the hypothesis space for cats might express that both *fluffiness* and *sleepiness* are likely features of cats. It would also express an inductive preference for conjunctive feature combinations, especially for *fluffy and sleepy*.

Discussing the structure of the hypothesis space is useful when considering the basis of generalization: how does one select a generalization from the infinite space of alternatives? For example, categories for any collection of encountered examples can be made arbitrarily specific. Consider for instance the set of numbers $\{8, 10, 12, 14\}$: *even numbers from 8 to 14* and *the numbers 8,10,12,14* are both equally accurate hypotheses. Due to the finite nature of the example and the domain of numbers, the set of alternatives that one would consider is fairly small.

However, for any set of natural categories, the number of potentially relevant features is infinite[1]. The Bayesian approach to inference provides an elegant answer to this issue. If several hypotheses are equally likely a priori and are consistent with the evidence, the more specific hypothesis will be favored. Since more flexible hypotheses cover a greater variety of observations, not observing these data provides negative evidence. Thus, in a Bayesian framework, complexity does not have to be explicitly penalized but can be accounted for implicitly by assuming human inductive priors (Tenenbaum, 1999). Furthermore, if the prior over hypotheses is not flat, but some inductions are favored, these biases can constrain the space of alternatives to consider. This allows for stronger generalizations, as more mass is concentrated on the a-priori favored hypotheses. These inductive biases have been studied widely in cognitive science and have been proposed as the source for children's rapid learning and far-ranging generalizations of word meaning, causal inference, or category induction (for an overview, see Griffiths et al., 2010). In Chapters 5 and 7, I will present the results of experiments discussing how people balance the complexity of the hypothesis against instruction and prior biases when extrapolating learned continuous patterns.

2.2   Representation and the Target of Generalization

We can distinguish different forms of generalization by closely examining the target of generalization. If the target is a discrete category label, the generalization amounts to applying a previously learned mapping to a novel situation and assigning that instance the most consistent category label. We can abstractly state that inference as $p(y = c | \mathcal{H}, x)$, that is, the probability of the new observation y

---

[1]Even in the number game, one can always construct ad-hoc hypotheses, such as *the number of chicken fingers one can order at the local chicken shop*, or *the lucky numbers I drew in a fortune cookie.*

belonging to category c given previous knowledge x and a hypothesis space $\mathcal{H}$ (for example, see Sanborn et al., 2006). However, if the target y is a continuous quantity, for example, $y \in \mathbb{R}$, the task amounts to function learning.

Another type of inference closely related to generalization is analogical transfer. Research in analogy or transfer tends to examine problem-solving, where information about one domain has to be generalized to a new domain. While this task is very similar to generalization, research in analogical reasoning and transfer usually uses less obvious, often structural similarities, and the domains or tasks between which one has to generalize are often more disparate (Gick and Holyoak, 1980). Structural analogies, similar to rule-learning, require the problem solver to detect a shared underlying structure between two or more instances; for example, deducing "An electric battery is like a reservoir"(Gentner, 1983).

Similar ideas have been discussed from a probabilistic point of view, expanding the notion of what is represented in the hypothesis space, and how learning operates. First, many of the problems we encounter elicit particular representations, and these representations might be crucial for the flexible ways we can generalize. For example, the representation of colors is widely assumed to be ring-like. In contrast, biological kinds are often represented in hierarchical, often tree-like structures, whereas social and political relationships often form graphs, with edges between nodes signifying interrelations between the actors. Other relationships are directed; for instance, causal relationships are often represented as directed graphs, where edges are causal relations. Even for seemingly simple categorization tasks, the representation is crucial; for example, see Figure 2.3. A theory of learning and generalization thus has to explain how these structures can be inferred and applied. Computational models of probabilistic structure learning have shown that structures, such as trees, rings, and graphs, can be inferred from data (Kemp and Tenenbaum, 2008; Lake et al., 2018).

This form of learning can be seen as learning about the hypothesis space

Figure 2.3: Materials from the Omniglot dataset (Lake et al., 2019). Lake et al. (2015) showed that people infer rich, generative models of characters given only few examples and generalize these patterns to new, unseen instances. Given the characters in the left box, which other characters belong to the same alphabet? Which characters belong to a fictitious alphabet? Can you locate the alphabets geographically? Would you be able to produce new characters of the alphabet given these examples? A model proposed by Lake et al. (2015) accounted for these rich behavioral patterns by inducing generative programs over *stroke sequences*, as opposed to the commonly adopted *pixel-based* representations.

itself: what is the structure of the domain, and how do features or entities combine? Knowledge about the structure of the hypothesis space is often referred to as *overhypotheses*, or hypotheses over hypotheses themselves (Goodman, 1983; Kemp et al., 2007). These abstract learning processes have been pointed out as instrumental for learning early in development (Xu et al., 2009; Kemp et al., 2007), and can sometimes precede more concrete learning, the so-called *blessing of abstraction* (Gershman, 2017a). Motivated by these results, I will examine overhypotheses in Chapters 6 and 7.

2.3 Generalization and Transfer in Function Learning

These are strange times to motivate function learning. With the coronavirus pandemic raging at the time of writing, time-series data and model predictions are discussed daily in the news. Figure 2.4 displays one such time-series: daily death-rates for the United States up to the 4th of May 2020. Function learning tasks amount to learning a mapping between two (or more) metric quantities, the predictor(s) and the outcome(s). In this case, the predictors are days from early March until the 4th of May, and the outcome is the daily death rate. Two types of generalizations are commonly discussed in function learning: interpolation and extrapolation. Interpolation amounts to predicting outcome values for predictor values within the range of known values, in this case, predicting death rates for a day up to the 4th of May. On the other hand, extrapolation amounts to predicting values outside the range of known values, in this case, values from the 5th of May onward. It requires no further motivation to see the immense importance of these predictions for policy, science, and everyday decision making.

Given the importance of accurate extrapolations, how can we predict, and given predictions, how can we judge their quality? Computationally, extrapolation consists in determining new values y_{n+1} for test values x_{n+1}, conditional on previously learned $\mathbf{x}_n, \mathbf{y}_n$ and a prior belief $\mathcal{F}$ over possible functions. The type of prediction then crucially depends on the prior belief over functions.

Once predictions are derived, their quality can be evaluated by measuring how closely interpolations match the seen data, and, once new data arrive, how well predictions match the new information. In Figure 2.5, we can see two predictions for the coronavirus data. The predictions in Figure 2.5a were created by a senior advisor to the White House and are (likely) predictions of a 3rd-order polynomial, $y = \sum_{i=0}^{3} \beta_i x^i$ (Washington Post, 2020). The model predicts an extreme decline of
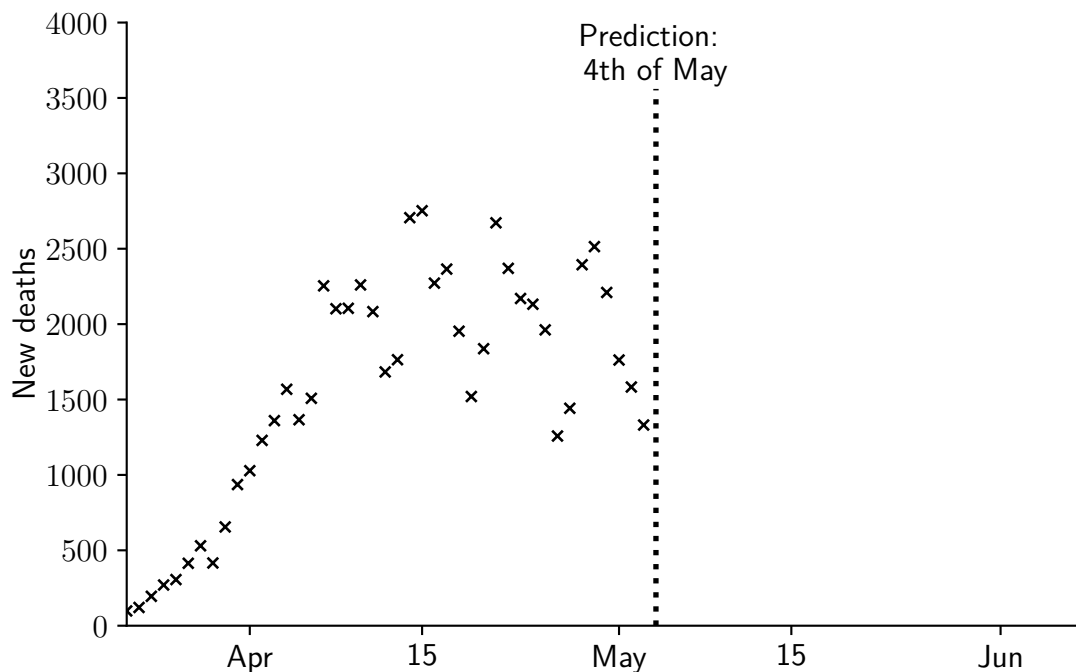
Figure 2.4: Daily coronavirus death rates in the United States (accessed from the New York Times on 9th of June 2020).

daily deaths (in fact, it predicts negative deaths past 15th of May). In Figure 2.5b, we can see the predictions of a different model, a Gaussian process (for a short introduction to Gaussian processes, see Appendix A). This model also predicts a decline in deaths, but at a more gradual rate. Furthermore, both models express different views of the underlying data-generating process — the Gaussian process explicitly assumes that the data exhibits smooth changes with additional periodic fluctuations. In contrast, the cubic model has no such assumption. In Figure 2.5c, we can see that the extreme drop in cases predicted by the cubic model has not occurred. Instead, the number of deaths has slowly decayed with some additional fluctuation[2].

---

[2]This is not to say that the Gaussian process captures the complex underlying causal process. Both the cubic model and the Gaussian process do not account for populations, their demographics, or contagion factors. Both models thus are of little use for evaluating policy interventions. However, the Gaussian process allows us to express prior assumptions, and derive posterior estimates for abstract features of the function, such as the extent of short-term changes, or the long-range decay of case-numbers.
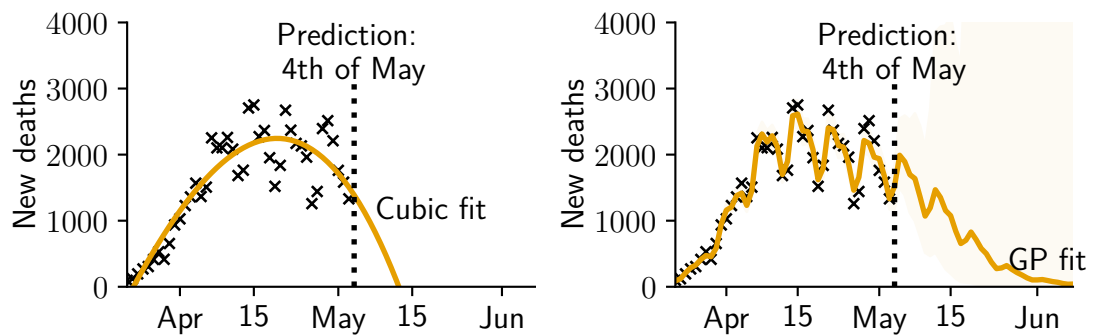
(a) Predictions of a cubic function.

(b) Predictions of a Gaussian process.
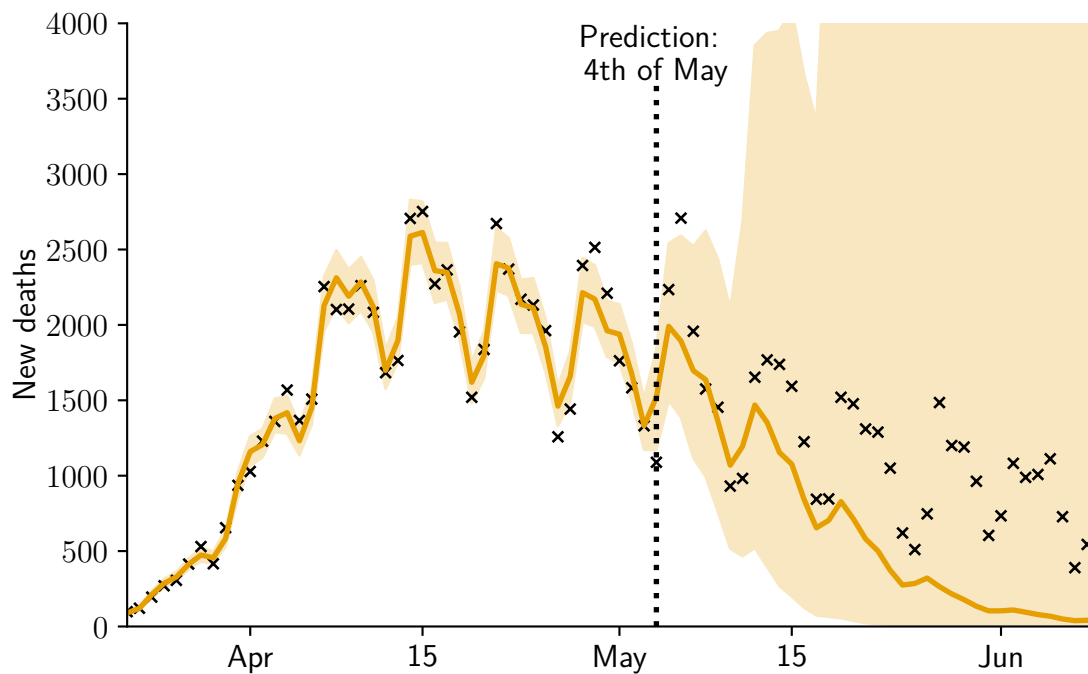
(c) Actual daily deaths (reported at the time of writing) and the median predictions, and 90% uncertainty intervals of the Gaussian process.

Figure 2.5: The predictions of a cubic function and a Gaussian process, as well as the actual daily deaths.

We have seen how function extrapolation crucially depends on the assumptions one makes about the data-generating process, and how different functions can produce widely different predictions. While the development of better predictive models is of fundamental importance for science and policy, the interest of this thesis is in how laypeople perceive and learn relationships in data, and how they generalize these relationships to new data.

Human function learning research has a much shorter and sparser history than categorization research. Research into human function learning started with work by Carroll (1963), that showed that participants could learn functions from data and extrapolate to new values, see Figure 2.6. This work also established that functions were easier to learn than random patterns, and linear functions were easier to learn than non-linear functions. Subsequent studies have confirmed these
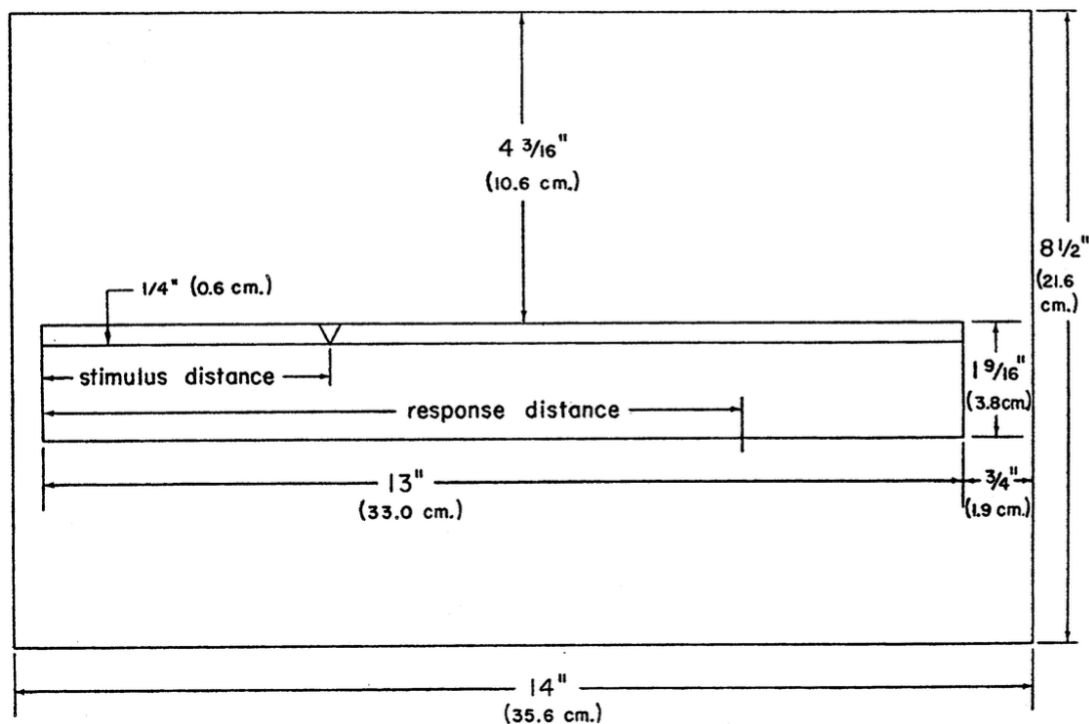

Figure 2.6: Experimental setup adopted in the first study of function learning. In Carroll (1963), participants received test-books and had to learn the relationship between stimuli (V mark on the mid-left) and response distances. Afterward, participants had to predict the response magnitude for a series of old and new stimulus magnitudes.

early results. People learn linear relationships more quickly (Brehmer et al., 1985; Byun, 1995) and have difficulty making non-linear and especially non-monotonic extrapolations (Brehmer, 1974; Brehmer et al., 1985; Byun, 1995; Kalish, 2013). These strong inductive biases for linearity have also been found in control experiments. Berry and Broadbent (1984) found that participants tasked to control a complex dynamical system struggled when the underlying dynamics were exponential, rather than linear.

When extrapolating, people show a strong bias toward inferring linear functions, particularly linear functions with matching x and y values. As an illustration, consider the aggregate results obtained from DeLosh et al. (1997) in Figure 2.7. While mean interpolations for linear, exponential, and quadratic functions match the training range (center of the plots), extrapolations deviate in informative ways. For linear functions, the extrapolations underestimate the slope, whereas exponential and quadratic extrapolations suggest linear patterns.
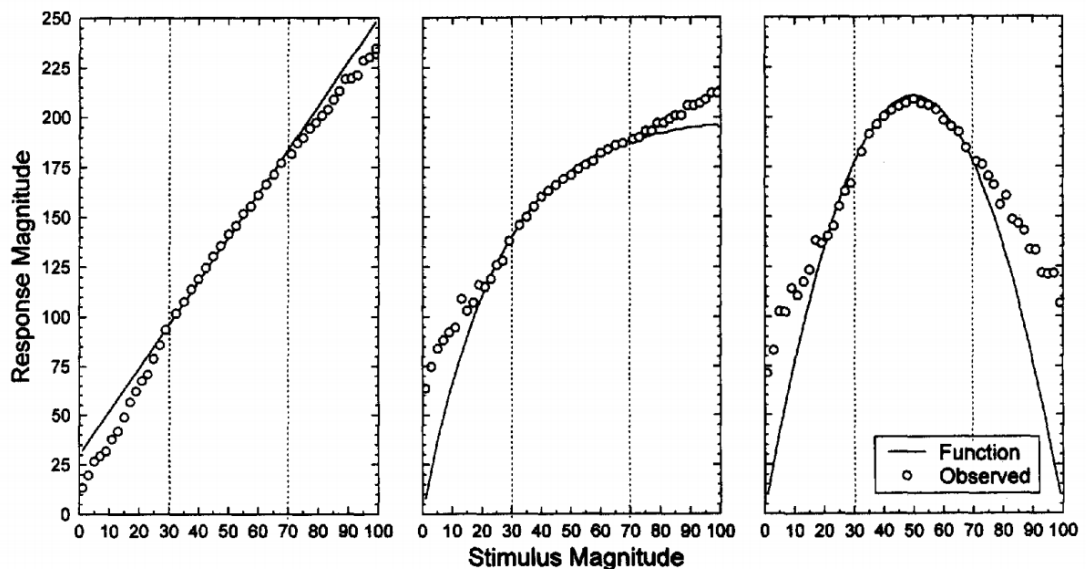

Figure 2.7: Figure reproduced from DeLosh et al. (1997). Mean interpolation and extrapolation results in DeLosh et al. (1997) suggest that participants often extrapolated linearly, even if the training data and their interpolation judgements were non-linear.

These results have been widely reproduced (Brehmer, 1971, 1976; DeLosh

et al., 1997; Kalish et al., 2004; Brown and Lacroix, 2017; Kwantes and Neal, 2006). Moreover, the strong preference for linear functions has also been found in iterated learning experiments. In iterated function learning experiments, participants obtain extrapolations of previous participants as training. Kalish et al. (2007) showed that these iterations quickly converge to linear patterns, even if the initial training was negative, U-shaped, or random (see Figure 2.8).
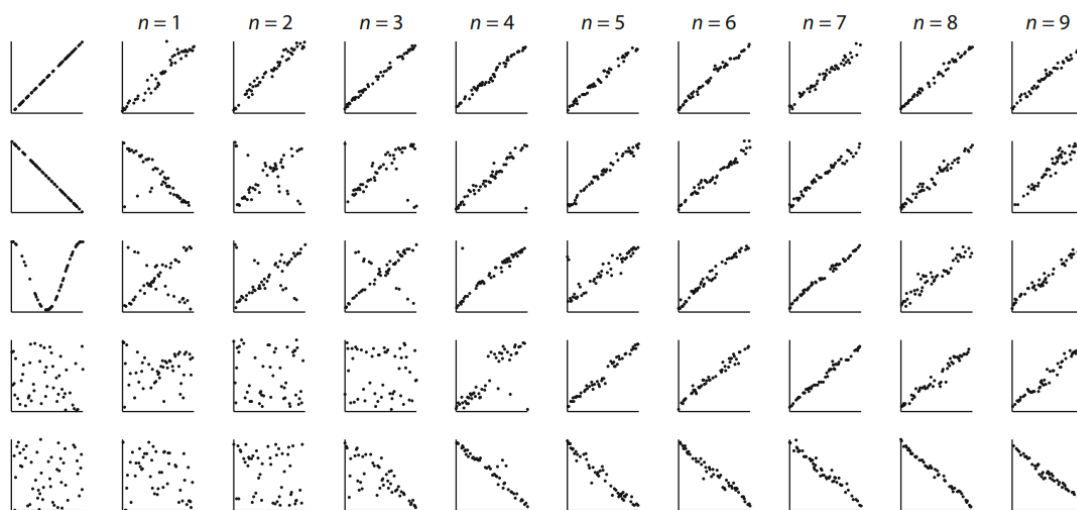
Figure 2.8: Figure reproduced from Kalish et al. (2007). In iterated learning experiments, participants receive extrapolations performed by previous participants as training data. Experiments by Kalish et al. (2007) showed that participants quickly transition to positive linear functions, even if the original training was not linear, or even random.

These results have led to the development of models that attach a special representational status to linear relationships (Kalish et al., 2004), or assume that people have a strong inductive bias favoring linearity (Brehmer et al., 1985). These models are typically evaluated by comparing their predictions to averaged human judgments, either via direct correlations, relative error rates, or qualitative features, such as single or multiple modes in judgments (Kalish et al., 2004) or non-monotonicity (Bott and Heit, 2004; Kalish, 2013). While previous research has consistently shown that people are biased toward linearity, I suggest that

these biases can be overruled by experimental paradigms (Chapter 3) or task instructions (Chapters 5 and 7).

2.3.1 Experimental Paradigms

The experiments presented so far all followed the classical experimental paradigm in function learning. Participants learn relationships from sequentially-presented pairs of points and then, after enough training, have to either interpolate or extrapolate from the learned data. At no point do participants see the whole data set, instead they have to learn the relationship implicitly.

More recent experiments have presented all data simultaneously, in setups resembling scatter plots. These experiments are far less taxing in terms of memory demands, and all data is readily available to the participant. For a comparison of the two approaches, see Figure 2.9.
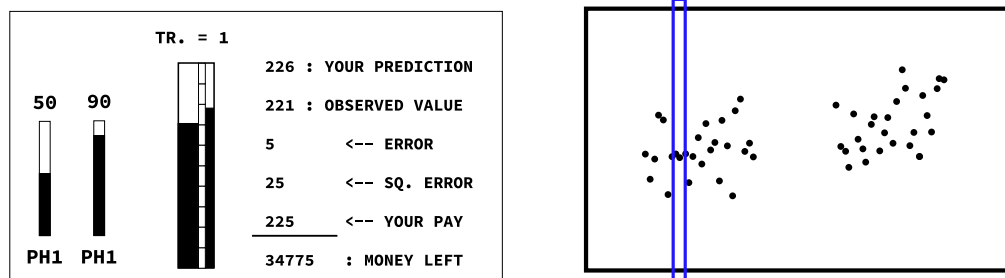
(a) Setup in Busemeyer et al. (1993).   (b) Setup in Little and Shiffrin (2009).

Figure 2.9: Examples of the two paradigms in function learning. In Busemeyer et al. (1993) the two bars on the left indicate the two predictive variables, the larger bar on the right the predicted and observed dependent variable. In this particular experiment, all quantities are also represented numerically, and the participant receives feedback on their accuracy. In Little and Shiffrin (2009) participants receive all training data, presented as a scatter plot. They then have to predict the values of the dependent value, the currently probed value highlighted by the blue rectangle.

Experiments in the scatter plot paradigm have demonstrated that human

learners can discover and extrapolate a larger set of relationships, including complex non-linear trends (Wilson et al., 2015; Schulz et al., 2017; Lucas et al., 2015; Little and Shiffrin, 2009). However, a crucial question is if both these paradigms tap into the same sort of inductive biases and if both paradigms test the same cognitive capacity. In Chapter 3, I directly compare these paradigms and find that not the presentation, but the memory demands of the task, significantly influence learning.

2.3.2   Models of Function Learning and Generalization

Models of human function learning can be divided into rule-based, similarity-based, and hybrid. Rule-based models postulate that humans learn explicit functions (such as polynomials) by finding the best fitting function and parameters from a small set of functions (Carroll, 1963; Brehmer, 1971, 1974). These models can accommodate diverse patterns of extrapolation by specifying flexible rules as hypothesized functions. However, standard rule-based models cannot explain the human ability to learn and reproduce arbitrary functions, since all applicable rules must be incorporated a priori.

In contrast, similarity-based models do not assume an explicit underlying function, but instead, suppose that humans learn to associate pairs of values. These models capture the flexibility with which human learners interpolate. However, similarity-based extrapolations tend to be linear or converge to a constant value (Busemeyer et al., 1997), even when human learners extrapolate in non-linear ways.

Given the complementary strengths of both approaches, hybrid models have been suggested. These models tend to incorporate the ability to extrapolate according to any of a fixed ensemble of rules while retaining similarity-based models' flexibility in interpolation (Kalish et al., 2004; McDaniel and Busemeyer, 2005).

More recently, Gaussian processes have been proposed as a unifying rational model that can represent both similarity and rule-based approaches while explaining human behavior on various tasks (Lucas et al., 2015). Furthermore, these models have shown promise as a means for reverse-engineering the inductive biases that human learners use to extrapolate from arbitrary and sometimes complicated relationships, capturing patterns of learning that cannot be accounted for with traditional similarity or rule-based approaches (Wilson et al., 2015).

2.3.3 Open Questions in Function Learning and Outline of the Thesis

While some previous studies have examined patterns in individuals' inferences (Wilson et al., 2015; Schulz et al., 2017; Kalish, 2013), these studies still neglect essential questions about the representations and tacit beliefs behind participants' judgments. For example, while models that take a distributional approach to function learning have successfully explained human behavior, there is little direct evidence that people track uncertainty or variability when faced with function learning problems. There are also open questions about individual differences, as most analyses have relied on aggregated judgments, or assume that individual inductive biases are broadly similar (Kalish et al., 2007). In Chapter 4, I present a first experiment examining if people track the variability of training.

A second issue regards how the hypothesis space over functions and the individual functions themselves are represented and learned. One possibility is that humans represent functions as parametric forms (Brehmer, 1974). A parametric representation allows very efficient storage of the learned relation, as only a few parameters have to be maintained. However, these approaches suffer from a lack of flexibility of learning and also raise the question about which parametric forms can be learned. On the other hand, the more popular associative approaches are liable to criticisms about memory constraints, since, in principle, all data points

are required to interpolate or extrapolate. While it is possible to introduce no-
tions of memory decay, much like in exemplar models of categorization, precise
decay mechanisms have not been suggested. In Chapter 3, I contrast conventional
experimental paradigms in function learning and examine the effect of memory
demands on generalization.

A related question is how the hypothesis space can be modified and updated
in light of new evidence. A common characteristic of all computational models is
that they rely on specific families of parametric relationships, extremely flexible
one-size-fits-all inductive biases, or both. As a result, they tend to be unable to
explain the human ability to extrapolate in tremendously varied ways, many of
which can be expressed as simple rules (Lucas et al., 2012). While overhypotheses
have received ample attention in categorization research in recent years, function
learning has mostly approached the hypothesis space as a rigid ranking of alter-
natives. One notable exception is the recent work by Schulz et al. (2017). In
several experiments, Schulz et al. (2017) showed that participants could learn the
compositional structure inherent in relational patterns. Participants preferred
compositional over non-compositional patterns and extrapolated in ways better
described by a compositional model. Finally, participants could also remember
compositional patterns better than non-compositional alternatives, suggesting
that participants represent these functions more efficiently. In Chapters 6 and
7, I will return to the question of how functions in the hypothesis space can be
learned. I argue that people perceive and often apply compositional principles to
form complex functions, but that this ability is less general than the results in
Schulz et al. (2017) suggest.

Finally, while function learning research has focused on learnability and ex-
trapolation, it is plausible that people can perceive and exploit deep and abstract
similarities between learned relationships. For instance, take the graphs presented
in Figure 2.10. All eight patterns correspond to infection cases for the Spanish flu

between 1918 and 1919 for eight cities in the UK, as collected by He et al. (2013). While all patterns are characteristically different in their absolute numbers and vary slightly in the exact onset and offset of their peaks, the overwhelming perceptual impression is of a three-peak pattern. This perceived similarity is more than just a mere visual impression. It can be the basis for more profound causal inference, trying to uncover the latent structure that results in such a widespread repeated pattern. These perceived structural similarities can be used to inform current policies, even when the context varies somewhat. For instance, convincing arguments for lock-downs, social distancing, and "flattening of the curve" were made based on case numbers of US cities 100 years earlier, for a pandemic of a related, but potentially very different virus (National Geographic, 2020). In Chapters 5 and 7, I explore how people use knowledge about a shared generative process to produce extrapolations.

This thesis will extend work in function learning by examining hypothesis spaces more closely. I will treat function extrapolations as generalization tasks. I refer to more abstract learning as transfer learning. An instance of transfer learning could be learning about the type of function applicable in a particular domain, or the compositional structure inherent in a task. While this notion of transfer is much more specific than the types of transfer discussed in analogical reasoning experiments, this distinction is intended to separate extrapolation or generalization in one particular task, from learning at the level of the hypothesis space and transferring this abstract knowledge to a subsequent task. I will approach the functions themselves from a unifying computational perspective, using Gaussian processes as a modeling tool.
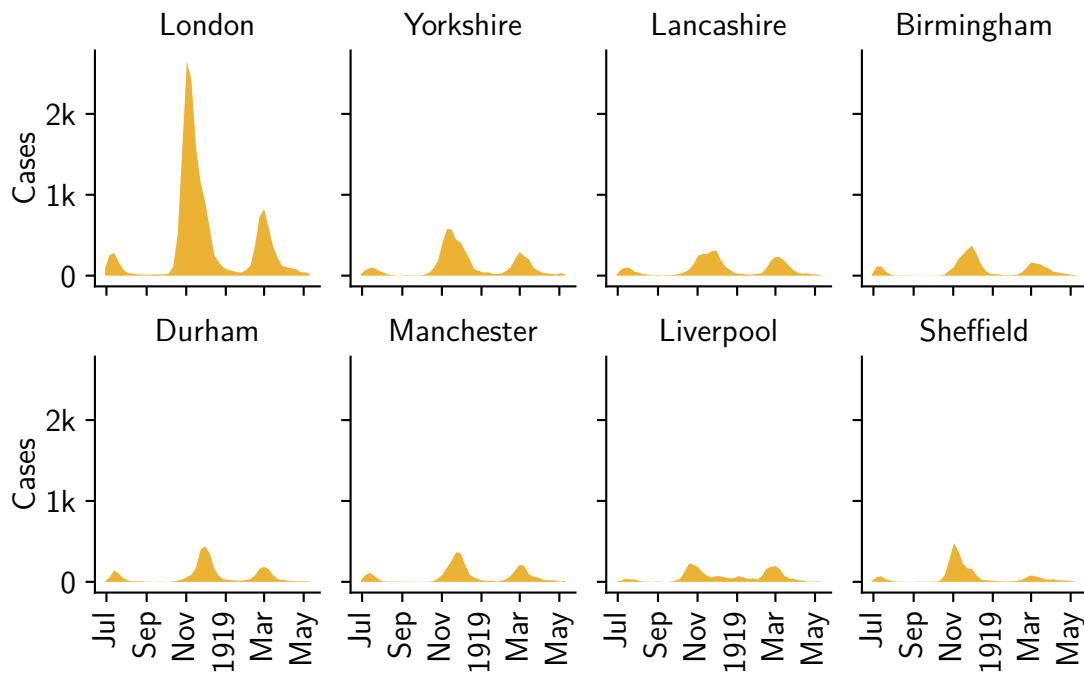
Figure 2.10: Data collected by He et al. (2013) presents the number of infected individuals with the Spanish flu for UK cities and regions. While all eight patterns presented here, and all other regions in the dataset, exhibit different absolute case numbers and slight variations in the onset and offset of infections, the overall pattern of three main outbreaks is striking.

Chapter 3

Function Representation and Generalization

In this chapter, I examine how experimental presentation affects function learning, and what role memory requirements of previous function learning tasks play in the subsequent function extrapolation.

Previous experiments have painted a mixed picture of the human ability to generalize functions. Early work where participants had to learn relationships from sequentially-presented examples has highlighted strong biases for linear extrapolations. Moreover, cyclic or non-smooth relationships could only be inferred after prolonged training with visual aids such as ticks or numeric values. These results have led to the development of models that attach a special representational status to linear relationships (DeLosh et al., 1997; Kalish et al., 2004), or assume that people have a strong inductive bias favoring linearity (Brehmer et al., 1985). We will refer to experiments following the sequential paradigm as *function learning* tasks. In contrast, more recent work, in which participants are presented all data simultaneously, which we will call *function estimation*, has shown that human learners can discover and extrapolate from complex non-linear trends (Wilson et al., 2015; Schulz et al., 2017; Lucas et al., 2015).

How do we reconcile these results? One possibility is that participants generalize in these experiments in different ways, for reasons that may be perceptual, cognitively innate, or experience-dependent. An alternative possibility is that the same inductive biases and cognitive processes underlie both paradigms, and differences between these tasks can be attributed to differences in their memory demands.

In function learning experiments, participants have to maintain learned data in memory and update and evaluate the appropriateness of a representation against alternatives. In contrast, function estimation allows an effortless recall of the data. Given that only a subset of the data is maintained, extrapolations will resemble inductive biases in the absence of data. In contrast, having all data visually available, as in function estimation, allows to counteract inductive biases and facilitates extrapolations resembling richer functions.

3.1   Experiment

We set up an experiment to contrast extrapolations in function learning and function estimation. To distinguish experimental presentation from memory requirements, we introduced a new condition that shared presentational-, but not memory-related characteristics with function estimation. In this new condition, data were presented as scatter plots, but data points disappeared from display immediately after the participant submitted her choice. Since the condition exhibits similar characteristics to classical function learning tasks, we predicted that extrapolations should more closely resemble function learning conditions, as participants will have to rely on the recollection of the presented data for their extrapolations. We will refer to the scatter plot condition presenting the full data as Scatter$^+$ and the new condition as Scatter$^-$. We will refer to the traditional function learning conditions as Bar.

In both scatter plot conditions we did not connect the presented data points, as previous research has shown that line plot presentations induce stronger biases towards inferring sequential dependencies (Theocharis et al., 2019).

3.1.1 Participants

We recruited 322 participants via Amazon's Mechanical Turk service[1]. Participants received \$0.40 for participation and took an average of eight minutes to complete the experiment ($M = 8.01$, $SD = 4.38$). Participants had to have completed more than 50 approved tasks with an approval rate of 95% or higher. Participants were randomly assigned to one of the nine conditions $\{f_{\text{lin}}, f_{x^2}, f_{\cos}\}$ $\times$ $\{\text{Scatter}^+, \text{Scatter}^-, \text{Bar}\}$, as described below.

3.1.2 Materials

The data presented in the experiment was generated by one of three functions: linear (f_{lin}), quadratic (f_{x^2}), or periodic ($f_{\cos}$). These functions were chosen to allow for informative error patterns resulting from human inductive biases. Since previous research reported strong biases for linear functions with zero intercepts and 1/1 slope (we will refer to this function as $f(x) = x$), we selected a shallower positive slope. The quadratic function was relatively flat in the training block to test if participants would revert to linearity or choose non-linear alternatives. Finally, a periodic function was used to evaluate if participants were able to extrapolate in non-monotonic fashion. To allow space for extrapolation beyond the function ranges, we normalized the data to span $(0,1)$ in both the x and y axis and then rescaled and centered the graph such as to span half the y-axis. For the full set of materials after transformation, see Figure 3.1.

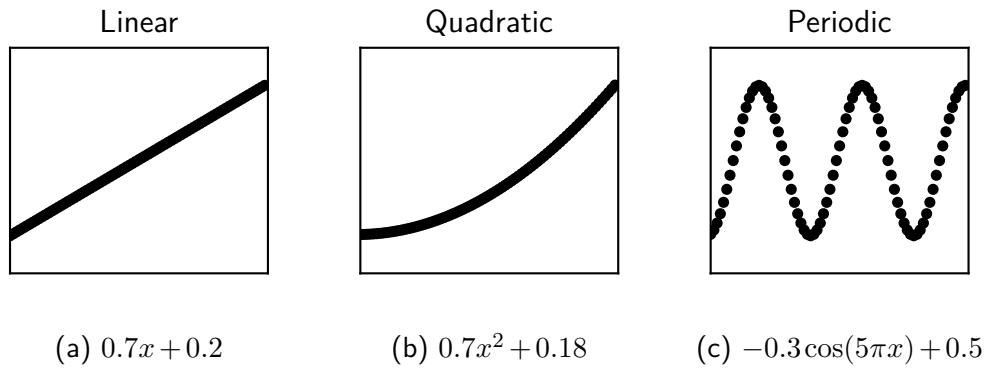[1]We did not collect any information on gender or age in this experiment.

Linear	Quadratic	Periodic
(a) $0.7x + 0.2$	(b) $0.7x^2 + 0.18$	(c) $-0.3\cos(5\pi x) + 0.5$

Figure 3.1: Three functions, f_{lin}, f_{x^2}, f_{cos}, generated the underlying data.

3.1.3   Procedure

For the full experimental procedure, see Figure 3.2. Participants were instructed that they would be presented with data and that they had to predict new values given their understanding of the relationship in the data. Then, participants proceeded to a block of training trials (the training block).
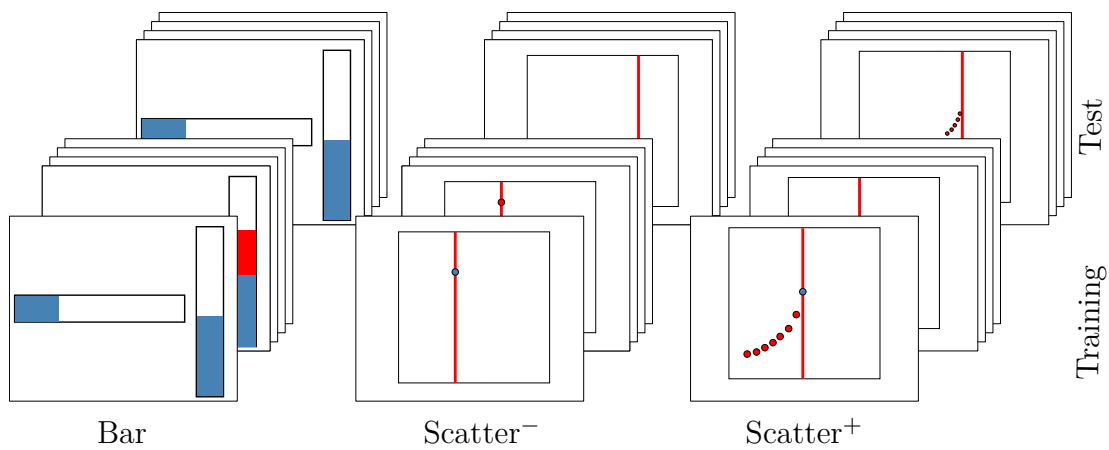

Figure 3.2: Participants were randomly assigned to one of the nine experimental conditions. All participants performed a training block consisting of 40 value pairs with feedback followed by a test block of 40 extrapolations without feedback.

3.1.3.1 Training Block

In the Scatter$^+$ and Scatter$^-$ conditions, the current test value (x) was marked with a red line spanning the whole vertical range. Participants were prompted to select a y value by clicking on the line. Once selected, the input value was highlighted with a blue point. Selected points could be updated by re-selecting a y value.

The selected values were submitted by pressing the space key. In the Bar condition, current x values were presented as the width of a bar on the left of the screen, and participants selected values by choosing the height of a bar on the right. As in the Scatter$^+$ and Scatter$^-$ conditions, participants could readjust these values. In all conditions, x-values were presented sequentially in ascending order. If the selected y value was within the error margin (± 0.05 of the true y), the true value was shown in red for 600 milliseconds. Afterward, a message indicating that the choice was correct and the remaining number of trials was shown.

If the selected value was not inside the margin, the message indicated an unsuccessful submission. Then, the selected value was removed, and participants had to resubmit. After erroneous submissions the true y was displayed as a red bar (Bar) or a red dot (Scatter$^+$, Scatter$^-$). Participants had to resubmit values until an admissible y was chosen. Participants received 40 points in total during the training block.

3.1.3.2 Test Block

The test block followed the same procedure as the training block, but no feedback was provided. After submitting 40 values in the test block, participants concluded the experiment by completing an optional short survey. For screenshots of the experimental stimuli and instructions, see Appendix B.

3.2   Results

3.2.1   Functions and Presentation Form

Consistent with previous findings, mean absolute error (MAE) in the test block was largest for f_{cos} ($MAE = 0.24$, $SD = 0.1$, $n = 108$). Errors for f_{x^2} and f_{lin} were small, with f_{lin} exhibiting the smallest error ($MAE_{x^2} = 0.14$, $SD_{x^2} = 0.09$, $n_{x^2} = 106$, $MAE_{lin(x)} = 0.11$, $SD_{lin(x)} = 0.1$, $n_{lin(x)} = 108$).

The errors in the presentation conditions were compatible with our hypothesis, with Scatter$^+$ lowest ($MAE = 0.1$, $SD = 0.1$, $n = 106$), and Scatter$^-$ and Bar at similar, higher levels ($MAE_{Bar} = 0.19$, $SD_{Bar} = 0.12$, $n_{Bar} = 110$; $MAE_{Scatter^-} = 0.19$, $SD_{Scatter^-} = 0.11$, $n_{Scatter^-} = 106$). For all errors in the subgroups of function and presentation conditions, see Figure 3.3.
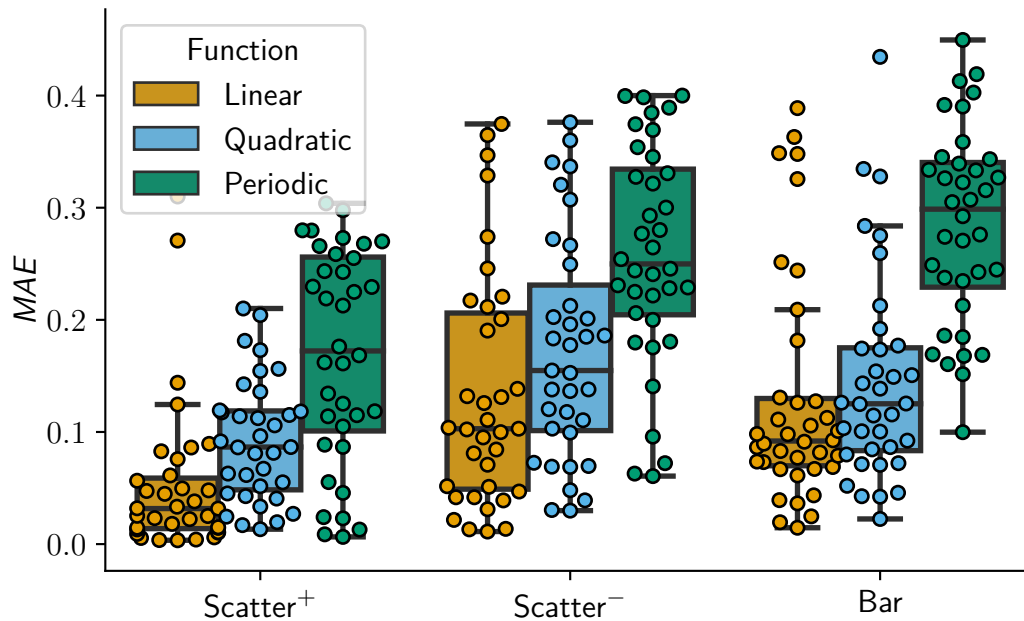

Figure 3.3: In all presentation conditions, participants exhibited the lowest errors for linear, followed by quadratic and periodic functions. Boxplots display first, second (median) and third quartiles. Whiskers show the ± 1.5 interquartile range (IQR). Each point represents the MAE of one participant.

3.2.2 Data Availability and Presentation

To assess the effects of data availability (DA, a binary variable denoting if the condition was Scatter$^+$, or either Scatter$^-$ or Bar) and function (f_{lin}, f_{x^2}, f_{cos}) on errors, while controlling for the effect of presentation ($Scatter$ denoting if the presentation condition was either Scatter$^+$ or Scatter$^-$, or Bar), we fitted a generalized linear model (GLM): $Y_{MAE} \sim \beta_0 + \beta_f \times (\beta_{Scatter} + \beta_{DA})$. The GLM was specified with an identity link function and allowed for interactions between $Scatter$ and function as well as DA and function.

In concordance with previous findings, f_{cos} had a significant positive effect on error. As predicted by our hypothesis, data availability had a significant, small negative effect on error, but presentation ($Scatter$) was non-significant. No other main effect and none of the interaction terms had a significant effect. For the full GLM results, see Table 3.1. For all extrapolations performed by the participants, see Figure 3.4.

Interim Discussion

The results are consistent with our hypothesis that differences in errors are attributable to differences in data availability. However, a stronger test of our hypothesis lies in the *patterns* of extrapolations that participants make. Do these patterns differ systematically between presentation conditions, or are differences explainable in terms of condition-independent biases?

In the final section, we will explore how differences in availability imposed by our experimental design are reflected in the participants' extrapolations. To analyze these extrapolations, we compared human extrapolations to two Bayesian models, one with low available data and one considering all available data.

These models allow us to capture our assumptions about the underlying representation learned in the two types of experimental conditions. Examining the

Table 3.1: Results of the GLM model assessing if function type (f_{lin}, f_{x^2}, f_{cos}), presentation (*Scatter*), or data availability (*DA*) were predictive of MAE in the test block. The f_{cos} condition had a significant positive effect on MAE. In addition, having all data available (*DA*, corresponding to condition $Scatter^+$) had a significant, small negative effect.

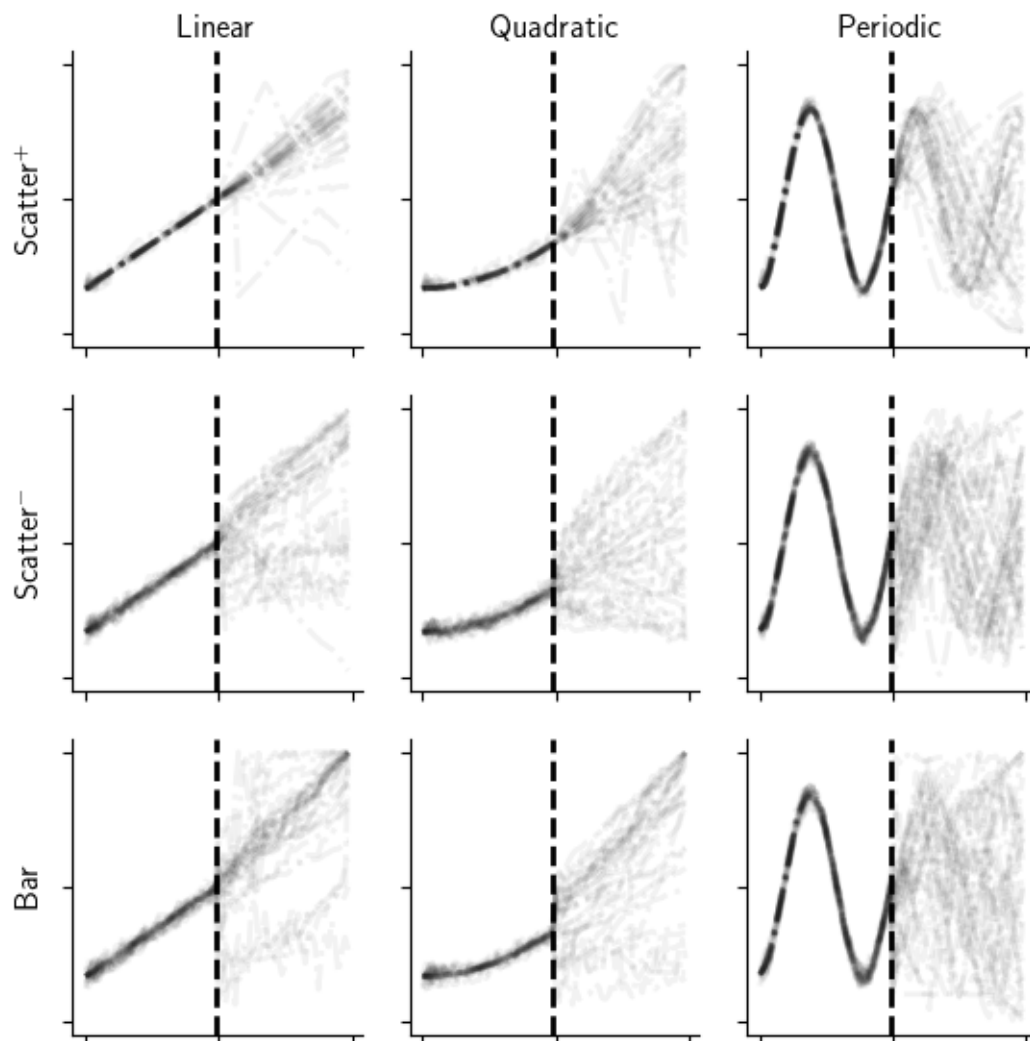|	β	SE	z	$P > |z|$	95%CI
β_0	0.13	0.02	8.78	$p < 0.001$	$0.1, 0.16$
$f_{cos(x)}$	0.15	0.02	7.28	$p < 0.001$	$0.11, 0.2$
f_{x^2}	0.02	0.02	0.75	0.45	$-0.03, 0.06$
Scatter	$\beta < 0.01$	0.02	0.22	0.82	$-0.04, 0.05$
DA	-0.08	0.02	-3.75	$p < 0.001$	$-0.13, -0.04$
$Scatter \times f_{cos(x)}$	-0.03	0.03	-1.02	0.3	$-0.01, 0.03$
$Scatter \times f_{x^2}$	0.02	0.03	0.75	0.46	$-0.04, 0.08$
$DA \times f_{cos(x)}$	-0.01	0.03	-0.24	0.8	$-0.07, 0.05$
$DA \times f_{x^2}$	$\beta < 0.01$	0.03	0.04	0.97	$-0.06, 0.06$

Figure 3.4: All participant extrapolations (gray lines) in the nine experimental conditions. Submissions within the admissible error for the training block are displayed on the left-hand side of the dotted vertical line. Extrapolations for the test block are displayed on the right-hand side. Function conditions are presented by column, presentation conditions by row.

models extrapolations and contrasting them with the experimental data allows us to assess if the participants extrapolations corresponded to our experimental manipulation.

3.3   Modeling Function Extrapolations

The computational problem faced in extrapolation tasks amounts to predicting new values y_{n+1} for test values x_{n+1}, conditional on previously learned $\mathbf{x}_n, \mathbf{y}_n$ and a prior belief $p(f)$ over possible functions. We will adopt a Gaussian process perspective on regression, an approach that has been applied successfully in previous function learning research (Lucas et al., 2015; Schulz et al., 2017).

A Gaussian process specifies a distribution over functions $f(x) \sim GP(\mu, k)$, where $\mu(x) = E[f(x)]$ and k is the covariance kernel $k(x, x') = cov(f(x), f(x'))$. The kernel specifies how much values of x' depend on the other values x and specifies a similarity measure over x. For a brief introduction into Gaussian processes, see Appendix A. We assume that two sets of priors can capture participant extrapolations in our study — a prior over kernel types describing the space of possible functions $f_i \sim \mathcal{F}$, and a prior for individual kernel parameters θ_{f_i}.

3.3.1   Human Function Priors

To specify a plausible prior over functions $\mathcal{F}$, we closely followed Lucas et al. (2015). The prior over functions proposed in Lucas et al. (2015) was motivated by previous empirical results and models trained with this prior could account for a wide range of experimental results in function learning. We used the same prior probabilities for functions $\mathcal{F}$, favoring $f(x) = x$ (Linear$^+$) over negative linear functions (Linear$^-$), and linear functions over other monotonic functions (RBF, the radial basis function kernel). Since our experiment included periodic data that we did not want to exclude a priori, we added a periodic kernel (Periodic)

with good coverage over the range of x, y. We chose a low prior weight for the periodic to account for the difficulty in learning non-monotonic functions (Bott and Heit, 2004; Kalish, 2013). For a full list of parameter priors, θ, see Table 3.2; for samples of the prior functions, see Figure 3.5.

With the priors $\mathcal{F}$ and θ we can express the task faced by our participants in general terms:

$$p(y_{n+1}|x_{n+1}\mathbf{x}_n, \mathbf{y}_n, f) = \int_f p(y_{n+1}|x_{n+1}, y, f) p(f|\mathbf{x}, \mathbf{y}) df \qquad (3.1)$$
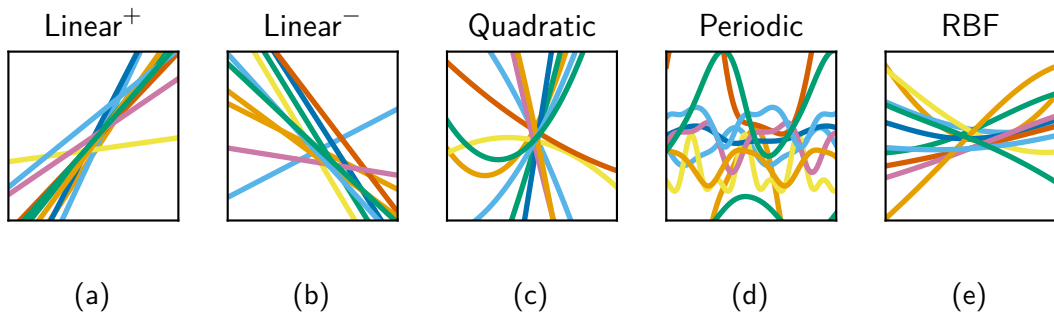

Figure 3.5: Ten samples for each of the functions constituting the prior over functions $\mathcal{F}$. The prior consisted of: (a) a linear kernel biased towards $f(x) = x$, (b) a negative linear, (c) a quadratic, (d) a periodic, and (e) a RBF kernel. All kernels had additional intercept terms. The distribution over functions $\mathcal{F}$ was chosen to closely match Lucas et al. (2015) and was proportional to $8, 1, 0.1, 0.01, 0.01$.

Given appropriate priors and Equation 3.1 a variety of human inductive biases can be accounted for, from strong biases for $f(x) = x$, to results in iterated learning experiments (Lucas et al., 2015).

However, this model assumes that all previously encountered data are equally available and inform posterior inference. In some function learning experiments where participants repeat training until they achieve a very low error rate, these assumptions may be appropriate. In other contexts, including many sequential function learning problems in the natural world, they are less plausible.

Table 3.2: Priors used to specify the two models (curvature of the quadratic kernel, θ_c; period of the periodic, θ_π; lengthscale of the RBF,θ_l). All models had a fixed noise variance of 0.0025, which matched the admissible error in the test set. All models hat an intercept with a prior covering the training range. Both linear kernels had slope parameters, expressing the prior preference for positive and negative slopes respectively.

	σ^2	Intercept	Slope	θ_c	θ_l	θ_π
Linear$^+$	$Exp(\frac{1}{6})$	$\mathcal{N}(0,\frac{1}{2})$	$\mathcal{N}(1,\frac{1}{10})$	–	–	–
Linear$^-$	$Exp(\frac{1}{6})$	$\mathcal{N}(1,\frac{1}{2})$	$\mathcal{N}(\text{-}1,\frac{1}{10})$	–	–	–
Quadratic	$Exp(\frac{1}{6})$	$\mathcal{N}(\frac{1}{2},1)$	$\mathcal{N}(0,1)$	$\mathcal{N}(0,2)$	–	–
Periodic	$Exp(\frac{1}{6})$	$\mathcal{N}(\frac{1}{2},1)$	–	–	$\mathcal{N}(1,\frac{1}{4})$	$\mathcal{N}(\frac{1}{2},\frac{1}{4})$
RBF	$Exp(\frac{1}{6})$	$\mathcal{N}(\frac{1}{2},1)$	–	–	$\mathcal{N}(1,\frac{1}{4})$	–

3.3.2 Modeling Data Availability

We contrasted the predictions of a model trained on the full data set (the 40 training points) with models that had only a sparse set of data available. As a first approximation of the effect of data availability, we assumed that only the last $k \in \{1, 5, 10, 20\}$ points in the training block were available in the Bar and Scatter⁻ conditions.

While the amount of data underlying participants' extrapolations might differ systematically, our analysis is not particularly sensitive to the size of the subset. In general, larger subsets will emphasize the training data, while smaller sets will result in posteriors emphasizing prior inductive biases since the likelihood of the data plays a diminished role. To contrast the effect of the sparsity of the data with the role of the function prior, we evaluated a full and a sparse $(k = 5)$ model that did not favor any particular type of function (flat prior). For the posterior probability over functions for these models, see Figure 3.6.

3.3.3 Posterior Mass for Functions

The models trained on linear data generally assigned high posterior mass to positive linear functions. Only if the model did not exhibit any strong prior preference for linearity or when only a few data points were assumed to be available ($k \leq 5$) did the models assign some posterior mass to other options.

Our analysis revealed that the training data presented was not strongly indicative of a non-linear trend for quadratic conditions. As a result, nearly all models exhibited large posterior mass for positive linear functions. Only when all data was assumed to be available was the evidence sufficient to overwrite the strong prior preference for linearity and assign some posterior mass to quadratic functions.

For periodic conditions, our results exemplify the trade-off between the strength of prior inductive biases and the amount of evidence available. Assuming no strong prior preference for a particular type of function, even a sparse model ($k = 5$) already posits non-negligible posterior mass for a periodic relationship. In contrast, models with priors exhibiting strong inductive bias towards linearity require larger amounts of evidence ($k \geq 20$) to posit a periodic function.

Similar to the role of abstract priors over function types, $\mathcal{F}$, the posterior over function parameters θ_f will depend on the sparsity of the data. In the absence of sufficient evidence, posterior parameters will reflect the prior uncertainty. To exemplify the effect of the amount of evidence and the resulting inferred functions, consider Figures 3.7 and 3.8. While for sparse data, the posterior distributions over function parameters are largely reflective of the prior, the posterior for the full model is heavily peaked on characteristic values for the data.
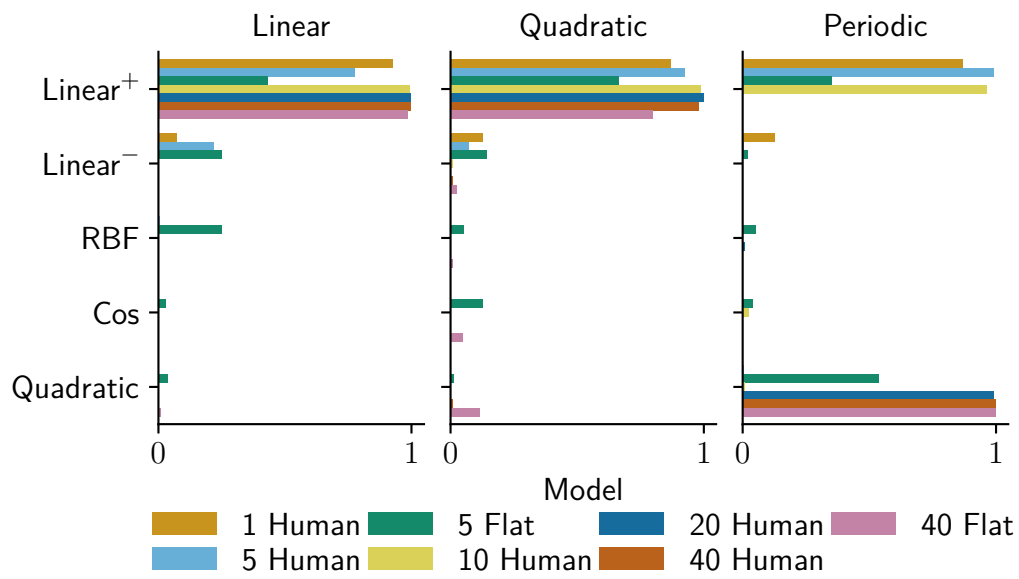
Figure 3.6: The inferred posterior probability for the five functions in each condition.
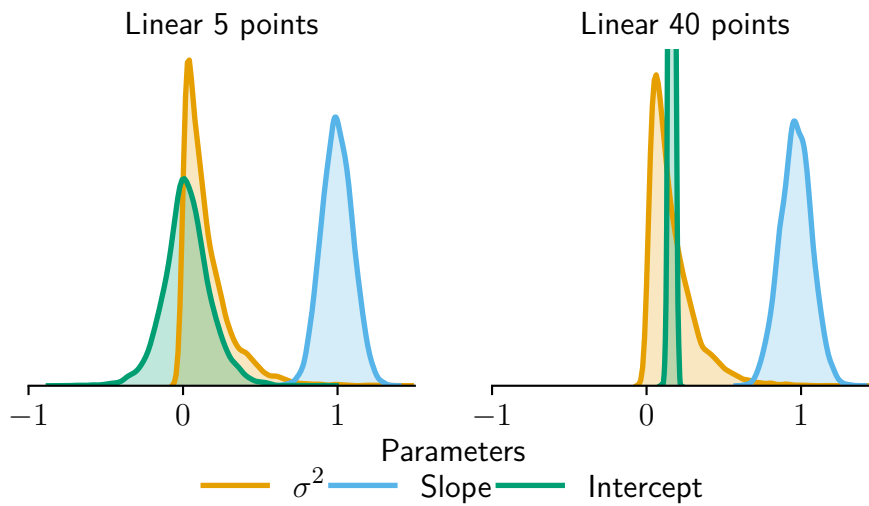

Figure 3.7: Given larger amounts of evidence, the posterior distributions for the intercept will concentrate on the true intercept.
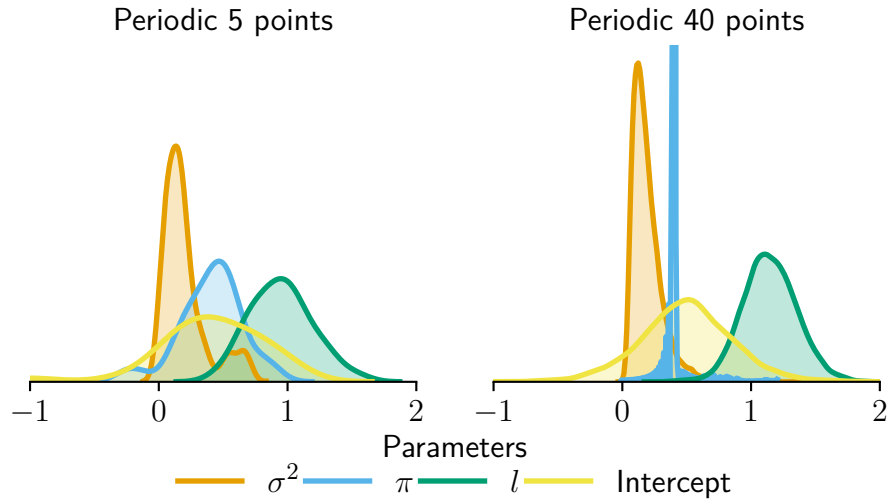
Figure 3.8: Conditional on the full training data the posterior distribution for the periodic function strongly concentrates on the true period.

3.3.4 Posterior Model Extrapolations

While the model's posterior mass over function types reveals which kind of abstract function the model infers, we are also interested in the corresponding extrapolation patterns. Since each model specifies a hierarchical distribution over function types and corresponding function realizations, posterior model predictions amount to posterior densities. For the resulting posterior densities for the full model, as well as a sparse model with a uniform prior over function types and a sparse model with a strong prior preference for linear functions, see Figure 3.9.

In general, sparse and full models captured the strong inductive biases for positive linear functions. Furthermore, our sparse model predicted the strong inductive bias for $f(x) = x$ in Scatter⁻ and Bar conditions, aligning well with the participants' data. The sparse model that did not prefer linear functions a priori exhibited similar inductive biases, but was also more diffuse, potentially due to the contribution of RBF or periodic functions.

For f_{x^2}, both full and sparse models reflected the strong prior preference for positive linearity. As a result, the full model did not capture the extrapolations

of participants in Scatter$^+$. While the model extrapolated from the available data linearly, participants performed steeper, quadratic-like extrapolations (see Figure 3.4). The sparse model was more predictive of the participants' extrapolations in Scatter$^-$ conditions, extrapolating in a steep linear fashion. Similar to models trained on linear data, the model with a uniform prior over $\mathcal{F}$ resulted in similar, albeit more diffuse, posterior density.

For f$_{cos}$, a sparse model with a strong bias towards linearity did not capture the participants' extrapolations well. While the model did favor positive linearity and extrapolated accordingly, many participants exhibited non-monotonic, high-variance extrapolations. In contrast, the full model captured the highly periodic extrapolations in the Scatter$^+$ condition and closely resembled human extrapolations. A sparse model that did not assume a strong prior preference for linear functions resulted in more diffuse posterior density, which was better aligned with participants' extrapolations.

3.3.5   Recovering Experimental Conditions from Likelihoods

To further evaluate how well our models captured characteristic differences between the experimental conditions, we attempted to recover participants' assigned experimental conditions from their extrapolations. We classified participants as either belonging to full or sparse experimental conditions according to the likelihood of their extrapolations given our models. We labeled a participant as belonging to a sparse condition if the model with the highest likelihood had $k \leq 10$. We then contrasted this classification with the true experimental condition. For confusion matrices for this classification procedure, see Figure 3.10. For the three participant extrapolations that had the highest likelihood for each model, see Figure 3.11.

This classification validated our general modeling results. While our method recovered participants' experimental condition reasonably well for linear data (65
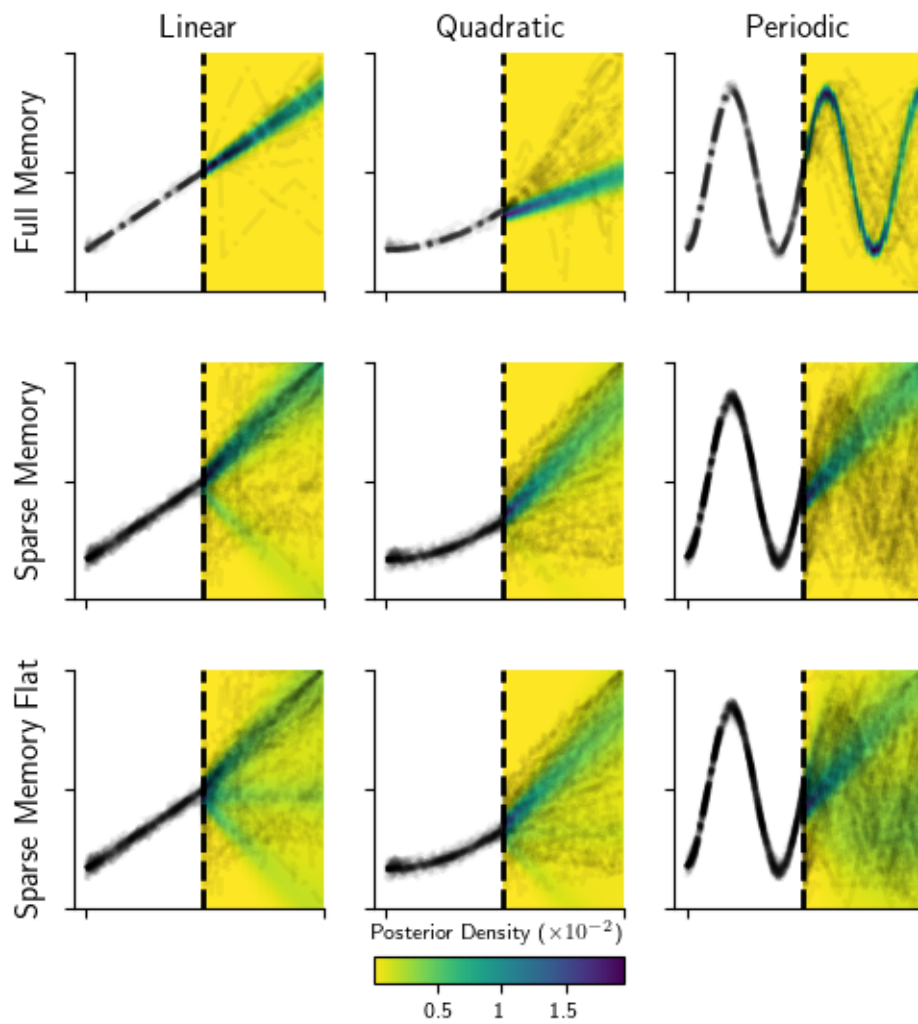
Figure 3.9: Posterior density for the models, with darker colours corresponding to higher posterior density. In concordance with our hypothesis, participants in the Scatter$^+$ condition kept a full representation of the training data available, corresponding to the full model (top row). The mid row present the sparse model posterior density. The bottom row presents the sparse model with a uniform prior distribution $\mathcal{F}$.

out of 108 classified correctly, $p_{correct} < 0.5^2$), for quadratic (46 out of 106 classi-fied correctly, $p_{correct} = .93$) and periodic data (41 out of 108 classified correctly, $p_{correct} > .99$), our method exhibited asymmetric confusions. For quadratic data our method resulted in high misclassification of Scatter$^+$ and chance level per-formance for Scatter$^-$ and Bar. Similarly, for periodic data, our model highly favors Scatter$^+$, consistent with a large proportion of participants performing cyclic extrapolations.
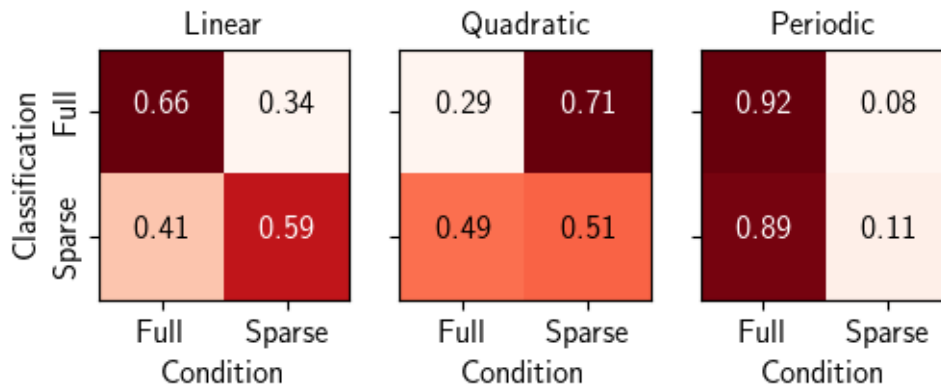
Figure 3.10: We contrasted our classification with the true experimental conditions. Our classification captured the effect of data availability for f_{lin}. However, it exhibited systematic mis-classification for f_{x^2} and f_{cos}. In f_{x^2}, we were at chance level classifying participants as belonging to Scatter$^-$ or Bar, and failed to recognize Scatter$^+$. In f_{cos}, our procedure mis-classified participants in the sparse conditions, but captured extrapolations in the Scatter$^+$ condition.

3.4   Discussion

We hypothesized that differences between function learning and function estima-tion experiments could be attributed to participants having direct access to all data points in the latter. More precisely, we sought to test the idea that the same inductive biases are at work in both settings, but that the reduced access

---

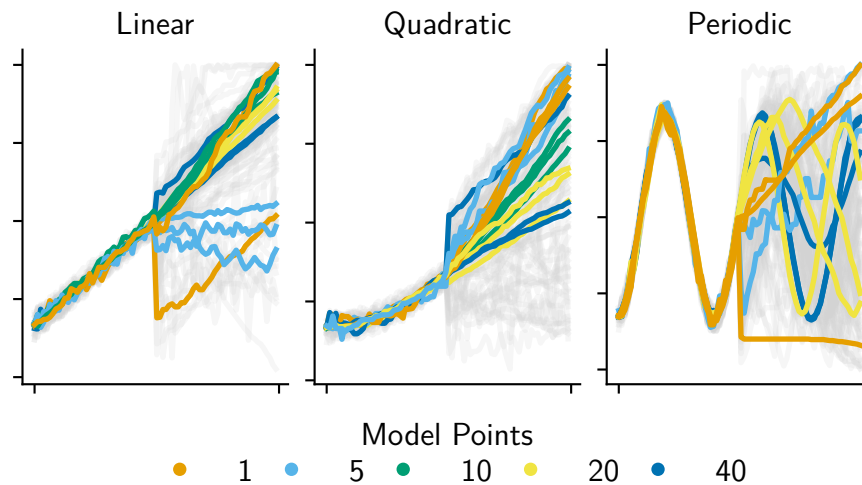[2]All test here are one-sided exact Binomial tests against chance (0.5).

Figure 3.11: Three extrapolations with the highest likelihood in each function condition. We categorized participants' extrapolations by contrasting the likelihood of our full and sparse models.

to data in function learning designs causes these biases to play a stronger role in shaping participants' extrapolations. As we anticipated, participant behavior in both Scatter⁻ and Bar was almost indistinguishable, demonstrating the same qualitative patterns. Furthermore, both Scatter⁻ and Bar were clearly different from extrapolations in the Scatter⁺ conditions.

However, we found mixed support for the more detailed hypotheses reflected in our Bayesian model. Behavior in both linear conditions was as predicted, with participants in Scatter⁻ and Bar conditions tending to extrapolate according to the $f(x) = x$ function, that past research has shown to be favored a priori, rather than the true function. In the quadratic conditions, our model did not capture participants' behavior, especially in the Scatter⁺ condition where participants were *more* likely than the model to infer a non-linear relationship. There are many possible explanations, one of which is that our simplistic assumptions about participant memory failed to capture the loss of precision in the locations of points.

Perhaps the most interesting deviation between the model's predictions and

participants' judgments is in the f_{cos} conditions. Contra the model's predictions — as well as our expectations — individual participants were quick to infer non-monotonic functions, even in the Scatter⁻ and Bar conditions. This result also admits several explanations, but one intriguing possibility is that people are better at tracking high-level, qualitative properties of functional relationships than the details of those relationships' parametrization.

In the next chapter, we will continue exploring what kind of abstract properties people perceive and track when they learn continuous relationships.

Chapter 4

A Distributional Space of Functions

The previous chapter has highlighted that people track high-level, qualitative properties of the functions they learn. This chapter explores this result by examining what kind of abstract properties people track when learning about functions. More specifically, I will analyze if people can track the variability of encountered functions.

While previous research has shed light on function learning and the representations and inductive biases that make it possible, some fundamental questions remain. For example, models that take a distributional approach to function learning have successfully explained human behavior. However, there is little direct evidence that people track distributional information — uncertainty or variability — when faced with function learning problems. This question has been unanswerable in previous work that relied on aggregated judgments or assumed that individual inductive biases are broadly similar (Kalish et al., 2007). Even the few studies that focus on inference patterns (Kalish, 2013; Wilson et al., 2015; Schulz et al., 2017) still neglect questions about the tacit beliefs behind participants' judgments. Only recently, experiments have started to explore the role of uncertainty in function learning. In Schulz et al. (2015), participants judged functions to be more predictable when they were smooth or exhibited low variance,

following the preferences of a probabilistic model. Similarly, Stojic et al. (2018) showed that participants' predictive accuracy in a function learning task correlated with their confidence ratings, again resembling the uncertainty estimated by a probabilistic model.

Here, we expand on this work and attempt to characterize how people represent uncertainty when they learn functions.

4.1   Markov chain Monte Carlo with people

To uncover the psychological space that participants learn when learning functions, we apply Markov chain Monte Carlo with people (hereafter MCMCP, Sanborn et al., 2010).

MCMCP is motivated by Markov chain Monte Carlo (MCMC), a method in statistics that can generate samples from an arbitrary target distribution. In MCMC, samples from the target distribution are produced by iteratively comparing a new sample to the current sample (or state of the sampler), probabilistically selecting the more likely sample under the target distribution. The selected (or accepted) sample then becomes the sampler's new state. If the selection procedure satisfies a set of mathematical criteria, it can be proven that, after enough time, the accepted samples will correspond to samples of the target distribution.

Sanborn et al. (2010) showed that Markov chain Monte Carlo can be used as an experimental method to elicit posterior distributions from people using a simple forced-choice task. As in MCMC, in MCMCP, participants have to select between two samples, one corresponding to a new proposed instance of a category of interest, the other being the sampler's current state. Participants are asked to select the more representative sample and, based on their choice, the state is updated. This choice can be shown to correspond to a statistically valid acceptance procedure Sanborn et al. (2010); thus, given enough of these forced

choices, samples obtained from MCMCP will be samples from the participants' implicit category representation.

Thus, MCMCP offers a method to explore the psychological representational space and has been successfully applied to elicit the representations of complex stimuli, such as peoples' representation of facial emotional expressions (Martin et al., 2012). Previously, MCMCP has been used in a function learning setting[1] to examine if participants prefer compositional over non-compositional functions (Schulz et al., 2017). Since Schulz et al. (2017) were interested in preferences for types of functions (compositional vs. non-compositional), the samples presented consisted of discrete varieties of functions and did not explore the distribution of function parameters.

In contrast, in this work, we directly explore the distributional space of the parameters governing linear functions. This approach lets us uncover how learned functions are represented without constraining participants' choices to pre-specified sets of materials. For an overview of how the forced-choice task results in the posterior distribution, see Figure 4.1.

Adopting MCMCP also allows us to explore new questions — do participants represent variability in the training relationships? Do they form a single, deterministic functional relationship, or do they form posterior distributions over parameters, reflective of the variability in training? In turn, this question about representation, can inform more general future questions: do typical extrapolation patterns amount to maximum a posteriori estimates for a learned function? Or do they correspond to samples from a range of probable parametrizations?

---

[1]Function learning has been more extensively studied in a closely related paradigm, *iterated learning*. Iterated learning experiments can elicit participants' shared expectations and have revealed strong inductive biases for positive linear functions (Kalish et al., 2007).

(a) Traceplots for intercept & slope.   (b) 2-D trajectories & marginals.
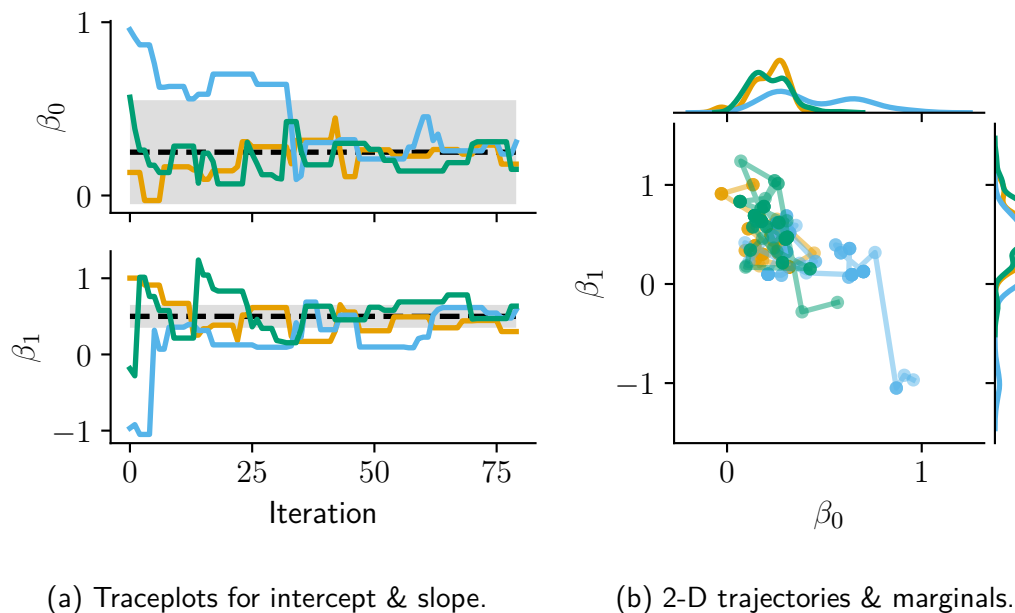
Figure 4.1: The 240 choices submitted by the participants corresponded to three Markov chains. By accepting or rejecting proposed parametrizations for the functions, participants traverse this representational space and eventually converge to a region reflecting the posterior over parameters. The process starts at a random position and slowly moves towards the typical set of the posterior distribution (a). Samples from the typical set correspond to posterior draws from the target distribution (b). For this participant, the chains converge after 35 iterations for β_0 and after 15 iterations for β_1. The corresponding distribution after this burn-in period closely matches the true relationship learned in the training phase, in terms of both its mean and variance (a, dashed line and gray range).

4.2 Experiment

In this experiment, we examine how participants represent linear functions when presented with sets of training examples. We hypothesize that participants learn both the parameters generating the function and the variability of the relationship, i.e., they will learn how much slopes and intercepts vary, while also learning the specific modes of slopes and intercepts. Therefore, we expect participants to form posterior distributions over the training parameters, with the variance of that posterior reflective of the training.

We distinguish between training functions with positive and negative slopes since previous research has highlighted strong inductive biases for these relationships. Similarly, it has been shown that people are biased to extrapolate matched linear functions, and their extrapolations are influenced by data-boundaries (De-Losh et al., 1997). In areas of the extrapolation range that are close to zero, participants adjust slopes towards the boundary (Brown and Lacroix, 2017; Kwantes and Neal, 2006). Thus, we contrast steep and shallow linear functions to test how different offsets and degrees of steepness are represented. We expect that highly salient functional relationships, like matched positive functions, will be easier to learn and result in more peaked posterior distributions if the training exhibits low variability. For high variability training, and especially if the function is not favored as strongly (for instance, a function with a shallow negative slope), we expect broader, less peaked posteriors. Finally, we hypothesize that, especially in high variability conditions, some participants will not exhibit uni-modal posterior distributions and might consider several potential generating functions broadly consistent with the learned function.

Contrasting these functions resulted in a $2 \times 2 \times 2$ between-subjects design (direction of the function: positive or negative, steepness: shallow or steep, the variability of the training data: low or high).

4.2.1   Participants

We recruited 454 participants ($M_{\text{age}} = 33$, $SD_{\text{age}} = 8.63$; 91 female, 176 male, 1 other, 186 refused information on gender) on Amazon Mechanical Turk. Participants had to have completed more than 50 approved tasks with an approval rate of 95% or higher. They received \$1.33 for participation and took an average of 17 minutes ($M = 17.25$, $SD = 8.59$) to complete the experiment. Participants were randomly assigned to one of the eight conditions.

4.2.2   Materials

The parameters generating the functions in the experimental conditions differed in the direction of the slopes, as well as in their steepness. In addition, parameters in the training set exhibited either low or high variance for intercepts and slopes. For the full set of experimental conditions, see Table 4.1.

Table 4.1: The materials were draws from sets of linear functions with intercepts β_0 and slopes β_1 drawn from high- and low-variance normal distributions $(SD_{\beta_0}, SD_{\beta_1})$.

	β_0	SD_{β_0}	β_1	SD_{β_1}
$C_{0.5,\text{low}}$	0.25	0.05	0.5	0.025
$C_{1.0,\text{low}}$	0	0.05	1	0.025
$C_{-0.5,\text{low}}$	0.75	0.05	−0.5	0.025
$C_{-1.0,\text{low}}$	1	0.05	−1	0.025
$C_{0.5,\text{high}}$	0.25	0.3	0.5	0.15
$C_{1.0,\text{high}}$	0	0.3	1	0.15
$C_{-0.5,\text{high}}$	0.75	0.3	−0.5	0.15
$C_{-1.0,\text{high}}$	1	0.3	−1	0.15

To create the 25 training sets, corresponding to independent and identically

distributed (i.i.d.) realizations of $\beta_0, \beta_1 \sim \mathcal{N}(\mu, \sigma)$, with μ and σ matching the experimental condition, we systematically sampled 10,000 pairs and selected the most normal and uncorrelated sets[2]. Then, we generated the corresponding linear function for a range of 15 points for x in 0–1 for all sets. One of those 15 values was picked at random and constituted the interpolation target. For the resulting materials, see Figure 4.2.
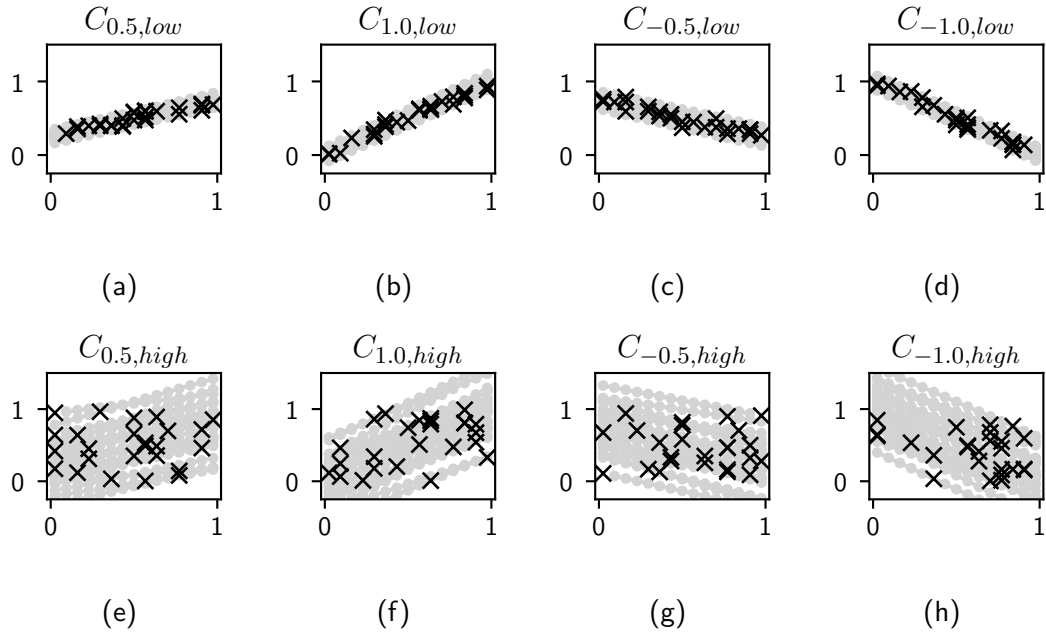

Figure 4.2: The eight conditions differed in the functions presented during training. Functions were either positive (a, b, e, f) or negative linear relationships (c, d, g, h). Participants had to extrapolate one randomly selected data point (marked with an x) in 25 successive interpolation tasks.

4.2.3 MCMCP

Proposals were generated by two symmetric Gaussian distributions, $\mathcal{N}(\mu, \sigma)$, to allow for both local and distant proposals $\sigma_{\beta_0} \in [0.14, 0.98]$, $\sigma_{\beta_1} \in [0.21, 1.47]$, respectively. At each iteration, the local proposal was selected with a probability

---

[2]All Shapiro-Wilk tests yielded $p > 0.99$, and all correlation coefficients were in -0.01–0.01

of 0.8 and a distant proposal with a probability of 0.2. Proposals were further restricted to be in bounds $\beta_0 \in [-0.5, 1.5]$, $\beta_1 \in [-1.5, 1.5]$, and if less than four points of the function realization were visible on the screen, the proposal was automatically rejected, and a new proposal was resampled.

Participants traversed three different, interleaved chains since multiple chains allow a wider application of convergence diagnostics and reduce the impact of the particular starting state. The starting values for these chains were obtained by k-means clustering of pilot data ($n = 8$, one participant per condition). This resulted in the following starting values $\beta_0 = \{0.12, 0.1, 0.58\}, \beta_1 = \{0.92, -0.94, -0.28\}$ for chains one to three.

4.2.4   Procedure

Participants were instructed that they would learn the relationship between two proteins, Zenopin and Mepradin. Participants were told that the concentration of Zenopin was related to Mepradin, but that the extent varied between humans. Participants were also instructed that they would be presented with examples of the relationship, as observed in different people and that they would be asked to interpolate the relationship. They were then instructed that after the training phase, they would be presented with pairs of proposed relationships, all observed for a new person, and would have to choose which of the two was more likely to resemble the learned relationship. After reading the set of instructions, the participants were tested on their comprehension. If participants did not respond correctly in the questionnaire, they had to restart the instructions.

4.2.4.1   Training Phase

In the training phase, participants were presented with 25 interpolation tasks presented as scatter plots. In each task, they were instructed that the scatter plot depicted the relationship between the two protein concentrations for a new

person. They then had to guess the concentration of the protein by selecting the height of the corresponding value on the plot (on the y-axis). Participants were shown the correct value as feedback for one second, and if their choice deviated by more than ± 0.05 from the actual value, they had to readjust their selection.

4.2.4.2 Test Phase

The test phase consisted of 240 forced-choice tasks, corresponding to 80 interleaved iterations of the three Markov chains. On each trial, participants were presented with two adjacent scatter plots, one corresponding to the current state of the chain and the other reflecting the proposed new state (in randomized order). Participants had to select the plot they believed most likely to depict the relationship in the training phase. After the test phase, participants completed a short survey, were debriefed, and compensated. See Figure 3.2 for a depiction of both training and test phase. For screenshots of the experimental stimuli and instructions, see Appendix C.

4.3 Results

We excluded participants from the analysis if their chains did not converge to the stationary distribution. Many criteria for convergence checks have been suggested in the literature; here, we applied one of the most commonly used evaluations, $\hat{R}$ (Gelman et al., 2013; Vehtari et al., 2019). $\hat{R}$ estimates the ratio between within-chain variances and between-chain variance and thus provides a measure of how (self-)similar chains are.

In general statistical practice, $\hat{R}$ should not exceed a value of 1.1. However, such a strict application of this diagnostic is not realistic in most MCMCP experiments, since human judgments might exhibit more correlated choices and the number of iterations in experiments is usually considerably lower than in statis-
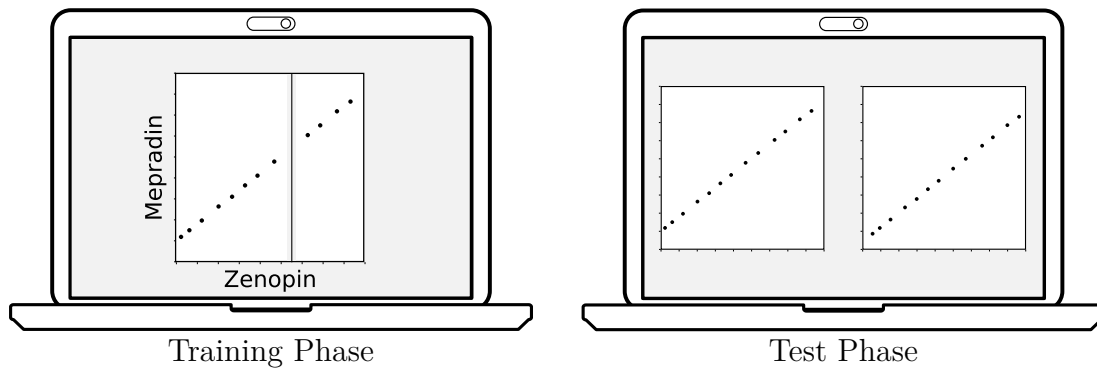
Figure 4.3: Participants had to complete a training phase and a test phase. In the training phase, they were asked to interpolate the concentration of a fictitious protein for 25 different people (with feedback). In the test phase, they were presented with 240 forced-choice tasks, for which they had to choose the scatter plot that most resembled the relationship in the training phase. The choices were presented in random order and corresponded to a Markov chain, in which the participant implemented the acceptance function.

tical applications. Therefore, we incrementally calculated $\hat{R}$ values for chains for each participant and selected the lowest overall $\hat{R}$, with the additional constraints that the first 20 samples of the chain were always discarded and the resulting chains had to be at least 20 iterations long. We then used the maximum of the intercept and slope $\hat{R}$ values to apply exclusion criteria and determine burn-in.

Similar to Ramlee et al. (2017), we excluded participants who exhibited $\hat{R} \geq 2$. Furthermore, we excluded participants who required more than one correction in the interpolation task. Given that the interpolation function was deterministic, most participants did not require many corrections ($Mdn = 0$, $IQR_{1-3} = [0,1]$, $max = 44$).

In total, these methods excluded 262 participants (convergence exclusions: 224, interpolation exclusions: 72). This high number of exclusions was expected given the correlated, bivariate parameter space, and previous results (Sanborn et al., 2010). For group sizes after exclusions, see Table 4.2.

4.3.1   Determining Burn-in

To determine how many trials were required on average for the Markov chains to converge, we used the iteration for which $\hat{R}$ was optimal for each participant. On average, chains required 33 iterations to reach optimal burn-in and the resulting optimal $\hat{R}$ values were well below 2, $M_{\hat{R}} = 1.4$, $SD = 0.2$. Conditions did not differ considerably in terms of the optimal iterations or the resulting $\hat{R}$ values. For the full list of burn-in values, see Table 4.2. For all subsequent analyses, we discarded all points of the chain before the individual burn-in index.

Table 4.2: Condition sizes before (N_{total}) and after exclusion (N). We calculated the optimal burn-in iteration for each participant $M_{\text{burn-in}}$, $SD_{\text{burn-in}}$ and the resulting acceptance probabilities (M_{acc}, SD_{acc}).

	N_{total}	N	$M_{\text{burn-in}}$	$SD_{\text{burn-in}}$	M_{acc}	SD_{acc}
$C_{0.5,\text{low}}$	48	25	34.88	14.49	35	17
$C_{1.0,\text{low}}$	63	21	31.37	12.01	42	10
$C_{-0.5,\text{low}}$	52	19	34.37	13.73	37	13
$C_{-1.0,\text{low}}$	64	22	29.59	11.59	38	15
$C_{0.5,\text{high}}$	59	35	32.29	13.22	38	14
$C_{1.0,\text{high}}$	57	26	32.08	12.24	45	9
$C_{-0.5,\text{high}}$	56	29	35.66	12.75	42	13
$C_{-1.0,\text{high}}$	55	15	29.40	10.67	36	12

4.3.2   Acceptance Probabilities

Acceptance rates for MCMC samples should range between 20–40% (Roberts et al., 1997). Mean acceptance probability was in that range, $M = 39\%, SD = 13$, indicating that the proposals were wide enough to traverse the parameter space.

Between conditions, the mean acceptance probabilities for participants varied, ranging from 35–45%; for all acceptance probabilities, see Table 4.2. For each condition, acceptance probabilities for each chain did not vary substantially and were similar to the general acceptance rates (not shown).

4.3.3   Posterior Distributions

Slopes differed significantly between conditions, with participants trained on negative slopes preferring negative slopes, $M_{\beta_1} = -0.16$, $SD_{\beta_1} = 0.53$, and participants trained on positive slopes preferring positive slopes, $M_{\beta_1} = 0.19$, $SD_{\beta_1} = 0.45$, $t(165.33) = -4.74$, $p < .0001$[3].

For conditions with negative slopes in the training sets, steep and shallow conditions exhibited significantly different posterior slopes, with lower slopes for steep compared to shallow conditions ($M_{-0.5} = -0.05$, $SD_{-0.5} = 0.45$; $M_{-1.0} = -0.29$, $SD_{-1.0} = 0.59$; $t(65.58) = 2.08$, $p < .05$). For conditions with positive slopes in the training sets, there was also a significant difference in posterior slopes. However, this difference was not in the predicted direction, as slopes in the shallow condition were on average larger than in the steep condition, $M_{0.5} = 0.29$, $SD_{0.5} = 0.4$, $M_{1.0} = 0.05$, $SD_{1.0} = 0.47$, $t(89.75) = -2.89, p < .01$. Posterior intercepts in conditions with negative training slopes ($[-0.5, -1.0]$) did not differ significantly between steep and shallow conditions, $M_{-0.5} = 0.52$, $SD_{-0.5} = 0.21$, $M_{-1.0} = 0.6$, $SD_{-1.0} = 0.3$, $t(62.84) = 1.38$, $p > .1$, nor for conditions with positive training slopes, $M_{0.5} = 0.35$, $SD_{0.5} = 0.2$, $M_{1.0} = 0.5$, $SD_{1.0} = 0.25$, $t(88.71) = 3.31$, $p < .05$.

Equally, per-participant *SD*s for slopes did not differ significantly between high and low variability conditions, $M_{low,\beta_1} = 0.49$, $SD_{low,\beta_1} = 0.26$, $M_{high,\beta_1} = 0.55$, $SD_{low,\beta_1} = 0.25$, $t(180.07) = -1.39$, $p > .1$. However, for intercepts, per-participant *SD*s did differ significantly between high and low variability con-

---

[3]All tests are unequal variance, two-sided *t*-tests.
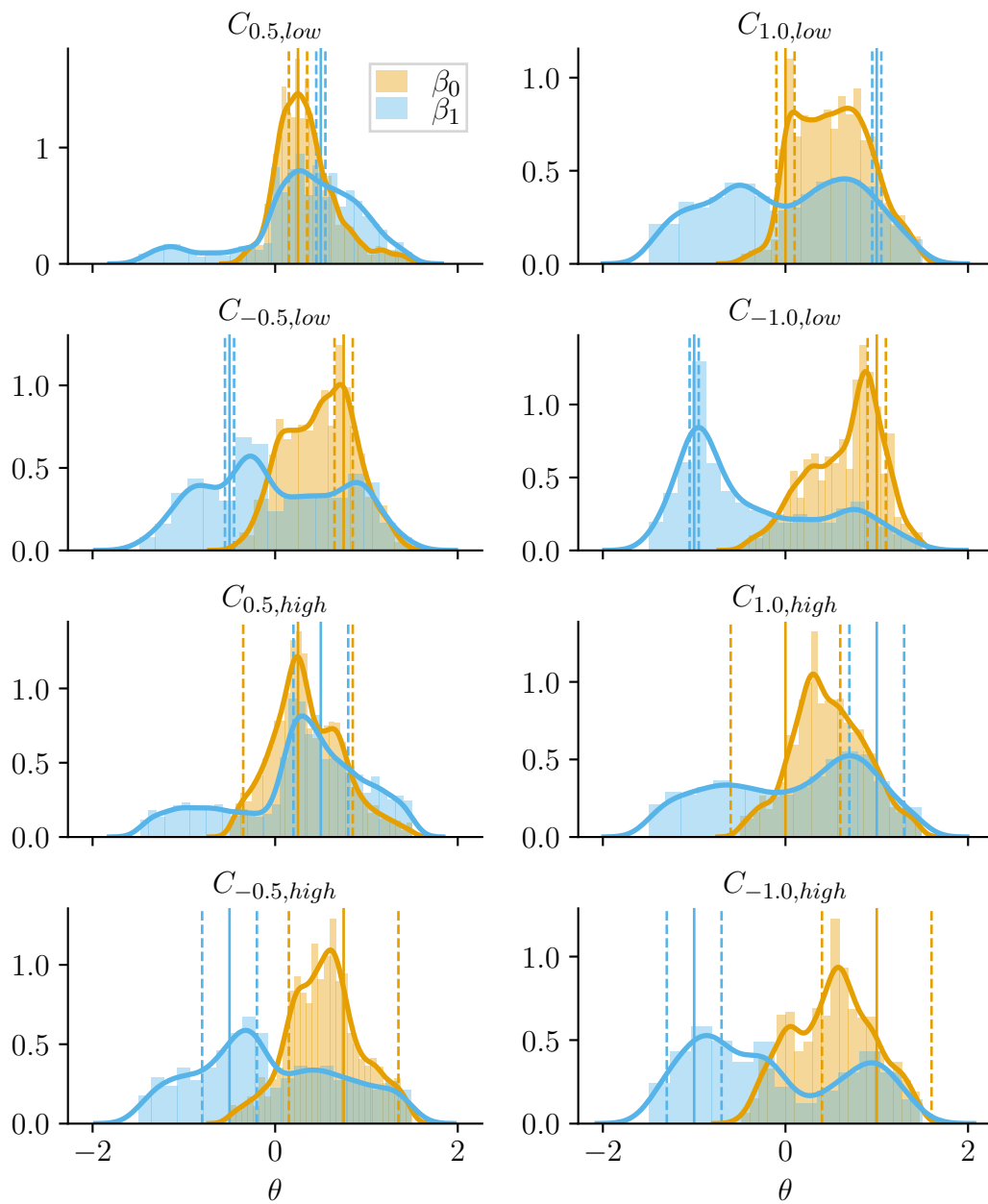
Figure 4.4: The posterior densities and the true training means (solid lines) and standard deviations (dashed lines). The posterior densities exhibited multiple modes, some centered in close proximity of the true parameters.
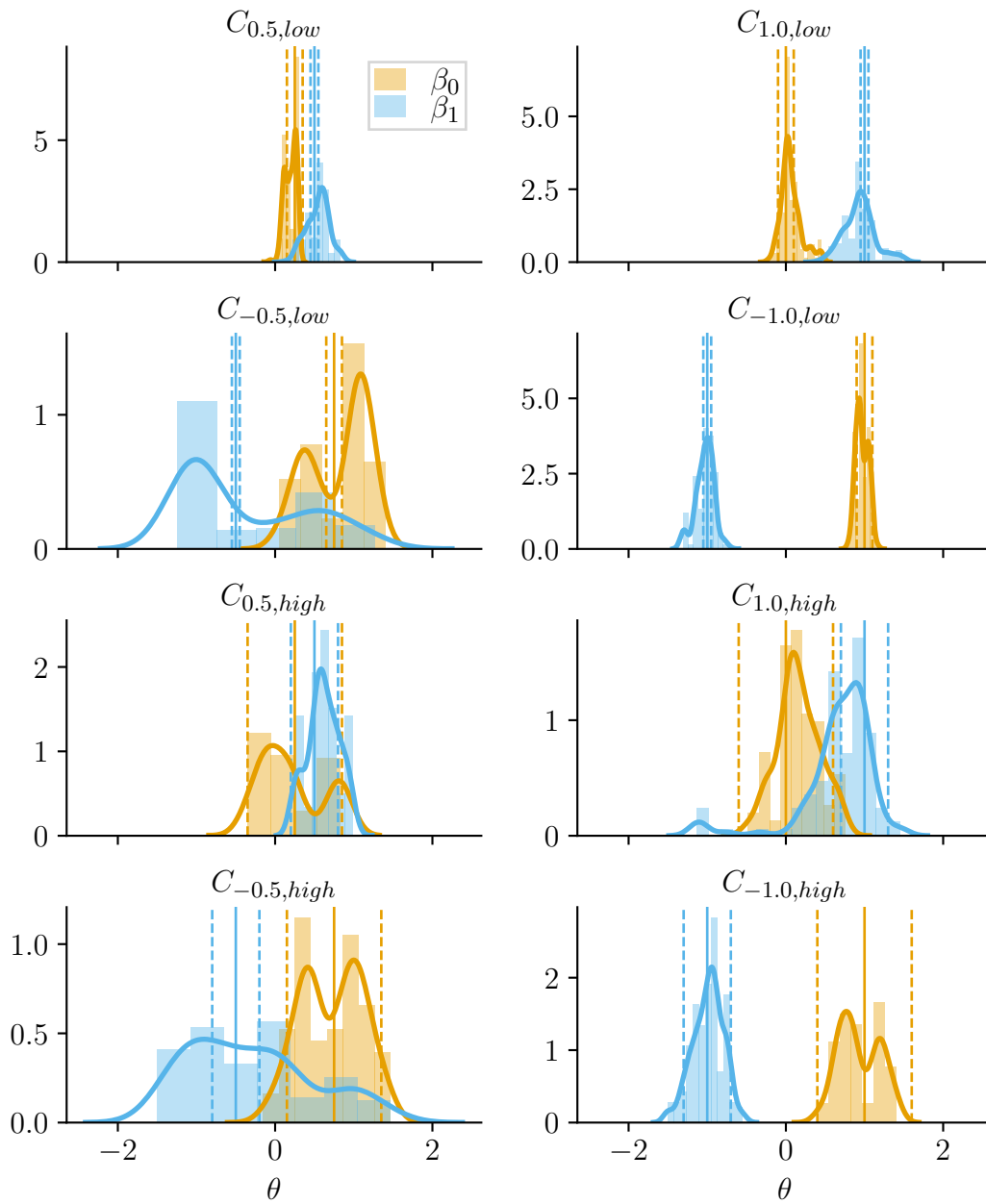
Figure 4.5: Posterior densities for one participant in each condition. Lines represent the true values and standard deviations (dashed lines) in the experimental conditions.

ditions, with high variance conditions resulting in higher SD, $M_{low,\beta_0} = 0.26$, $SD_{low,\beta_0} = 0.11$, $M_{high,\beta_1} = 0.31$, $SD_{low,\beta_1} = 0.11$, $t(182.48) = -2.46$, $p < .05$.

Visual inspection revealed that posterior distributions in all conditions were multimodal and heavily skewed, which complicated the analysis. In general, the posterior densities suggested that the modes of the posterior distributions were often close to the learned parameters (see Figure 4.4; for a selection of posterior distributions for one participant in each condition, see Figure 4.5).

Table 4.3: Posterior means and variances per condition, for function intercepts (β_0) and slopes (β_1).

	M_{β_0}	SD_{β_0}	M_{β_1}	SD_{β_1}
$C_{0.5,\text{low}}$	0.34	0.32	0.32	0.62
$C_{1.0,\text{low}}$	0.52	0.40	0.00	0.78
$C_{-0.5,\text{low}}$	0.49	0.37	−0.02	0.73
$C_{-1.0,\text{low}}$	0.65	0.40	−0.40	0.77
$C_{0.5,\text{high}}$	0.35	0.39	0.27	0.71
$C_{1.0,\text{high}}$	0.47	0.40	0.07	0.81
$C_{-0.5,\text{high}}$	0.54	0.40	−0.07	0.77
$C_{-1.0,\text{high}}$	0.52	0.44	−0.20	0.83

Since the means and standard deviations of multimodal, heavily skewed distributions are not good representations of the underlying data, and because we were interested in characteristic modes of the distributions, we used mixture models to identify dominant modes of the posterior distributions.

4.3.3.1 Estimating Posterior Density Clusters

We estimated Gaussian mixture models that best described the distributions for each experimental condition. We incrementally increased the number of com-

ponents and selected the model with the lowest Bayesian Information Criterion $(BIC)^4$.
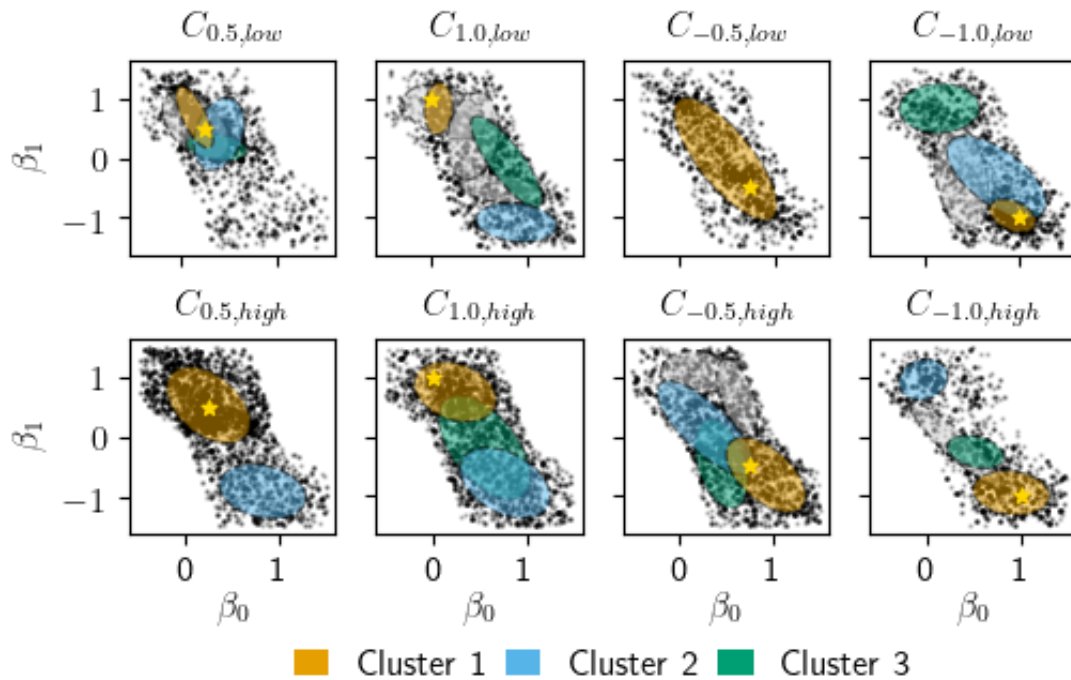
Figure 4.6: Clusters obtained by fitting a Gaussian mixture model (oval shapes). The top three clusters (colored shapes) accounted for a large proportion of the data and in general matched the distribution learned in the training phase well (mean parameters of the true distribution in yellow).

The clustering produced a moderate number of clusters, reflecting the multi-modal nature of the data. In general, each condition was estimated to correspond to a mixture of 1–8 clusters $(M = 4.5, SD = 2.56)$, and the largest clusters closely matched the different training conditions. For Kullback-Leibler (KL) divergences between training distribution and the inferred clusters, see Table 4.4; for the number of clusters, weights, means and covariances for the largest clusters, see Table 4.5; for plots of the clusters, see Figure 4.6.

---

[4]Estimating the mixtures with a Bayesian Dirichlet process mixture model yielded very similar results.

Table 4.4: KL-divergence between the training distribution and the three largest clusters. In general, one of the largest clusters corresponded well to the training distribution.

	$KL_{c=1}$	$KL_{c=2}$	$KL_{c=3}$
$C_{0.5,\text{low}}$	2.18	1.1	1.95
$C_{1.0,\text{low}}$	1.74	42.1	5.85
$C_{1.0,\text{low}}$	1.35	–	–
$C_{-1.0,\text{low}}$	0.31	1.76	24.16
$C_{0.5,\text{high}}$	0.83	10.37	–
$C_{1.0,\text{high}}$	1.06	9.76	2.95
$C_{-0.5,\text{high}}$	1.49	2.66	4.25
$C_{-1.0,\text{high}}$	1.02	59.27	11.09

Table 4.5: The total number of clusters (N_c) assigned was generally low and the weight of the largest clusters was relatively large (16–100%).

	N_C	$w_{c=1}$	$\mu_{\beta_{0,c=1}}$	$SD_{\beta_{0,c=1}}$	$\mu_{\beta_{1,c=1}}$	$SD_{\beta_{1,c=1}}$
$C_{0.5,\text{low}}$	8	0.2	0.15	0.02	0.69	0.14
$C_{1.0,\text{low}}$	8	0.17	0.07	0.01	0.84	0.1
$C_{-0.5,\text{low}}$	1	1.0	0.49	0.14	-0.01	0.53
$C_{-1.0,\text{low}}$	4	0.42	0.93	0.03	-0.98	0.04
$C_{0.5,\text{high}}$	2	0.81	0.24	0.10	0.54	0.21
$C_{1.0,\text{high}}$	3	0.46	0.24	0.1	0.75	0.13
$C_{-0.5,\text{high}}$	5	0.31	0.93	0.09	-0.65	0.2
$C_{-1.0,\text{high}}$	5	0.39	0.9	0.08	-0.95	0.07

4.3.3.2 Per-Participant Clusters

To evaluate if the source of the multimodality in our data was due to averaging over diverse cohorts of participants, or if individual participants produced multimodal posteriors, we performed the same clustering procedure on a per-participant basis. Participant posterior distributions were characterized by 1–12 clusters ($M = 3.11, SD = 1.96, IQR_{1-3} = [1, 4]$), suggesting that individual participants exhibited multimodal distributions. Furthermore, some participants with optimal $\hat{R}$ (≤ 1.1) also exhibited multiple clusters, indicating that the multimodality was not simply due to poor convergence ($M = 1.89, SD = 1.36, N_{\hat{R} \leq 1.1} = 9$).

The number of clusters did not differ significantly between low-variance and high-variance conditions, $M_{low} = 2.98, SD_{low} = 1.94, M_{high} = 3.1, SD_{high} = 1.57,$ $t(164.24) = -0.49, p > .3$. Neither did the slope variance differ significantly in the largest cluster, $M_{low} = 0.1, SD_{low} = 0.13, M_{high} = 0.1, SD_{high} = 0.11, t(172.43) = $ 0.11, $p > .5$. However, for intercepts the variances of the largest clusters were significantly different, with smaller cluster variances for low-variance conditions, $M_{low} = 0.04, SD_{low} = 0.03, M_{high} = 0.05, SD_{high} = 0.04, t(189.85) = -2.09, p < $ .05.

4.4   Discussion

We found some evidence that participants represent the functions learned in training as distributions over parameters. Furthermore, the modes of these distributions were, in many cases, aligned with the true parameters. Also, for intercepts, but not for slopes, these distributions were affected by differences in training variability. Finally, our results suggest that the learned distributional spaces over function parameters can exhibit multiple modes.

The multimodality in the posterior distributions allows for two interpretations. First, it is possible that participants truly evaluated distinct candidate

representations, and thus multimodal posterior distributions characterized their hypothesis space. It is plausible that a priori strongly favored relationships, in addition to the implied parameters in training, constitute the psychological space when learning sets of varying functions. Second, the multimodality might also arise from our experimental method. One issue could be the number of iterations. Theoretically, MCMCP is well suited to discover complex, multimodal distributions, but practically many more samples could be necessary to achieve convergence to the posterior distribution. Since vast numbers of iterations might not be feasible from an experimental perspective, one practical test of our results could be starting the chains of later participants at the endpoints of previous participants (Martin et al., 2012).

Future research should clarify the source of multimodality, for instance, by comparing our results with results obtained by multidimensional scaling (*MDS*). MDS provides an alternative experimental method to obtain participants' representations. Unlike MCMCP, MDS uses similarity ratings between stimuli to construct the internal representations. Thus, MDS results could provide additional evidence for multimodality, based on an alternative experimental paradigm.

If such a comparison corroborates our results, these insights into the structure of psychological spaces could, in turn, provide invaluable guidance for future generalization research. MDS would also allow us to address two shortcomings of the current study: its exclusive focus on linear functions, and the potential influence of perceptual similarity of functions on participants' forced choices. First, similarity judgments obtained via MDS could be used to determine if participants are well-described by linear models, or if non-linear representations underlie their judgments. These results would allow us to determine if the multimodal representations observed in our experiment were the result of a lack of satisfactory choices or a genuine characteristic of learning. Second, MDS would allow us to chart sets of perceptually similar samples. It is plausible that intercepts and slopes can af-

fect notions of similarity of linear functions differently. For example, if functions sharing the same slope but very different intercepts are judged more similar than functions with similar slopes and intercepts, such non-linear interactions could explain the multimodality observed in our experiment.

While more research is required, our results also highlight the importance of a plurality of experimental approaches and methods in the study of human generalization. Most previous research has focused on averaged errors or single extrapolations. Here, we suggest that to fully understand human generalization, we need to consider the interplay between errors, extrapolations, and the hypothesis spaces facilitating them.

Chapter 5

Transferring Functions and Parametrizations

Many everyday situations require us to generalize from experience, even if faced with a specific problem we have never seen before. For example, in cooking, one regularly has to infer the relationship between ingredients, ratios, or quantities, like the amount of sweetener and resulting pleasantness of a dessert, and generalize this relation to new recipes or ingredients. Often, we learn a general relationship that helps us understand related problems. When we learn that adding sugar to a dish will gradually increase its sweetness before saturating, we acquire knowledge that we can apply to similar relationships, such as deciding how much xylitol[1] to add to a cake.

The previous chapters have explored which representations allow us to generalize based on learned relationships. Here we expand this notion to situations in which past relationships themselves are generalized or transferred to a new situation. These situations expand the tasks humans face in classical function learning experiments and require further-reaching and more abstract inferences. Given a set of prediction tasks, how can we capitalize on statistical regularities

---

[1]A widely used sugar substitute.

to aid future prediction? If the tasks exhibit some shared structure, learning a representation capturing this latent structure of the environment (Gershman and Niv, 2010) or learning which aspects of a task change (Wilson and Niv, 2012) can enable the learner to perform wide-ranging and data-efficient generalization.

The value of transferring knowledge across different tasks is receiving growing attention in machine learning communities. For example, abstract learning and transfer have been successfully applied to challenging control tasks (Hamrick et al., 2017). From a cognitive science perspective, the study of such general learning mechanisms has a long tradition, (e.g., Harlow, 1949). Research in this tradition has highlighted how hierarchical representations can allow for the "blessing of abstraction" (Gershman, 2017b), where abstract knowledge is acquired faster than detailed information. Several proposals have been put forward on how hierarchical and structured inductive biases can be acquired through development and how they allow for rapid generalization (Goodman et al., 2008; Tenenbaum et al., 2011).

In function learning, the hierarchical and abstract representation of relationships has traditionally been reduced to mechanisms that allow generalizing a mapping from criterion to target. Here we will adopt a general perspective and express the task as Gaussian process regression. While Gaussian processes allow us to express inductive biases for functions in flexible, non-parametric fashion, only recently has more attention been given to structural and hierarchical aspects of function generalization. This work has emphasized the importance of inductive biases over different function types (Lucas et al., 2015; Wilson et al., 2015), the compositional structure of functions (Schulz et al., 2017), or the generalization of functions into dimensions outside the learned space (Lucas et al., 2012).

In Wilson et al., participants repeatedly extrapolated from a non-parametric smooth function. Participants progressively learned the features of this function and adapted their expectations to the learned data. However, while partici-

pants in Wilson et al. adapted their expectations, their final extrapolations were still skewed towards prior expectations for correlational structure. Similarly, in Reimers and Harvey (2011), participants had to forecast time series repeatedly. Participants adapted their predictions to the correlation structure in the data but exhibited strong inductive biases for positive autocorrelations. These results led Reimers and Harvey (2011) to suggest that participants update their prior expectations based on the structure in the data.

Here we expand on this line of research and propose that when humans learn relationships, they do not maintain sets of data, parametrizations, or fixed parametric forms. Instead, they form flexible and abstract hypothesis spaces. Based on this abstract encoding, they can capitalize on statistical co-occurrences of abstract information about the *type* of relationship learned. As a result, repeated exposure to similar functions should result in learning about the shared type of relationship and its relevant features. Such exposure should then facilitate extrapolation in sparse contexts and allow far-ranging generalization. We hypothesize that this application of past knowledge does not merely amount to remembering previous data, but productive extrapolation depends on the adaptation of the learned abstract function type to the context at hand.

5.1 Experiments

We ran two sets of experiments. In all experiments, participants were trained on a set of three function realizations. They either had to select from a set of candidate patterns the option they deemed the most consistent with this training (Forced-Choice Experiment) or had to extrapolate (Extrapolation Experiment). In the forced-choice experiments, we also contrasted the participants' preferences with a control condition in which no training was provided. We used both extrapolations and choices as dependent variables, as they provide complementary

insight into participants' inferred functions. Forced choices directly present participants with the alternative extrapolation patterns and, at the same time, allow us to adopt simple statistical tests to evaluate which patterns, from the limited set of candidates, were preferred. Extrapolation tasks are more flexible, but at the same time are more challenging to quantify, as participants tend to produce idiosyncratic patterns.

5.1.1 Procedure

Participants were instructed to learn the relationship between two nondescript substances, substance x and substance y. They were told that they would be presented with three sets of patterns, each depicting one realization of the same relationship and that they had to predict the relationship for ten new points. They also received a visual depiction explaining how they would predict the points. They were instructed that they would see one more pattern from the same relationship, consisting of three points. In the control task, participants were only instructed that they would be presented with a relationship between the two substances. Then, they immediately proceeded to the forced-choice test phase.

5.1.1.1 Training Phase

Each training block took the form of an extrapolation task: participants saw scatter plots and had to guess the value of the substance on the y-axis by selecting the height of the corresponding value on the plot. Participants were shown the correct value as feedback for one second, and, if their choice deviated by $\pm 2.5\%^2$ or more of the true value, had to readjust their selection. We presented the training data as an extrapolation task, since previous research has highlighted that testing aids extrapolation performance (Kang et al., 2011). Training blocks

---

[2]In relation to the total range of the extrapolation area.

were presented in randomized order.

5.1.1.2 Test Phase

After the training blocks, there was either a forced-choice task or an extrapolation task. Participants saw either the three- or one-point pattern and were instructed that this pattern belonged to the same relationship as the one in the training phase. In the forced-choice task, they then saw six scatter plot patterns corresponding to one conditional sample for each of the six candidate functions in randomized order. In the extrapolation task, they instead had to perform extrapolation given the pattern. In the forced-choice condition, participants had to select the pattern that they deemed the most likely extrapolation for the learned relationship. In the extrapolation task, participants received the same points that generated the conditional samples in the forced-choice condition and had to extrapolate for 30 values of x, without feedback, following the same procedure as in the training sets. The 30 extrapolation criteria were the same as those used to generate the forced-choices. After the test phase, participants completed a short demographic survey, were debriefed and compensated. For screenshots of the experimental stimuli and instructions, see Appendix D.

5.1.2 Materials

The functions in the six conditions corresponded to samples from Gaussian processes (GPs), with three different types of kernels and mean functions, each with two distinct parametrizations (see Table 5.1). To allow for characteristic periodic samples, we elected a "pure" cosine kernel, *Cos* with $k(r) = \sigma \times cos(r)$, $r(x, x') = \frac{(x-x')^2}{\ell_q^2}$, with an additional intercept. We generated linear samples from a linear kernel *Lin* with explicit slope and intercept terms. Finally, we used an Ornstein-Uhlenbeck kernel (OU) with an additional intercept to generate non-smooth samples.

Table 5.1: Kernels and kernel parameters generating the training data (variance σ, lengthscale λ, intercept β_0, and slope β_1). For all models, we set the residual variance to $\sigma = 0.01$.

	σ	λ	β_0	β_1
Lin_1	0.02	–	0.35	0.47
Lin_2	0.02	–	0.7	-0.47
Cos_1	0.05	0.1	0.5	–
Cos_2	0.05	0.04	0.5	–
OU_1	0.01	1	0.5	–
OU_2	0.08	1	0.5	–

5.1.2.1   Training Sets

We generated the training data by sampling three sets of 35 points, each in the range 0.05–0.95 for each of the six conditions. The first 25 points constituted the evidence provided in each training set. Participants had to extrapolate the target value for the last 10 points and received feedback for their choices. To ensure that samples were perceptible and the samples were distinct (within function type and between function types), we generated a set of 20 candidate patterns for the 18 sets. We then selected samples from these candidates for which all points were in the presentation range $[0, 1]$, which were ≥ 0.05 of the three transfer points. Finally, we also rejected visually uncharacteristic samples[3]. For a full list of kernels and kernel parametrizations, see Table 5.1; for the training data and the conditional samples, see Figure 5.1.

---

[3]For example, samples from the OU kernel that did not exhibit any discontinuities and were visually identical to linear relationships, or cos samples that had very low amplitudes.

(a) Positive linear function   (b) Slow periodic function   (c) Low-variance OU

(d) Negative linear function   (e) Fast periodic function   (f) High-variance OU
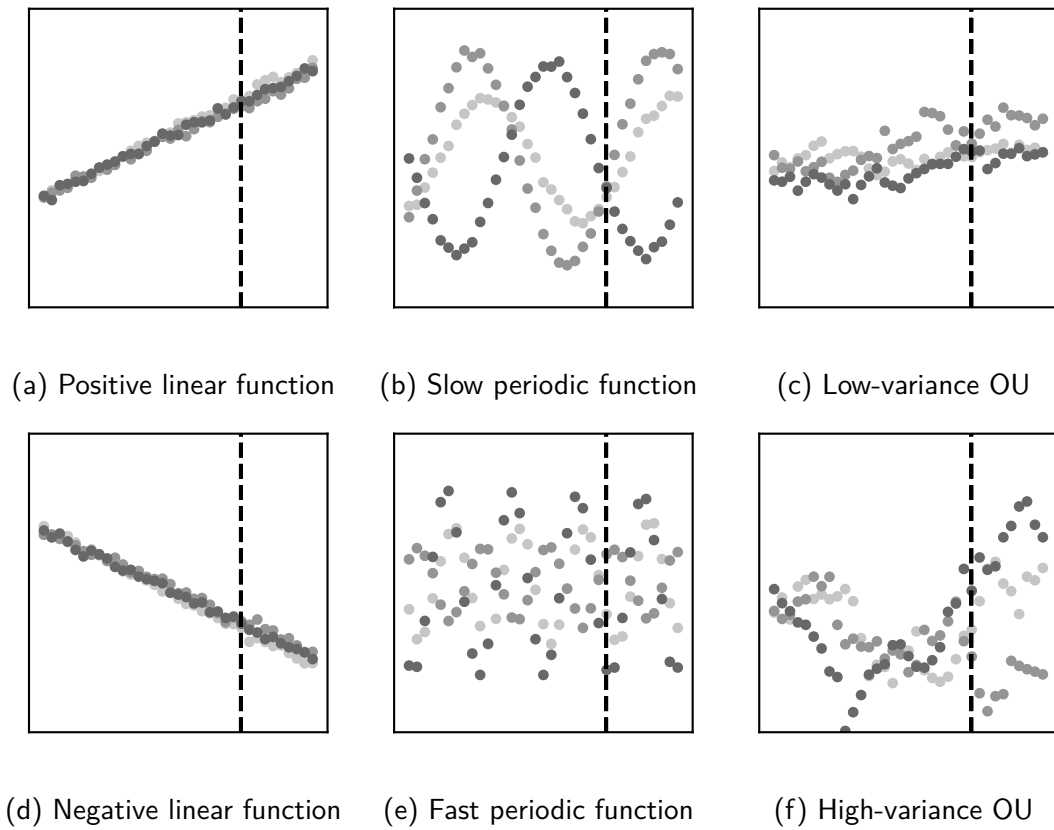
Figure 5.1: Training data in the six conditions. For each condition, there were three sets of points to be learned. Participants received the first 25 points and had to extrapolate for the ten remaining points. The dashed line is the cutoff between presented evidence and training.

5.1.2.2   Transfer Set

In the transfer set, either three points, $x = \{0.05, 0.1, 0.2\}$, $y = \{0.475, 0.525, 0.5\}$, or one point, $x = \{0.2\}$, $y = \{0.5\}$ were presented. These points were selected not to be strongly reflective of the training materials, in terms of specific point locations. We then generated three samples conditional on the transfer points for each of the six functions. Participants received one of these three samples at random for each of the six kernels in the forced-choice task. For the samples presented in the 3-point and 1-point forced-choice conditions, see Figure 5.2.

| (a) $\text{Lin}_1 - 3$ | (b) $\text{Lin}_2 - 3$ | (c) $\text{Lin}_1 - 1$ | (d) $\text{Lin}_2 - 1$ |

| (e) $\text{Cos}_1 - 3$ | (f) $\text{Cos}_2 - 3$ | (g) $\text{Cos}_1 - 1$ | (h) $\text{Cos}_2 - 1$ |

| (i) $\text{OU}_1 - 3$ | (j) $\text{OU}_2 - 3$ | (k) $\text{OU}_1 - 1$ | (l) $\text{OU}_2 - 1$ |
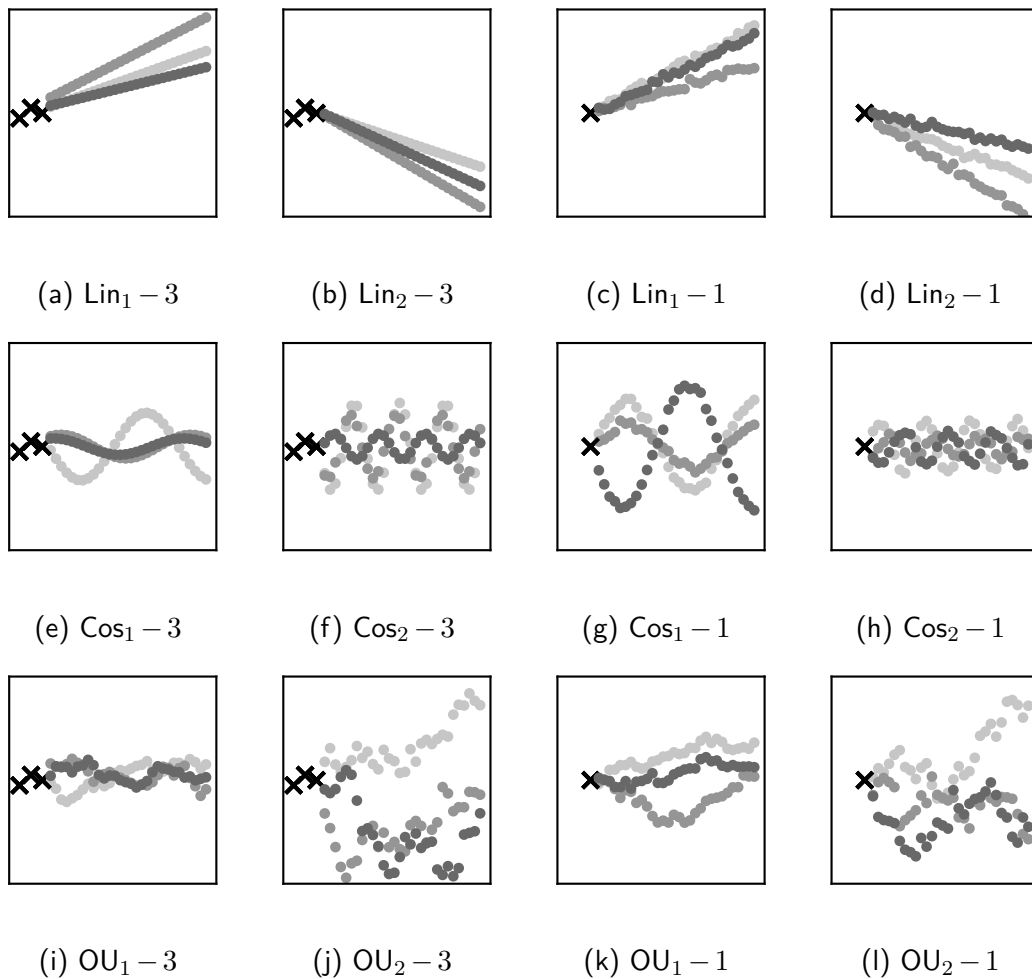
Figure 5.2: The materials presented in the 3-point and 1-point forced-choice experiments, as well as the control conditions.

5.1.3 Participants

5.1.3.1 Forced-Choice Experiments

We recruited a total of 291 participants ($M_{age} = 33.58, SD_{age} = 10.82$; 109 female, 181 male, 1 other) on Amazon Mechanical Turk. Participants had to have completed more than 50 approved tasks with an approval rate of 95% or higher. In total, 191 participants completed the experimental conditions ($N_{point=1} = 92, N_{point=3} = 99$) and 100 the control conditions ($N_{point=1} = N_{point=3} = 50$). Participants in the experimental conditions received \$0.55 for participation and took an average of 8 minutes ($M = 7.56$, $SD = 7.08$) to complete the experiment. Participants were randomly assigned to one of the six experimental conditions; for the resulting group sizes, see Table 5.2. In the control conditions, participants received \$0.20 for participation and took an average of 1.5 minutes ($M = 1.5$, $SD = 6.30$) to complete the experiment.

5.1.3.2 Extrapolation Experiments

We recruited a total of 184 participants ($M_{age} = 32.58$, $SD_{age} = 9.68$; 74 female, 110 male) on Amazon Mechanical Turk. Participants had to have completed more than 50 approved tasks with an approval rate of 95% or higher. Participants received \$0.65 for participation and took an average of 9 minutes ($M = 9.37$, $SD = 8.46$) to complete the experiment. Participants were randomly assigned to one of the six experimental conditions; for the resulting group sizes, see Table 5.2.

5.2 Results

5.2.1 Training Errors

We aggregated the training errors across forced-choice and extrapolation experiments, since participants were presented with the same training task for both

Table 5.2: Total number of participants in the forced-choice (N_{choice}) and extrapolation (N_{extrap}) conditions.

	Points	N_{choice}	N_{extrap}
Lin_1	1	16	16
Lin_1	3	16	15
Lin_2	1	16	16
Lin_2	3	17	16
OU_1	1	15	15
OU_1	3	17	16
OU_2	1	15	15
OU_2	3	16	14
Cos_1	1	15	15
Cos_1	3	17	15
Cos_2	1	15	16
Cos_2	3	16	15

conditions.

Mean absolute errors calculated on extrapolations before the participant had received feedback for that particular value differed considerably depending on the type of function presented in training. As expected, errors were lowest for linear conditions ($M_{\text{Lin}_1} = 0.02$, $SD_{\text{Lin}_1} = 0.01$; $M_{\text{Lin}_2} = 0.02$, $SD_{\text{Lin}_2} = 0.02$). The low-variance OU and the slow periodic condition also exhibited low mean errors ($M_{\text{OU}_1} = 0.03$, $SD_{\text{OU}_1} = 0.01$; $M_{\text{Cos}_1} = 0.03$, $SD_{\text{Cos}_1} = 0.03$). High-variance OU and fast periodic conditions exhibited the highest training errors, ($M_{\text{OU}_2} = 0.08$, $SD_{\text{OU}_2} = 0.03$; $M_{\text{Cos}_2} = 0.08$, $SD_{\text{Cos}_2} = 0.05$). For error quantiles per condition, see Figure 5.3.
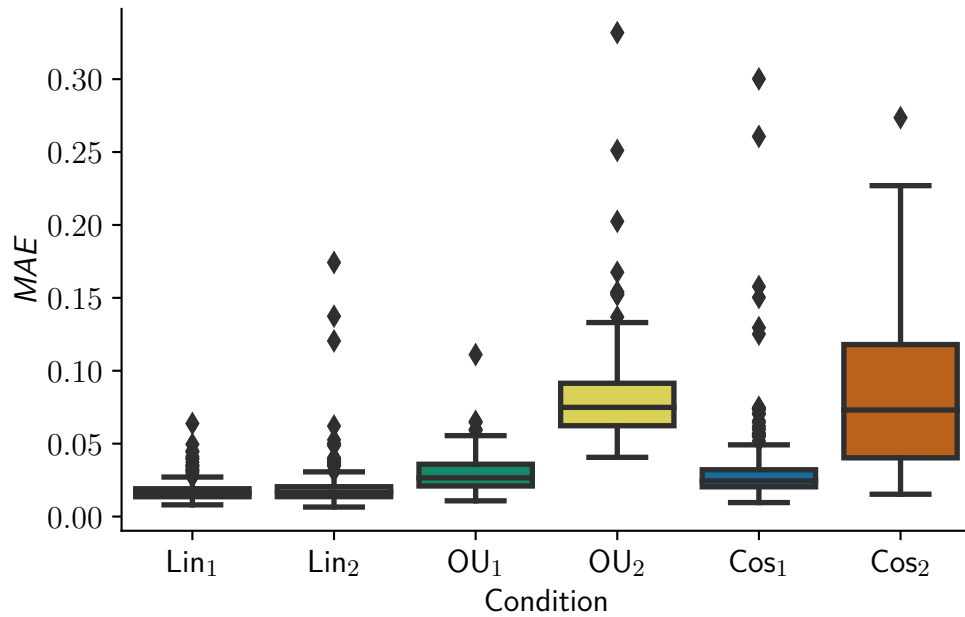
Figure 5.3: Averaged errors across the three training blocks. Errors were generally low and displayed low variability. However, for high-variance OU and the fast periodic function, errors were considerably higher and displayed larger variability. Boxplots display first, second (median) and third quartiles. Whiskers show the 1–5 interquartile range (*IQR*).

5.2.1.1   Error Decay

We can also assess the change in training error over the three training blocks. Note that in contrast to previous research, participants in our experiments did not see the same realization of a function. Instead, changes in error over the training blocks reflect the influence of an expectation for a particular type of function.
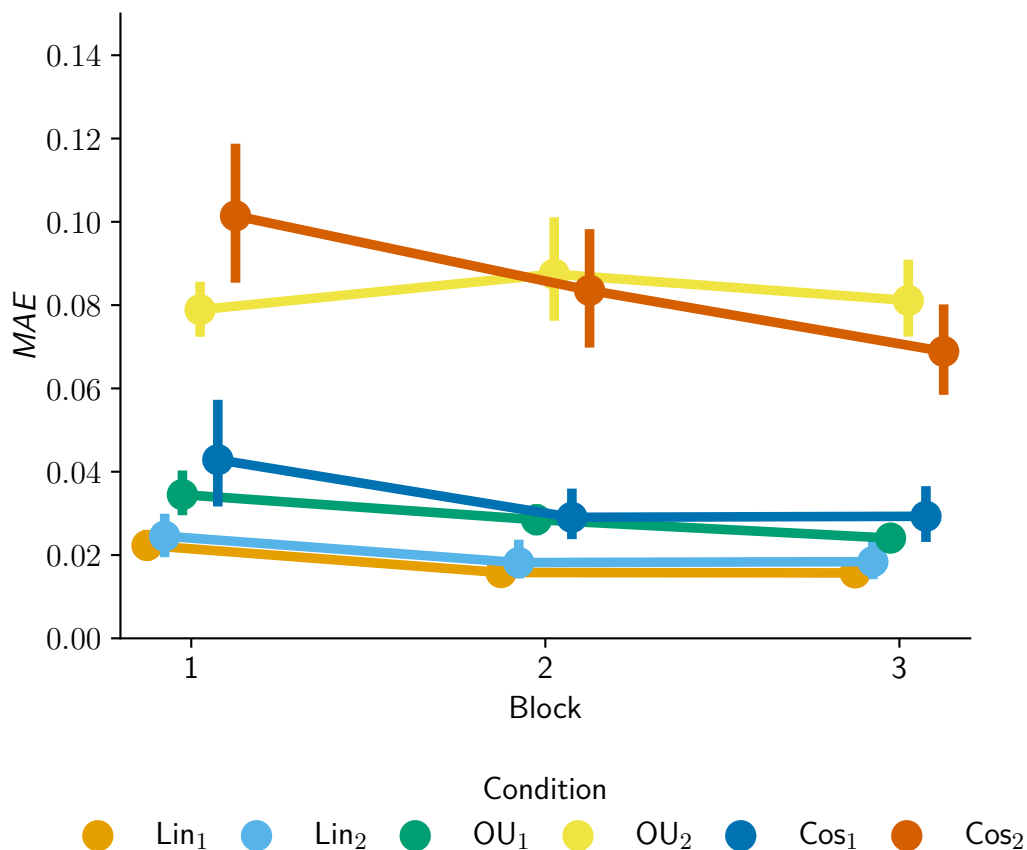
Figure 5.4: Mean absolute error across training blocks and conditions. Error bars display 95% bootstrapped confidence intervals. While both linear, low-variance OU and slow periodic functions were easy to learn, the fast periodic function and the high-variance OU conditions were not. While participants in the fast periodic condition improved somewhat over training, on average, participants in the high-variance condition did not decrease their training errors.

Previous work has modeled the error decay during training with a hierarchical

exponential-decay model (Kalish, 2013). We evaluated a variety of candidate models, including the exponential-decay model, and found that a hierarchical log-normal, $log(\text{error}) \sim \mathcal{N}(\mu, \sigma)$ model fit the data best.

As in Kalish (2013), our model is a hierarchical Bayesian model, with per-participant (s) variation in slopes and intercept $\mu = \beta_{0s} + block \times \beta_{1s}$. Participants' intercepts and slopes were drawn from two normal distributions, $\beta_s \sim \mathcal{N}_c(\mu_c, 1)$, pooled within their corresponding experimental condition c, with $\mu_c \sim \mathcal{N}(0, 10)$. For simplicity, per-participant variance was fixed ($\sigma_s = 1$) and overall error variance was shared, $\sigma \sim Cauchy^+(5)$. For details of the model and the parameter estimation, as well as comparison to alternative models, see Appendix D.2.

We obtained group-level intercepts β_{0c} and learning rates β_{1c}. Since we model error on the logarithmic scale, we transformed intercepts and slopes for interpretation, $\beta_0 = e^{\beta_0}$, $\beta_1 = e^{\beta_1} - 1$. Note that the slopes obtained amount to a percentage change in relation to unit changes in each block. Group-level intercepts for both linear conditions, the low-variance OU, and the slow periodic were generally low, and error decayed across blocks at about 12%. In contrast, for high-variance OU, initial errors were higher and did not change over training ($\beta_0 = 0.08$, $\beta_1 = 0.01$). Similar to high-variance OU, initial errors for fast periodic functions were high ($\beta_0 = 0.08$). However, these errors decreased significantly over training blocks at a rate of about 13%. For errors across the three training blocks, see Figure 5.4; for group-level intercepts and slopes estimated by our model, see Figure 5.5; for estimated parameters and highest posterior density intervals, see Table 5.3.

5.2.2 Choices

Overall, in the forced-choice experimental conditions, about 46% of the participants selected the correct function type and parametrization (88 out of 192) after training. In the 3-point experiment, approximately 35% chose the correct
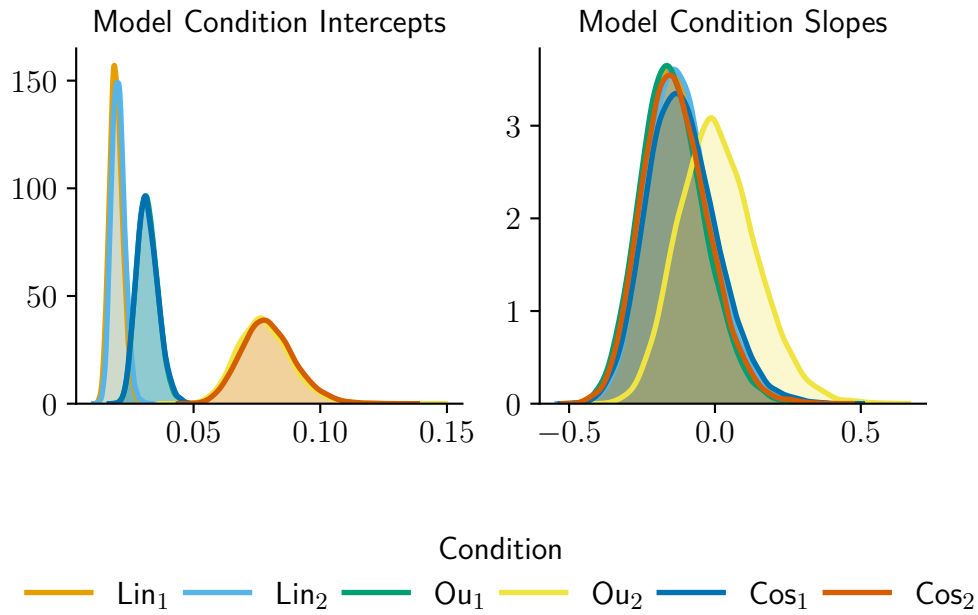
Figure 5.5: Group-level estimates of error intercepts and slopes estimated via the hierarchical Bayesian lognormal model. Both the high-variance OU condition and the fast periodic condition exhibit large initial errors in contrast to the remaining conditions. While the error for the fast periodic condition decreases over blocks, the error for the high-variance OU remains high.

Table 5.3: Group-level estimated means $\hat{M}$ for intercepts, β_0, and slopes, β_0, as well as 95% highest-posterior density intervals estimated via MCMC for the exponential decay model.

	$\hat{M}_{\beta_0}$	$\text{HPD}_{95}\beta_0$	$\hat{M}_{\beta_1}$	$\text{HPD}_{95}\beta_1$
Lin_1	0.02	[0.02, 0.02]	-0.12	[-0.34, 0.10]
Lin_2	0.02	[0.02, 0.02]	-0.12	[-0.34, 0.10]
OU_1	0.03	[0.02, 0.04]	-0.14	[-0.35, 0.08]
OU_2	0.08	[0.06, 0.10]	0.01	[-0.23, 0.29]
Cos_1	0.03	[0.02, 0.04]	-0.11	[-0.32, 0.12]
Cos_2	0.08	[0.06, 0.10]	-0.13	[-0.33, 0.11]

function and parametrization (35 out of 99).

In the 1-point condition, the proportion was higher, with approximately 57% of the participants selecting the correct function type and parametrization (53 out of 93). In the absence of training, in the 1-point condition participants preferred periodic functions (Cos_2 30%, 15 out of 50; Cos_1 28%, 14 out of 50) over OU (OU_1 18%, 9 out of 50; OU_2 14%, 7 out of 50). Only 10% of the participants selected the positive linear function (5 out of 50). The strong preference for periodic patterns suggest that participants interpreted the three points generating the sample to correspond to noiseless realizations of a low-amplitude periodic function, and not as intended noisy realizations of a flat linear function.

In contrast, in the 3-point control condition, 30% of the participants preferred positive linear functions (15 out of 50). They selected slow periodic functions in 24% cases (12 out of 50), and negative linear functions in 18% (9 out of 50). Low-variance OU and fast periodic functions were chosen in 12% (6 out of 50) and high-variance OU in 4% (2 out of 50).

The proportion of choice and the preference for particular functions differed

considerably across training function types and parametrizations. We fitted Dirichlet-Multinomial models to each condition to estimate the proportion of choices and contrast it with control preferences in both 3-point- and 1-point conditions. We contrasted the inferred true-option proportions with chance-level (1/6) and the corresponding proportions in the control condition. For more details on the model and the estimation procedure, see Appendix D.3.

5.2.2.1   3-Point Choices

In the linear conditions, participants preferred linear functions (Lin_1 69%, 11 out of 16; Lin_2 59%, 10 out of 17) and the correct parametrization specifically (Lin_1 44%, 7 out of 16; Lin_2 53%, 9 out of 17). In both conditions, they also chose low-variance OU, albeit at considerably lower rates (Lin_1 25%, 4 out of 16; Lin_2 24%, 4 out of 17).

In contrast, in the OU and periodic conditions, participants were not as homogeneous in their preferences. In the OU conditions, participants did select OU at slightly higher rates than alternatives (OU_1 47%, 8 out of 17; OU_2 56%, 9 out of 16) and selected periodic functions at similarly high rates (OU_1 41%, 7 out of 17; OU_2 38%, 6 out of 16). Participants preferred the correct parametrization in the high-variance OU condition (38%, 6 out of 16). They selected it over the low-variance alternative (19%, 3 out of 16), as well as both periodic options (each 19%, 3 out of 16). However, in the low-variance condition, participants predominantly selected the high-variance OUs and the slow periodic options (both 29%, 5 out of 17) over the low-variance option (17%, 3 out of 17).

For periodic conditions, periodic options were selected at higher rates than alternatives (Cos_1 59%, 10 out of 17; Cos_2 75%, 12 out of 16). For fast periodic functions, participants chose the correct parametrization in 50% of the cases (8 out of 16). They selected the slow alternative at lower rates (25%, 4 out of 16).

To confirm that the preferences for the option corresponding to the true func-

tion differed significantly from chance, we contrasted the proportion estimates of the Dirichlet-Multinomial model with random choice (1/6) and the control condition. For both linear conditions and the high-variance OU and the fast periodic, estimated proportions were considerably higher than chance, with $\geq 95\%$ of the estimated proportions larger than 1/6. Of those, all but Lin_1 were also larger than the proportion corresponding to the true choice proportion in the control condition ($\geq 99\%$). Only 50% of the estimates in the low-variance OU condition and 26% of the proportion estimates in the low-frequency periodic were higher than chance. When compared to the proportions obtained in the control condition, about 72% of the proportion estimates for Lin_1 and OU_1 were higher. Finally, less than 7% of the estimates in the low-frequency periodic condition were higher than the proportion in the control condition.

For the full set of posterior estimates, see Figure 5.7, and Table D.3; for proportions, see Figure 5.6.

5.2.2.2 1-Point Choices

In the 1-point linear conditions, participants preferred linear functions (Lin_1 69%, 11 out of 16; Lin_2 88%, 14 out of 16) and the specific parametrization (Lin_1 69%, 11 out of 16; Lin_2 69%, 11 out of 16). In contrast to the 3-point condition, they chose alternatives at much lower rates.

Similarly, for OU conditions, participants preferred OU-type functions (OU_1 80%, 12 out of 15; OU_2 50%, 8 out of 16). As in the 3-point condition, participants in the low-variance condition did not prefer the true OU parametrization (OU_1 40%, 6 out of 16) and chose high-variance OU equally often (6 out of 16, 40%). In contrast to the 3-point condition, they did not select low-frequency periodic options (7%, 1 out of 15) nor the other options frequently. In the high-variance OU condition, participants preferred the true parametrization (OU_2 44%, 7 out of 16) and less frequently chose high-frequency periodic options (25%, 4 out of
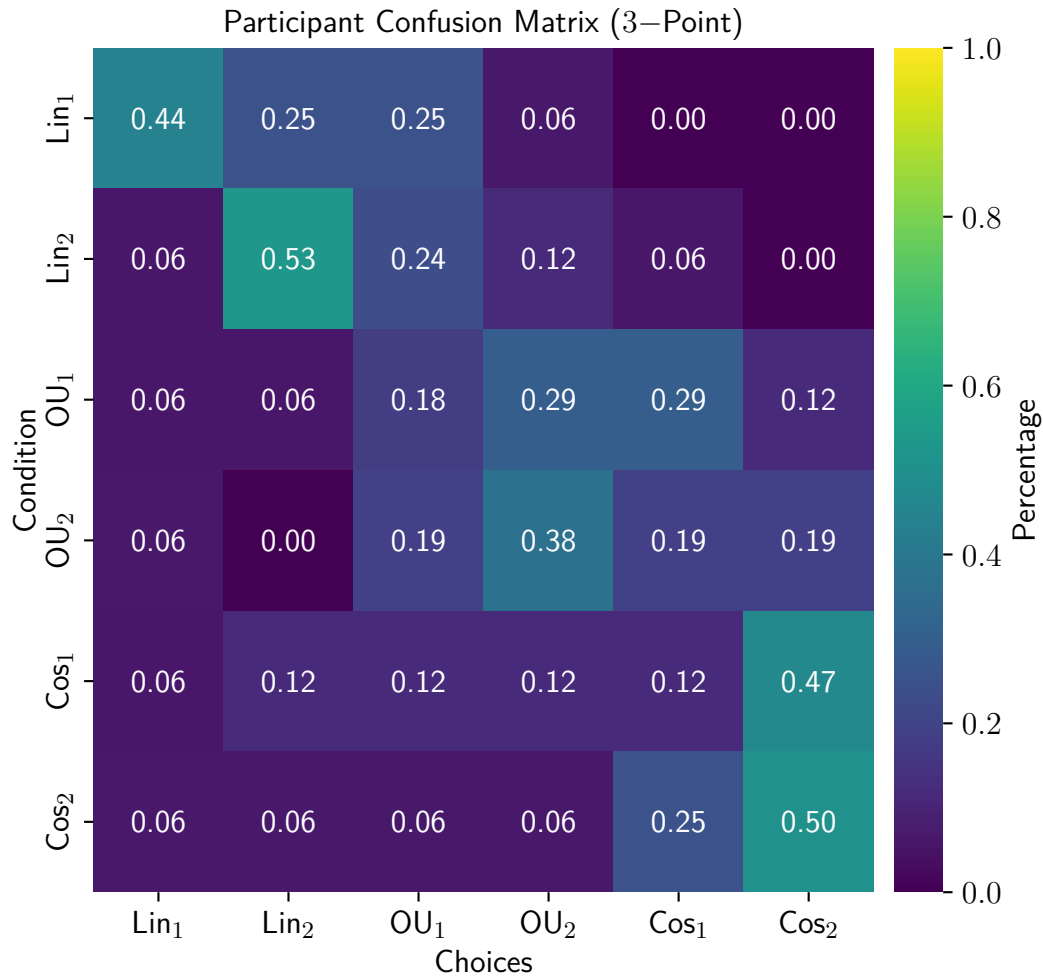
Figure 5.6: Proportions for each choice in the six experimental conditions. Partici-pants preferred high-frequency periodic samples over the true low-frequency samples. Similarly, participants in the OU_1 conditions preferred the high-variance samples, or even periodic samples, over the low-variance samples they saw in training.
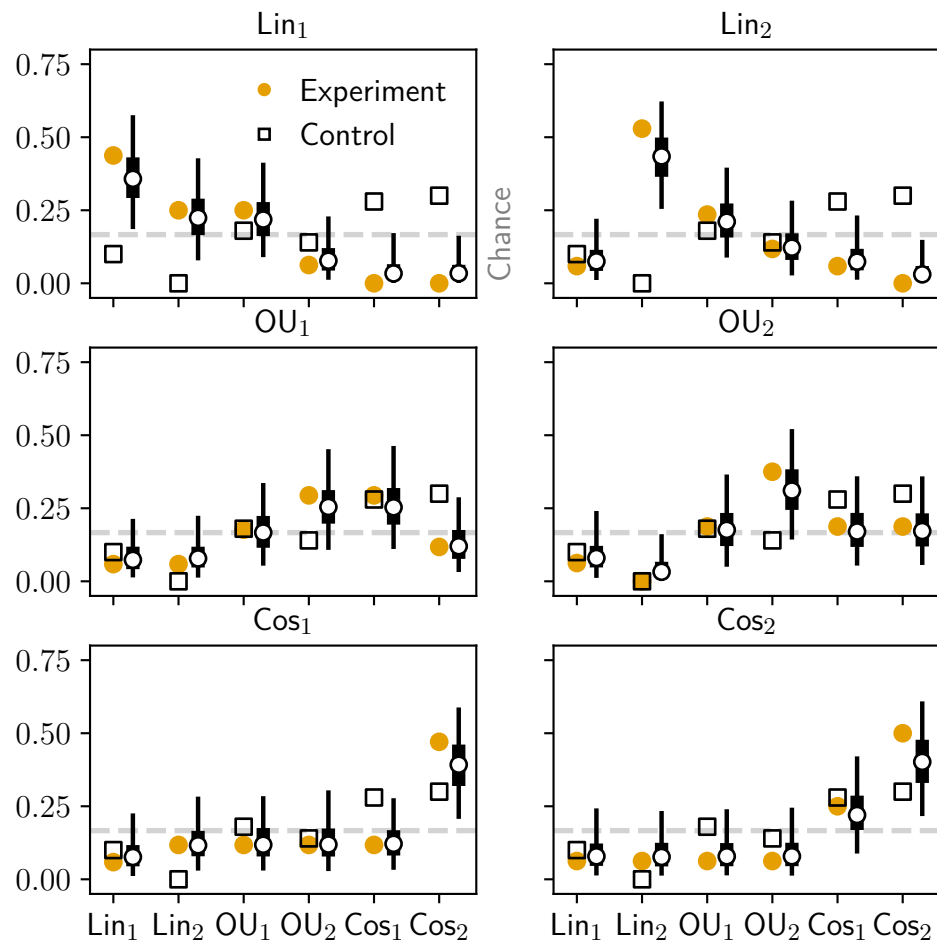
Figure 5.7: Proportion of options chosen in the six experimental conditions (round marks) and control (square marks) for the 3-point conditions. The trained function type and its specific parametrization were selected at rates higher than chance $(1/6)$. Proportion estimates for the options selected in the experimental conditions (round marks with 5–95% *IQR*) revealed that for the low-variance OU and the low-frequency periodic condition, participants did not prefer the true function at rates higher than the control condition. Instead, they preferred the alternative parametrization corresponding to the same function type.

16).

In the periodic conditions, participants preferred the true function types (Cos1 12 out of 15, 80%, Cos2 11 out of 15, 73%). In both conditions, they mainly selected the trained parametrization (Cos_1 67%, 10 out of 15; Cos_2 53%, 8 out of 15).

We again compared the models' estimated proportions to chance and the proportion in the control condition. The trained option was selected above chance in all conditions (95% of the estimated proportions in all conditions $\geq 1/6$). Furthermore, in all conditions, the estimated parameters were also higher than the control condition (95% $\geq$ control). For the full set of proportions, see Figure 5.8. For mean estimates and 95% highest posterior density intervals, see Figure 5.9, and Table D.4.

5.2.3   Extrapolations

Visual inspection of the extrapolations in both 3-point and 1-point conditions strongly suggested that the extrapolations reflected participants' training conditions. Slopes in the positive linear conditions were visually distinct from negative-slope Linear. These slopes suggested that participants extrapolated positively in the positive-slope condition and negatively in the negative-slope condition. Similarly, variance was visually higher in the high-variance OU conditions when contrasted with low-variance OU. Finally, periodic extrapolations in the fast conditions were visually suggestive of a higher frequency than the slow, low-frequency periodic conditions. For all extrapolations, see Figure 5.10.

5.2.3.1   Recovering Function Types from Extrapolations

To evaluate if these patterns were also well aligned with the generating models, and if samples reflected the differences in function parametrization, we performed maximum-likelihood estimation (*MLE*) for each participant and each generating
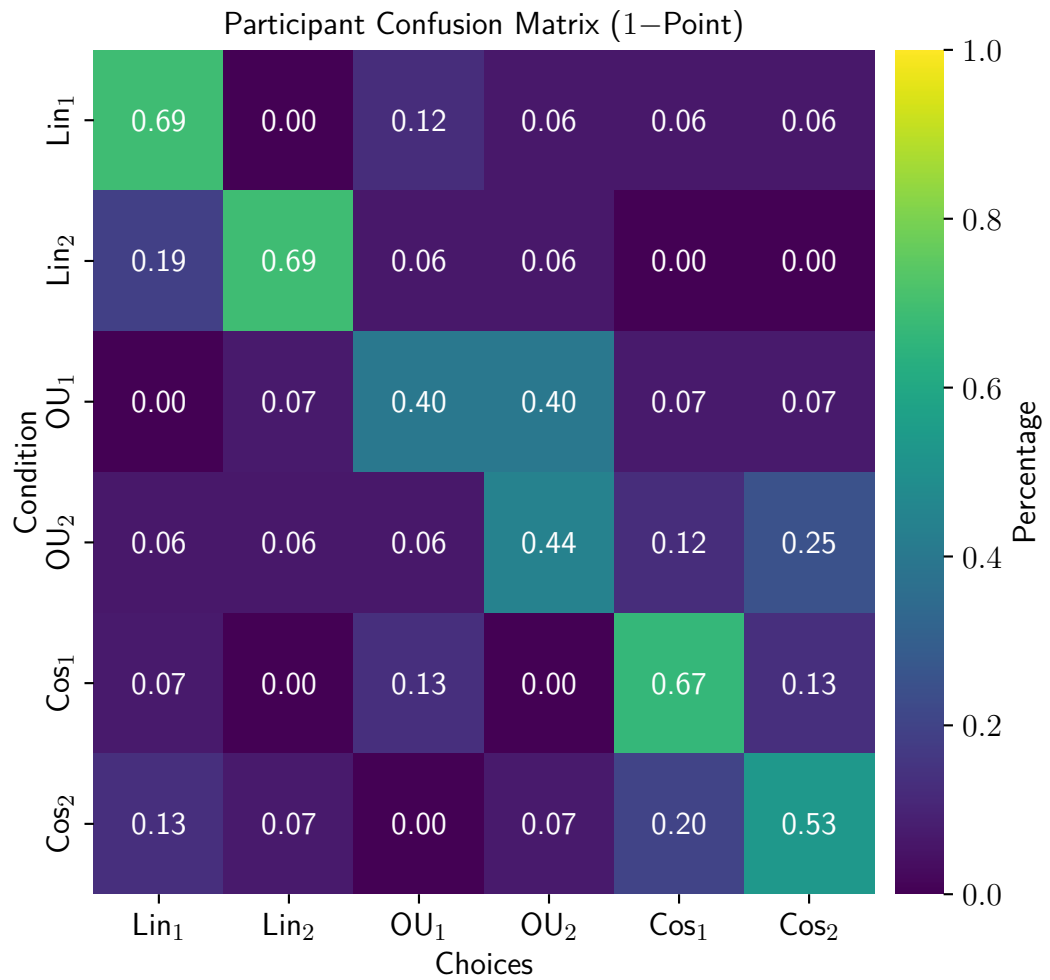
Figure 5.8: Proportions for each choice in the six 1-point experimental conditions. As in the 3-point condition, the trained option was selected above chance in all conditions. However, while both low-variance OU and fast periodic were selected more often than control, and at higher rates than in the 3-point condition, high-variance OU was chosen equally as often as low-variance OU.
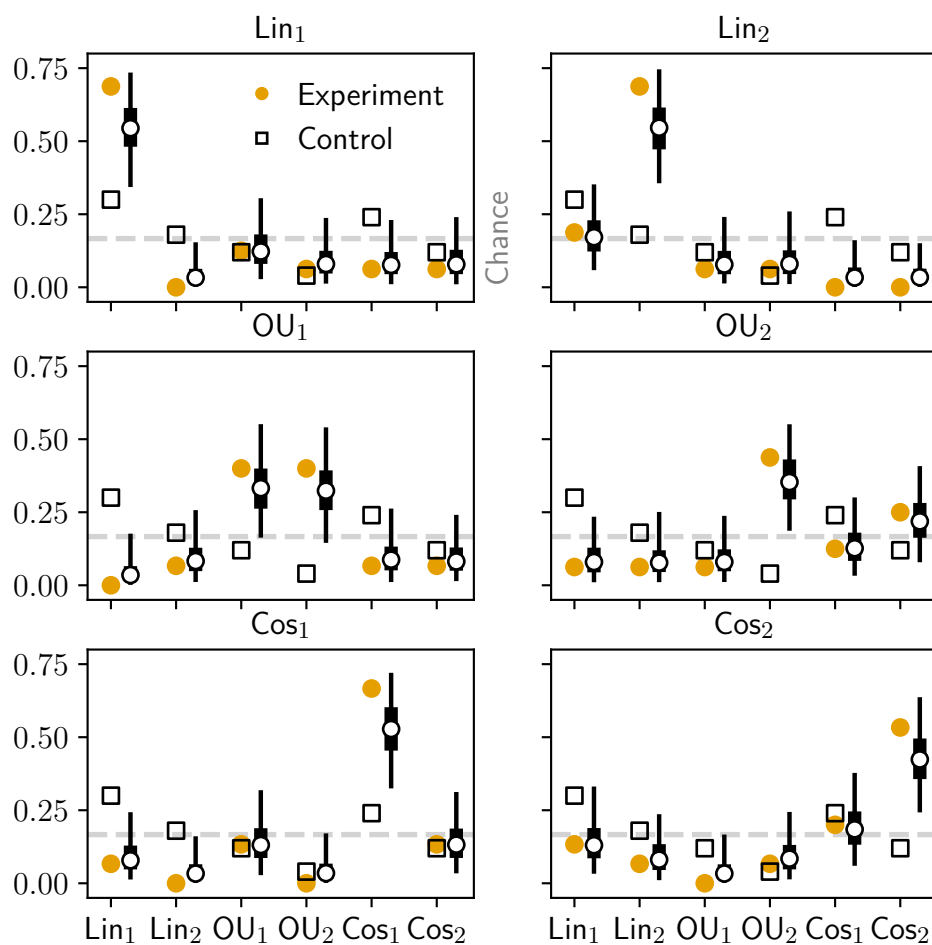
Figure 5.9: Proportion of options chosen in the six experimental conditions (round marks) and control (square marks) for the 1-point conditions. The trained function type and its specific parametrization was selected at rates higher than chance ($1/6$) in all conditions. Proportion estimates for the options selected in the experimental conditions (round marks with 5–95% IQR). For low-variance OU however, high-variance OU options were selected at equally high rates.
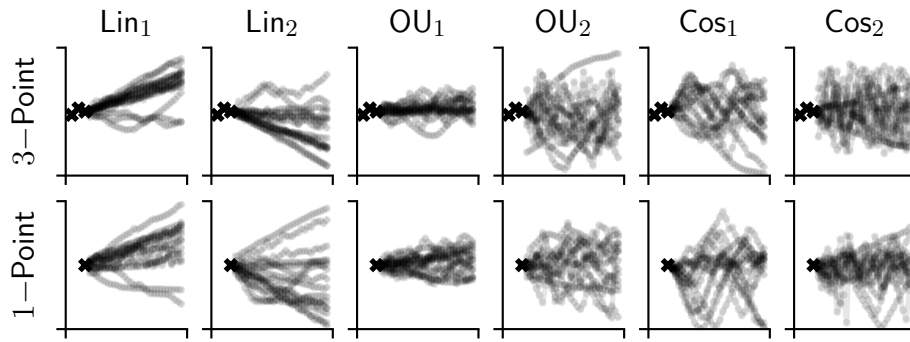
Figure 5.10: Extrapolations closely matched the learned function types, as well as concrete parametrizations.

GP[4]. That is, we fitted the three types of Gaussian process models to each participants' extrapolations. We then used the type of generating GP with the highest likelihood to predict which training samples the participant had received. This approach allowed us to evaluate if the experimental manipulation resulted in extrapolation patterns consistent with the generating GPs. Finally, we contrast the fitted parameters between conditions to assess if the extrapolations also reflect characteristic differences between the training conditions. For the five best-fitting extrapolations in each condition, see Figure 5.11.
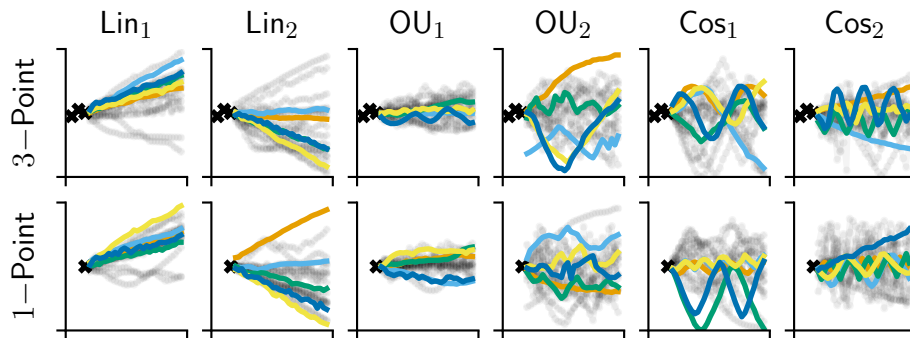
Figure 5.11: The five extrapolations with the highest likelihood scores in each condition.

Our classification procedure classified 8 out of 15 (53%) participants correctly

---
[4]We imposed range constraints for periodic and OU lengthscales $\in [0.001, 0.2]$ to aid the estimation and ran 50 L-BFGS optimization restarts.

in positive linear conditions, a proportion that was not significantly larger than expected by chance $(1/3)$, $p_{\text{Lin}_1} = .09^5$. In negative linear conditions, 8 out of 16 (50%) participants were classified correctly, a proportion that was not significantly larger than chance, $p_{\text{Lin}_2} = .13$. For OU, low-variance OU was classified correctly 10 out of 16 times (62%), $p_{\text{OU}_1} = .02$, and high-variance OU was classified correctly 12 out of 14 times (86%), $p_{\text{OU}_2} < .001$. Low frequency periodic functions were classified correctly 9 out of 15 times (60%), $p_{\text{Cos}_1} = .03$. Fast periodic functions were classified correctly 10 out of 15 times (67%), $p_{\text{Cos}_2} = .01$. For a confusion matrix of the classification, see Figure 5.12.
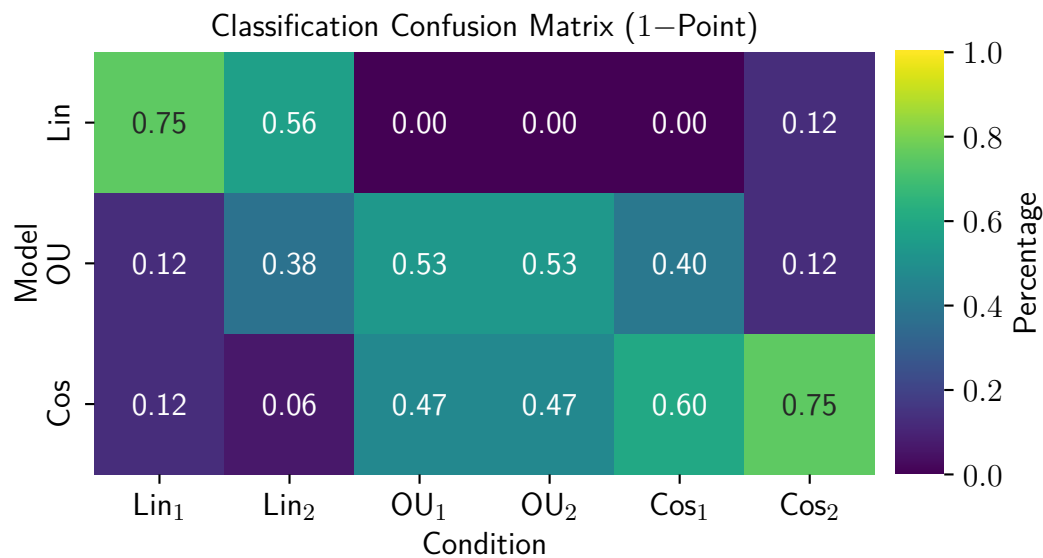
Figure 5.12: GP *MLE*s for each participant were used to predict which training samples the participant had been assigned to. This method was able to recover the training conditions.

For 1-point extrapolations, our classification scheme classified 12 out of 16 participants correctly in the positive linear condition (75 %), $p_{\text{Lin}_1} < .001$. For negative linear, only 9 out of 16 (56%) participants were correctly classified, a proportion that was only slightly larger than chance, $p_{\text{Lin}_2} = .05$. For both low- and high-variance OU conditions our scheme classified 8 out of 15 participants

---

^{5}All tests are one-sided, exact Binomial tests.

(53%) correctly, $p_{\text{OU}_1} = p_{\text{OU}_2} = .09$. For low- frequency periodic functions, 9 out of 15 participants (60%) were classified correctly, $p_{\text{Cos}_1} = .03$. For fast periodic functions, 12 out of 16 participants (75%) were classified correctly, $p_{\text{Cos}_2} < .001$. For a confusion matrix of the classification for 3-point extrapolations, see Figure 5.13.
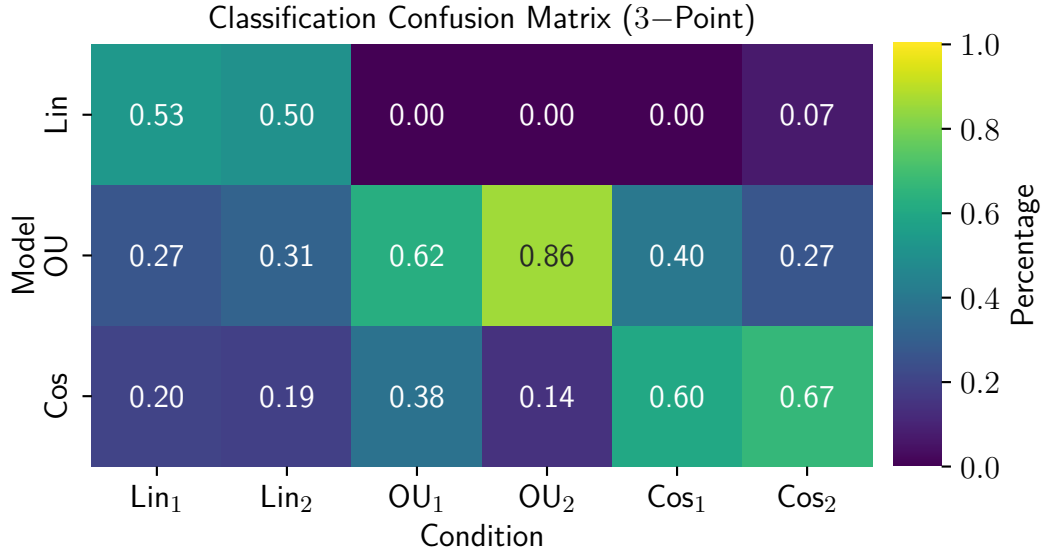
Figure 5.13: For 1-point extrapolations, *MLE*s were only able to recover the majority of participants in the linear and periodic conditions. Extrapolations in the OU conditions could not clearly be distinguished from those in periodic conditions.

5.2.3.2 Recovering Function Parameters from Extrapolations

To evaluated if the parameters of the best fitting model for each training function captured condition-specific parameter differences, we contrasted the parameters obtained via *MLE* for the true model. In the 3-point linear conditions, *MLE*-estimated parameters for slopes differed significantly between conditions, $M_{\text{Lin}_1} = 0.2$, $SD_{\text{Lin}_1} = 0.25$, $M_{\text{Lin}_2} = -0.1$, $SD_{\text{Lin}_2} = 0.34$, $t(27.49) = 2.82$, $p = .01$[6], with the signs of the inferred slopes matching the training. Neither intercept, variance or noise estimates differed significantly between conditions (all $p > .1$).

---

[6]All tests in this section are unequal variance, two-sided *t*-tests.

The inferred parameters for variance in the OU conditions did not differ significantly, but were reflective of differences in training, $M_{\text{OU}_1} < 0.01$, $SD_{\text{OU}_1} < 0.01$; $M_{\text{OU}_2} = 0.01$, $SD_{\text{OU}_2} < 0.01$; $t(15) = -1.95$, $p = .07$. The inferred lengthscale did not differ significantly between conditions, but was slightly higher for $OU1$, $M_{\text{OU}_1} = 0.38$, $SD_{\text{OU}_1} = 0.39$; $M_{\text{OU}_2} = 0.21$, $SD_{\text{OU}_2} = 0.28$; $t(26.31) = 1.46$, $p = .16$. Both intercept and noise estimates did not differ significantly between conditions (all $p > .5$).

The inferred parameters for periodic conditions did not differ significantly for lengthscale, $M_{\text{Cos}_1} = 0.08$, $SD_{\text{Cos}_1} = 0.06$; $M_{\text{Cos}_2} = 0.08$, $SD_{\text{Cos}_2} = 0.1$; $t(22.47) = 0.27$, $p = .79$. Instead, conditions differed significantly for variance, $M_{\text{Cos}_1} = 0.02$, $SD_{\text{Cos}_1} = 0.02$; $M_{\text{Cos}_2} = 0.01$, $SD_{\text{Cos}_2} = 0.01$; $t(20.1) = 2.25$, $p = .04$. Estimates for intercepts and noise were not significantly different between conditions (all $p > .1$).

The parameters for linear slopes in the 1-point condition again differed significantly between conditions, $M_{\text{Lin}_1} = 0.29$, $SD_{\text{Lin}_1} = 0.18$; $M_{\text{Lin}_2} = -0.2$, $SD_{\text{Lin}_2} = 0.22$; $t(28.94) = 7.09$, $p < .001$. In contrast to the 3-point conditions, intercepts did also differ significantly, $M_{\text{Lin}_1} = 0.19$, $SD_{\text{Lin}_1} = 0.04$; $M_{\text{Lin}_2} = 0.3$, $SD_{\text{Lin}_2} = 0.04$; $t(29.81) = -6.7$, $p < .001$. However, variance did not differ significantly, $M_{\text{Lin}_1} = M_{\text{Lin}_2} < 0.001$, $SD_{\text{Lin}_1} = SD_{\text{Lin}_2} < 0.01$, $t(29.79) = 0.34, p = .74$. For parameter estimates for 3- and 1-point conditions, see Figure 5.14.

MLE estimates for OU in the 1-point conditions did not differ significantly. While noise was higher for high-variance conditions, $M_{\text{OU}_1} < 0.001$, $SD_{\text{OU}_1} < 0.01$; $M_{\text{OU}_2} = 0.001$, $SD_{\text{OU}_2} < 0.01$, this difference was not significant, $t(14.02) = -1.4$, $p = .18$. For parameter estimates for 3- and 1-point OU conditions, see Figure 5.15.

Similarly, neither estimated intercepts, $M_{\text{OU}_1} = 0.52$, $SD_{\text{OU}_1} = 0.04$; $M_{\text{OU}_2} = 0.47$, $SD_{\text{OU}_2} = 0.12$; $t(16.82) = 1.6$, $p = .13$, nor lengthscales, $M_{\text{OU}_1} = 0.11$, $SD_{\text{OU}_1} = 0.08$; $M_{\text{OU}_2} = 0.13$, $SD_{\text{OU}_2} = 0.07$; $t(27.75) = -0.49$, $p = .63$, differed
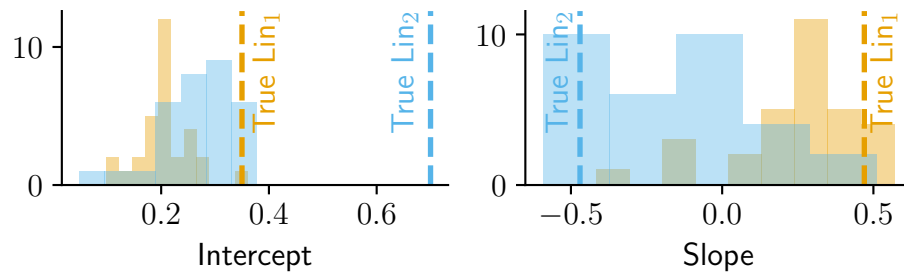
Figure 5.14: *MLE* estimates for the linear GP model for extrapolations in the linear 1- and 3-point conditions. Overall, the estimated parameters reflected the difference in slopes and intercepts in the training materials.
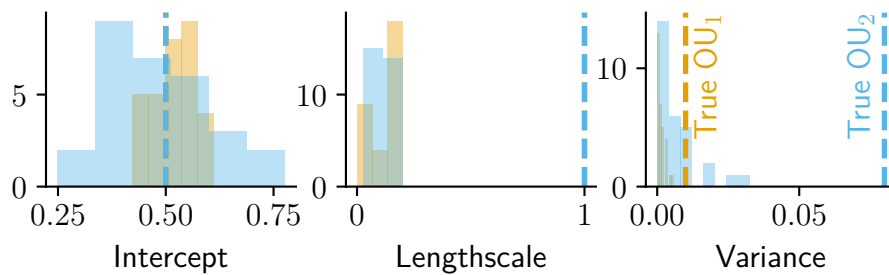
Figure 5.15: *MLE* estimates for the GP model for extrapolations in the OU 1- and 3-point conditions. The estimated parameters could not accurately distinguish the two experimental conditions.

significantly.

For periodic functions, there were again no significant differences between conditions. While lengthscales were slightly higher for Cos_1, $M_{\text{Cos}_1} = 0.85$, $SD_{\text{Cos}_1} = 0.05$; $M_{\text{Cos}_2} = 0.6$, $SD_{\text{Cos}_2} = 0.06$, this difference was not significant, $t(28.38) = 1.18$, $p = .25$. Similarly, neither variances, $M_{\text{Cos}_1} = 0.02$, $SD_{\text{Cos}_1} = 0.01$; $M_{\text{Cos}_2} = 0.01$, $SD_{\text{Cos}_2} = 0.01$; $t(28.87) = 0.54$, $p = .59$, nor lengthscales, $M_{\text{Cos}_1} = 0.08$, $SD_{\text{Cos}_1} = 0.05$; $M_{\text{Cos}_2} = 0.06$, $SD_{\text{Cos}_2} = 0.06$; $t(28.38) = 1.18$, $p = .25$, differed significantly. For parameter estimates for 3- and 1-point periodic conditions, see Figure 5.16.
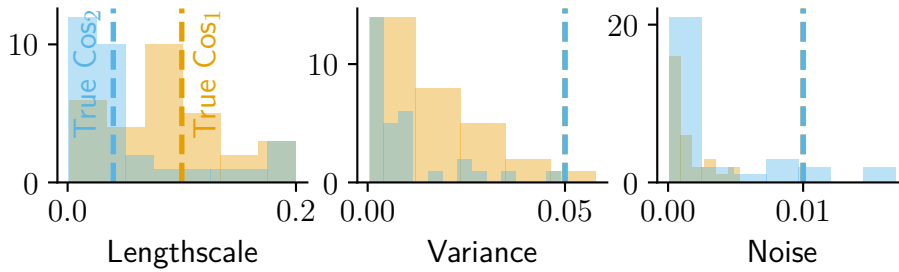
Figure 5.16: *MLE* estimates for the periodic GP model for extrapolations in the Cos 1-point and 3-point conditions. Only in the 3-point conditions did parameter estimates accurately reflect differences in conditions.

5.3   Discussion

Training difficulty reproduced previous results in function learning, with positive linear functions being easy to learn and non-monotonic and random-patterns being more challenging. However, our results also highlight that this difficulty is graded, with low-variance OU and low-frequency periodic functions exhibiting lower errors than their counterparts. These results suggest that differences in the ability to learn particular functions rest not only on the general functional form but also on the parametrization of those functions.

While our results regarding training difficulty expand previous notions of learning difficulty, the main contribution of this chapter is in the analysis of the choices and extrapolations after training. In general, we found that participants choose patterns and extrapolated in ways consistent with the learned function type. Across both 3- and 1-point conditions, participants selected the correct function type well above chance. These type-specific differences were also reflected in the transfer sets, where extrapolations reflected relevant features, such as trends, periodicity, or variance.

While participants' judgments generally reflected the functions they learned during training, our results also highlight inherent human biases. Most notably, in the 3-point choice condition, participants preferred fast periodic samples over the true low-frequency samples. Similarly, participants in the Ou_1 conditions preferred the higher variance samples, or even periodic samples, over the trained low-variance samples. In the 1-point choices, these biases were slightly less pronounced, but participants still chose high-variance OU at an equal rate as low-variance OU options. One explanation for these biases could be that people have a strong preference for particular functions because these parametrizations are well adapted to environmental regularities. As a result, these functions would be robust and applicable to a wide range of tasks in the environment. This explanation would be consistent with recent results in human exploration, where participants tended to undergeneralize spatial correlations. However, this undergeneralization resulted in comparable or even better performance than a ground-truth matching model (Wu et al., 2018).

Visually, extrapolations reflected the functional type and the corresponding parametrization of the training. However, recovering the experimental conditions based on model comparison proved challenging and was only entirely successful in the 3-point condition. For 1-point extrapolations, our models were not able to differentiate the OU and periodic extrapolations. One possible explanation for

this difficulty could be that participants struggled to express the randomness of
the OU conditions, and their extrapolations often exhibited periodicity. Similarly,
small errors or drifts in the periodicity would have proven difficult to capture
for our simple periodic model. Thus, comparing the inferred parametrizations
between function-types was limited by the general difficulty to capture the highly
variable and complex functional forms. Nevertheless, parameter estimates, while
highly dispersed, overall reflected differences in conditions.

   We take our results to suggest that when humans learn consistent sets of rela-
tionships, they can learn the abstract structure or *type* of a family of relationships,
and exploit that knowledge to improve their ability to learn and generalize in the
future, especially in the face of sparse or ambiguous data. Participants were able
to apply this learned regularity to perform rapid and flexible generalization based
on the shared abstract relations in the training data.

   Future research should more closely examine which statistical patterns can be
generalized and under which circumstances these generalizations are performed.
For example, while our experiment imposed that all patterns followed the same
relationship, in reality, this information is rarely available. Thus, future research
should examine under which circumstances task regularities are inferred to be
similar, and what kinds of notions of similarity can guide these generalizations.

   A second question relates to the space of functions people can infer. Previous
research has treated this space as a closed set of flexible hypotheses. However,
these approaches have to continually revise and expand the set of hypotheses
with ad-hoc alternatives. One exciting alternative prospect is to link notions
of hierarchical and compositional representations to function generalization. If
people can combine hypotheses to form complex functions, this would alleviate
the need to include all potentially relevant functions into the hypothesis space.
We will explore compositionality in function learning in the next two chapters.

Chapter 6

Generalizing Function Compositions

In the previous chapters, we have shown that function learning operates in highly structured and abstract hypothesis spaces. These spaces over functional forms provide reusable and generative models that can be rapidly adapted to new situations. So far, the structure of these hypothesis spaces has been flat – people might learn how to weigh their prior expectations about what kind of functions to apply, but the general structure of the space remains unchanged. Here, we examine this assumption in more detail and evaluate if people can perceive continuous relationships as compositions of simpler constituents, and if this knowledge can be used productively.

Function extrapolation amounts to inferring the underlying generative process producing the data, and using that process to predict new values. If these functions are drawn from a space of alternative functions — a hypothesis space — there are two levels at which people can learn: individual function parameters and prior expectations for function types.

First, people can learn the parameters of individual functions. For example, when repeatedly encountering moderately steep linear relationships, one might update prior expectations for linear relationships to reflect the steepness. Computational models that do not consider a hypothesis space of alternatives are limited

to this level of parameter learning, as they cannot distinguish learning about one type of function from learning about other types of functions. That is because these models adopt a one-size-fits-all mechanism to function learning, and thus all forms of extrapolations are performed by the same mechanism. Consider, for example, one of the most prominent function learning models, EXAM (McDaniel and Busemeyer, 2005). EXAM proposes that associative weights are learned for trained value pairs. If values outside of this range have to be predicted, EXAM extrapolates linearly given the closest known value, successfully capturing human data (McDaniel and Busemeyer, 2005). However, the model has no mechanism to store the learned weights and parameters. Thus, if a different pattern is encountered, the model must be retrained, and no previous information can be maintained.

Second, people can learn to weigh prior expectations about the functions. For instance, repeatedly encountering cyclic patterns could result in updating one's prior expectation of these relationships. Computational models, such as the (theoretical) model of Brehmer (1974) or the more recent models by Lucas et al. (2015), allow for this kind of learning.

While both types of learning would allow one to adapt their expectations about the types and shapes of functions in the world, in both cases, learning is domain-general. For example, if one learns that atmospheric measurements often exhibit seasonal patterns, updating one's beliefs about seasonality would influence one's expectations about all other relationships. The results of Chapters 3 and 5 suggest that people do not learn functions in domain-general ways, but in domain-specific ways. Introducing a higher-order structure to the hypothesis space would allow people to form expectations about functions in domain-specific ways. These beliefs abstract the current task and allow the agent to learn high-order generalizations, or overhypotheses (Goodman, 1983; Kemp et al., 2007).

Overhypotheses amount to hypotheses over hypotheses; beliefs about what

kinds of structures are plausible in a particular context or task. In turn, acquiring these expectations over alternative hypotheses enables the learner to effectively generalize to novel situations, as the space of plausible hypotheses, an in-principle infinite space, is heavily constrained. Overhypotheses allow children and adults to learn effectively and underpin the human ability to form complex structures like biological taxonomies, physical or psychological theories, and causal inferences (Gopnik and Wellman, 2012; Kemp et al., 2007; Lucas and Griffiths, 2010).

While the notion of a hypothesis space and its extension to domain-specific hypothesis spaces allows for flexible learning, these approaches cannot satisfactorily answer how people can infer a wide range of functions. For instance, experiments by Wilson et al. (2015) showed that people could learn unconventional functions, such as saw-like patterns. These functions would have to be included a priori in the hypothesis space, in addition to the many other functions that experiments have found people can learn and infer. Furthermore, many common patterns exhibit even more complex patterns. For example, the atmospheric measurements mentioned earlier do often exhibit seasonality and additional features, such as trends, changepoints, or further smooth or rugged variation. Treating the hypothesis space as a set of a priori available functions would thus have to include all of these combinations and variations, as parts of the space.

One solution to this issue is to suggest that the space of candidate function is not structured as a weighted set of hypotheses but is a generative compositional process. This process would take a set of atomic functions and compositional operators to produce atomic or composed functions as a hypothesis. Two potential compositional operators in function learning are the addition or multiplication of simpler functions, but more sophisticated operations would be possible.

Compositionality is the idea that complex structures can result from the combination of simple elements. Compositionality has been discussed most notably in linguistics and is the fundamental principle of syntax. In language, composi-

tionality is the mechanism by which a finite set of words can produce an infinite variety of sentences (Chomsky and Halle, 1965). Similarly, when interpreting other agents' behaviors, we naturally decompose them into goals, motivations, and beliefs (Jara-Ettinger et al., 2016), and classes of objects are decomposed into parts and functions (Kemp, 2012; Dechter et al., 2013; Lake et al., 2012). Compositionality is a crucial ingredient of intelligence — it allows productive and adaptive behavior. If phenomena in the world have compositional structure, it is advantageous for a rational agent to detect and adopt compositional representations, as previous experiences can be productively reused and recombined (Griffiths et al., 2009; Griffiths, 2017; Ullman et al., 2016).

As a motivating example, imagine trying to book accommodation ahead of a conference in an expensive city. To predict changes in rental prices as the conference approaches, one has to be able to generalize from past experiences and infer a mapping between the proximity of the conference date and the price. However, we cannot only use previous experiences about how time and price relate, but we can also decompose the inferred function into constituents. For example, we know that the prices fluctuate according to common temporal patterns; prices typically increase as the conference date approaches. Moreover, there might be changes from day to day such that it could be more expensive to book a place on a Sunday than on a Tuesday when most people are at work. Knowledge about which patterns combine and how quantities relate can aid flexible generalization. For example, if we assume that weekday prices and proximity of the date influence price additively, we can infer that the Sunday before the conference should be particularly pricey. Moreover, jointly learning relationships and the compositional rules underlying them allows us to make a wide variety of predictions about new patterns. For example, if we expect declining prices and daily fluctuations, we can infer what their composition will look like and predict accordingly.

Compositionality has not received much attention in previous function-learning

research, with notable exceptions of models that suggested that inferred functions were composed of local linear functions (Kalish et al., 2004), or hybrids of parametric and non-parametric processes (DeLosh et al., 1997). These approaches were compositional in that the learned functions were decomposed into local (linear) experts, or into different extrapolation processes responsible for interpolation and extrapolation. More recently, one paper explored the human ability to infer compositional structure in functions. Schulz et al. (2017) showed that participants generally preferred and extrapolated in ways that were better accounted for by compositional (Duvenaud et al., 2013) than non-compositional models (Wilson and Niv, 2012). Here we expand on that work, exploring if people can detect compositional structure in patterns and if the rules underlying these compositions themselves can be flexibly generalized to new situations.

We hypothesize that people see compositional structure in patterns and, given that they perceive compositional structure, they can perceive structural similarity between several patterns. For instance, if people perceive several complex patterns as governed by a linear trend and an additive seasonal pattern, they can infer this structure to be relevant to the class of patterns. As a result, they might assume that novel relationships belonging to the same pattern will also exhibit seasonality, a trend, and will compose additively.

6.1 Overview of the Experiments

Our setup is a rule-inference task in which the composition of patterns in a training set has to be applied to the two new instances in the test set, similar to studies typically conducted in compositional rule-learning domains (e.g., Piantadosi et al., 2016). We will call a successful generalization of the rule *one-shot* generalization since only one rule-application is presented in the training set. Our definition of one-shot learning is thus slightly different from the common

use in machine learning, where the term usually refers to the number of training instances presented (e.g., one category instance; Lake et al., 2015).

In Experiments 1-3, participants were first shown two patterns generated by sampling from two distinct GPs. Using GPs to generate the materials is well-suited to produce function compositions, as additions and multiplications of GPs result in valid GPs (see Appendix A), and these compositions are intuitively interpretable. Samples were introduced as the sales patterns over time for a particular alien plant on the intergalactic market. Then, participants were told that by combining the two plants, an offspring plant could be produced, and a pattern corresponding to the sales of that offspring was shown. The offspring corresponded to the additive or multiplicative combination of the two GPs producing the parent patterns. These three patterns constituted the learning set, and the composition applied to the patterns was the training-set composition (addition or multiplication). Then two new patterns comprising the test set were presented. Participants then had to infer what kind of sales pattern the combination of the two patterns in the test set would produce. We presented participants with a set of four candidate sales patterns: one pattern corresponding to applying the same composition as the training set to the two test-set patterns (the true function), a pattern generated by applying the same training-set composition to the test set (the alternative function), and two patterns corresponding to samples from the test set constituents. For screenshots of the experimental stimuli and instructions, see Appendix E.

6.1.1   Generating Functions

The sales patterns were functions sampled from a GP, with a mean of zero, and the kernel was sampled from the set of *base kernels* or an additive or multiplicative composition of two base kernels. The base kernels used in Experiments 1 and 2 were the radial basis function (RBF) kernel that produces smoothly
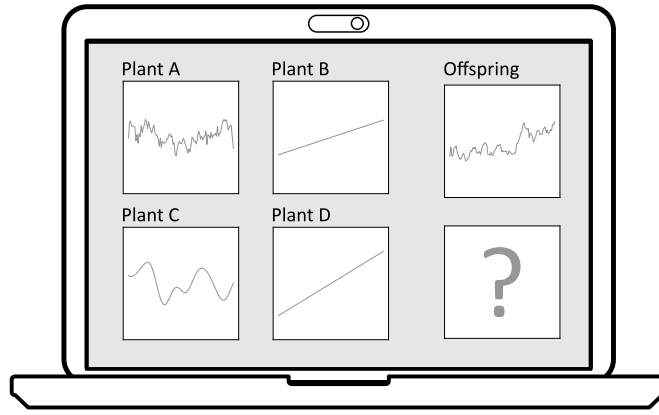
Figure 6.1: In the first three experiments, we presented sales patterns from two plants (plants A and B) and the sales pattern generated by their offspring. In this example, the patterns for plant A and B were sampled from GPs with OU and linear kernels. The offspring pattern was sampled from a GP with an OU+Linear kernel. After participants saw the three patterns in the training set, they were presented with patterns from two more plants (plants C and D) and had to infer the pattern of their offspring. In this example, the sales pattern for plant C was generated from an RBF kernel, and a linear kernel generated the pattern for plant D. If participants generalize the way plants combined in the training set to the test set, the offspring of plants C and D should resemble a sample from a GP with an RBF+Linear kernel.

varying patterns, the linear kernel, $k(x, x') = \sigma^2 (x - c) \times (x' - c)$ that produces linear functions, and a periodic kernel that produces smooth, periodic patterns, $k(x, x') = \sigma^2 \exp\left(-\frac{2 \sin^2(\pi x - x' \times p)}{\lambda^2}\right)$. In Experiments 3 and 4, the RBF kernel was replaced with an OU kernel. OU samples resemble rugged, random-walk-type patterns, $k(x, x') = \sigma^2 \exp\left(-\frac{x - x'}{2\lambda^2}\right)$[1]. All kernels had a variance σ^2 of 1 and kernel-specific parameters were set such as to produce discernible patterns ($\text{RBF}_\lambda = 0.1$, $\text{Periodic}_p = 1$, $\text{Periodic}_\lambda = 2$, $\text{Linear}_c = 2$, $\text{OU}_\lambda = 3$). For samples of the kernels and their compositions, see Figures 6.2 and 6.11.

---

[1]While the RBF kernel is infinitely differentiable, and therefore smooth, the OU kernel is non-differentiable and thus rugged.

(a) Linear+Cos   (b) Linear+RBF   (c) RBF+Cos

(d) Linear×Cos   (e) Linear×RBF   (f) RBF×Cos
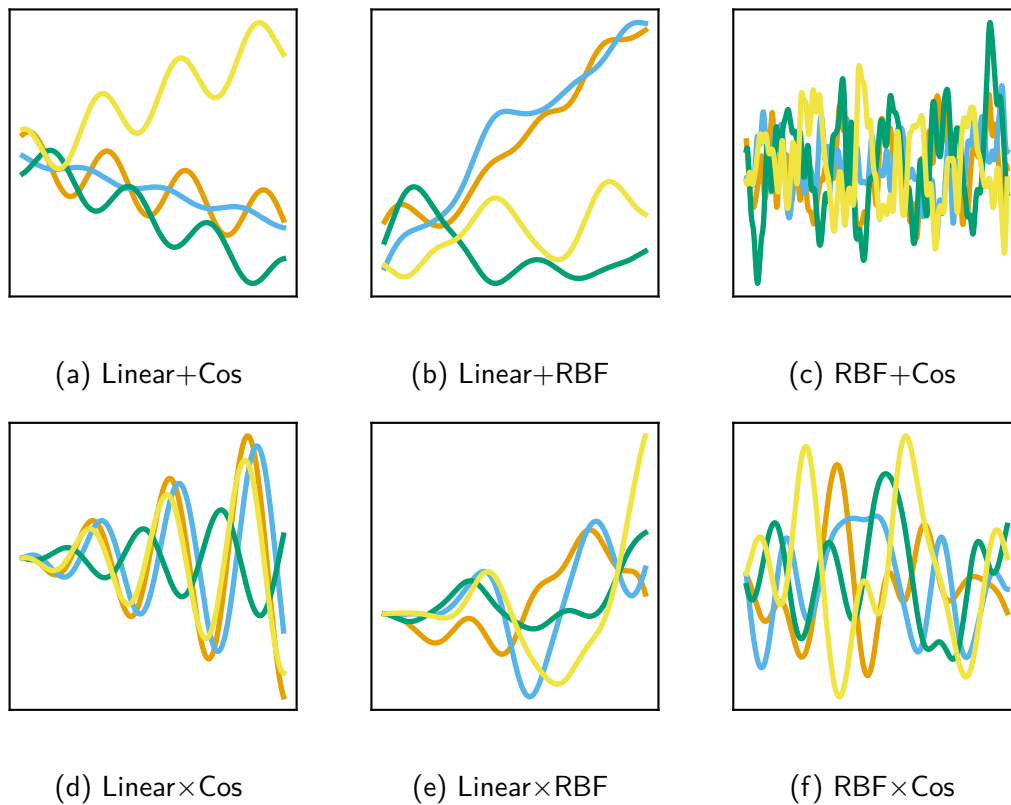
Figure 6.2: Samples resulting from adding a linear and a periodic kernel exhibit a linear trend with periodicity. Adding a periodic to a RBF kernel adds periodicity to smooth functions, whereas adding a linear to a RBF kernel leads to smooth functions with a linear trend. Linear × Periodic leads to periodicity with increasing amplitude, whereas RBF× periodic generates samples that are locally periodic. Linear × RBF leads to smooth functions with increasing amplitude.

6.2   Experiment 1: Distinguishing Compositions

The first experiment assessed whether participants could successfully identify a previously encountered composition from its single-kernel components and the alternative composition. Detecting the same compositional structure in a new pattern is a minimal requirement for assessing people's ability to generalize compositional functions.

6.2.1 Participants

We recruited a total of 50 participants ($M_{\text{age}} = 31.0$, $SD_{\text{age}} = 6.84$, 16 female, 34 male) through Amazon's Mechanical Turk web service. Participants had to have more than 50 approved tasks with an approval rate of 95% or higher. The experiment took about 5 minutes and participants received $0.30 for their participation.

6.2.2 Materials

We generated a set of 100 realizations from each of the base kernels and the compositions. Samples were evaluated over $\mathbf{x} = [0, 0.1, \ldots, 10]$. On each trial, two of the three kernels were sampled without replacement, and one composition rule (+ or ×) was chosen at random. In the test set, participants again saw one function sampled from each of the two kernels presented in the test set and then had to choose the most likely composition of the two kernels from a set of options. The four proposed options were samples from the true function, the alternative, and the two base kernels that generated the functions; see Figure 6.1. Note that while the kernels that generated the patterns were carried over from training to test set, each realization was a unique pattern.

6.2.3 Procedure

Participants were told that they had to reason about sales patterns of different fictitious alien plants on the intergalactic market. First, they were shown patterns from two different plants and their offspring. They were told that the x-axis marked the days over which a plant was traded and that the y-axis showed how well it sold on a particular day. Finally, sales patterns for two new plants were shown, and participants had to choose the sales pattern of their potential offspring. To familiarize participants with the task, participants first completed an

example trial and had to answer four comprehension checks successfully. They then saw the ten trials and were told that they should treat every trial as a new set of sales patterns.

6.2.4   Results

We modeled the task by fitting two hierarchical Bayesian models – one in which we aggregated across stimuli and focused on each individual participants' choice counts (correct, other, or single), and one in which we aggregated across participants and focused on per-condition counts. Both models had the same structure: the proportion of choices for each of the options was modeled as a Binomial distribution $c \sim Binomial(p)$, where the proportion of successes was $p \sim logit^{-1}(\theta_i)$, and each participants' or conditions' $\theta_i \sim \mathcal{N}(\mu, \sigma)$, mean and variance were pooled across groups, $\mu \sim \mathcal{N}(logit(p_0), 2)$. We centered the hyperprior μ on chance-level proportions ($p_0 = 1/4$ for correct and other proportions, $p_0 = 1/2$ for single-choice proportions) and specified a broad variance ($\sigma = 2$, corresponds to 95% of the prior distribution being between 2% and 98% for $\mu = 1/4$).

Participants selected the sales pattern generated by the true composition (204 out of 500, 41%) more often than the pattern generated by the alternative composition (126 out of 500, 25%. Estimates for the true pattern ($\hat{p} = 0.41$, $HPD_{95} = [0.36, 0.45]$) exhibited credibly higher proportions than the alternative ($\hat{p} = 0.25$, $HPD_{95} = [0.21, 0.29]$)[2]).

Participants chose the single component at higher rates than the alternative, but lower than the correct composition (170 out of 500, 34%). The estimated proportions were lower than the true proportion, but exhibited some overlap in their HPDs ($\hat{p} = 0.33$, $HPD_{95} = [0.27, 0.39]$).

We also analyzed how many of the participants chose the correct composition

---

[2]Two-sided Fishers' exact tests were generally consistent with our Bayesian analysis across this chapter.

(a) Posterior $\hat{p}$ for each option.

(b) Posterior $\hat{p}$ for the true function.
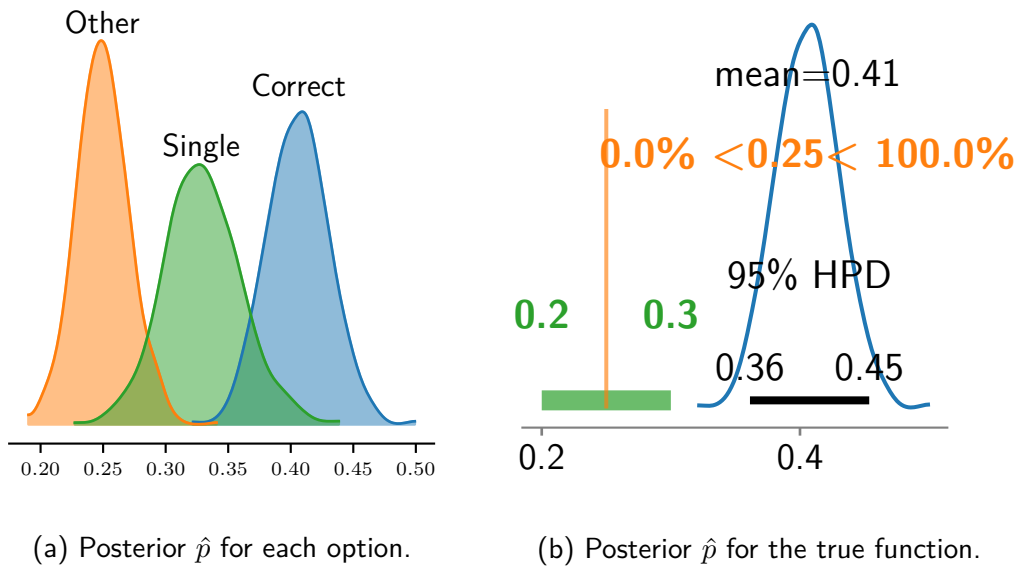
Figure 6.3: In Experiment 1 participants selected the true pattern over the alternative rule and the constituent patterns (Figure a). Figure (b) displays the reference value (1/4) and the surrounding ROPE ([0.2, 0.3]), as well as $\hat{p}$ ($M = 0.41$, $\text{HPD}_{95\%} = [0.36, 0.45]$). Since 100% of the HPD are larger than the ROPE (orange text), we can state that the proportion is credibly larger than $1/4$.

more frequently than expected at chance level. We contrasted HPD_{95} of our obtained estimates against chance level (1/4), including a 5% buffer [0.2, 0.3] to form a region of practical equivalence (ROPE). We can then accept or reject the hypothesis that our obtained proportions are different from chance. If the ROPE is entirely contained in the HPD_{95}, we accept the H_0, i.e., we find evidence for participants' choosing at chance-level. If the ROPE and the HPD_{95} form disjoint sets, we can reject the H_0. Finally, any other overlap is treated as undecided, with the percentage of overlap indicating how dissimilar the estimates were. For a recent introduction into hypothesis testing via ROPE, see Kruschke (2018).

All participants exhibited proportion estimates greater than 1/4. Furthermore, 38% of the participants' estimates were credibly higher than chance (19 out of 50 had no overlap between the ROPE and the HPD_{95}) and the remaining 31 had only negligible overlap with the ROPE ($M_{overlap} = 0.02\%$, $SD_{overlap} =$

$0.3, max = 0.12)$.
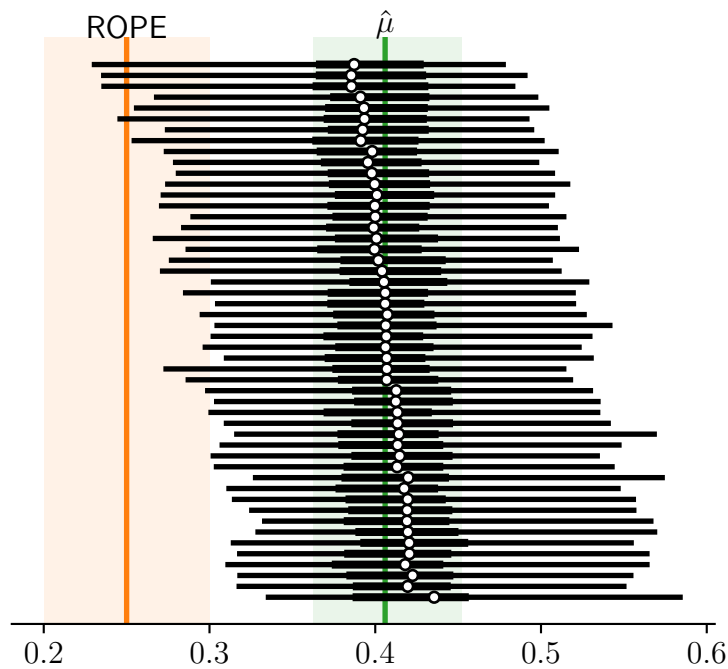
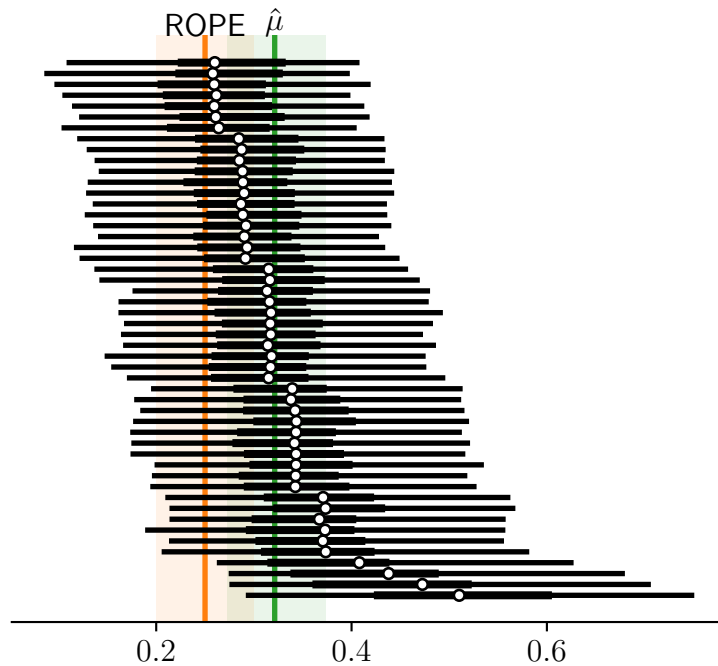

Figure 6.4: Per-participant median estimates (solid circles) in Experiment 1, 5% and 95% HPD intervals. In green, the overall participant estimate μ and its 95% HPD. In orange, the ROPE around chance level $1/4$.

Finally, we analyzed participants' performance, given the true underlying rule. Participants performed well if the true underlying rule was an additive (47 out of 84, 56%) or multiplicative composition of a linear and a periodic kernel (35 out of 80, 44%), and their HPD$_{95}$ was credibly above the ROPE, $\hat{p_+} = 0.53$, HPD$_{95+} = [0.43, 0.63]$; $\hat{p_\times} = 0.43$, HPD$_{95\times} = [0.33, 0.52]$ [3].

Similarly, participants selected the correct pattern at high proportions for additive (35 out of 71, 49%) and multiplicative (39 out of 91, 43%) combinations of linear and RBF at rates higher than chance. Again, HPD$_{95}$ were credibly above the ROPE, $\hat{p+} = 0.48$, HPD$_{95+} = [0.38, 0.59]$; $\hat{p\times} = 0.42$, HPD$_{95\times} = [0.33, 0.51]$. Performance was not better than chance for additive compositions of RBF and

---

[3] Exact binomial tests against $1/4$ were consistent with the Bayesian analysis throughout this chapter.

periodic kernels (20 out of 83, 24%, $\hat{p} = 0.27$, $\text{HPD}_{95} = [0.18, 0.37]$). Neither were multiplicative combinations selected credibly higher than chance (28 out of 91, 31%, $\hat{p} = 0.33$, $\text{HPD}_{95} = [0.22, 0.41]$). For the observed proportions, see Figure 6.5, for estimated proportions $\hat{p}$; see Figure 6.6.
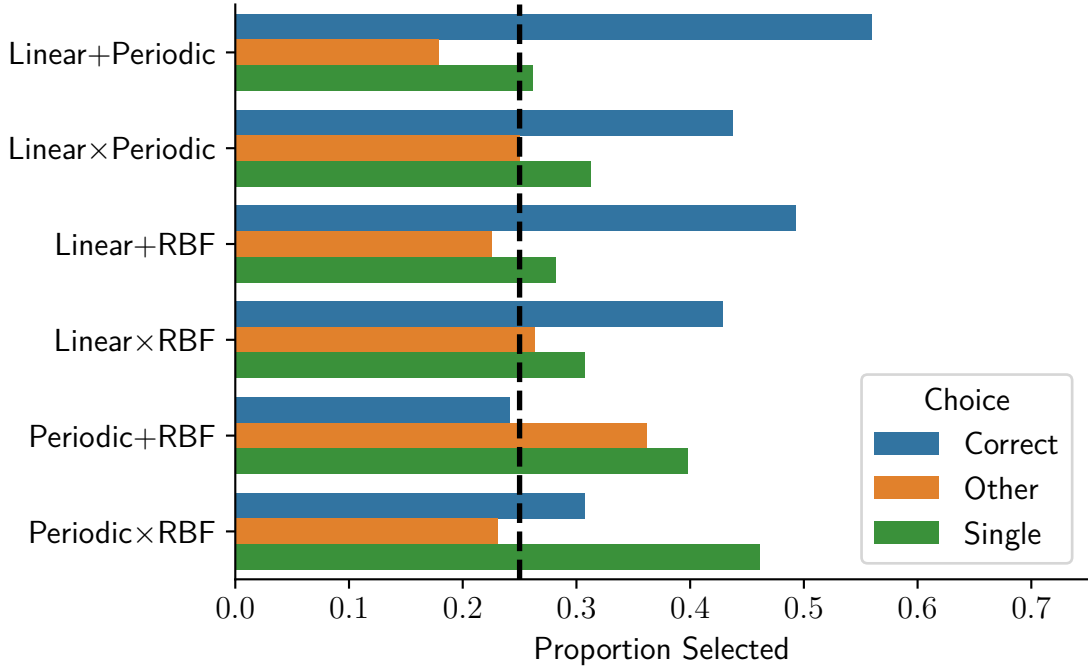
Figure 6.5: Choice proportions for each of the six underlying rules and chance level (dashed line) in Experiment 1. Note that the single proportions contain both constituent options.

Overall, Experiment 1 showed that participants were able to detect the rules in the test set and used that knowledge to generalize to novel instances. However, for combinations of RBF and periodic kernels, participants were less inclined to infer the true composition, potentially because the composition of these patterns was more challenging to distinguish from the alternative composition or the individual constituents.

While participants successfully generalized the rules from only one example, both training and test set corresponded to the same combination of kernels. Thus, this one-shot generalization could simply amount to generalizing the pattern with-
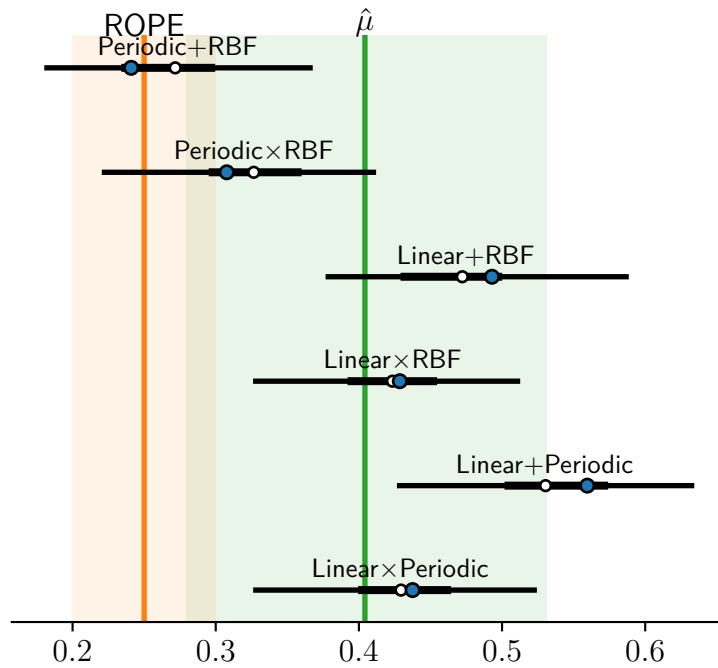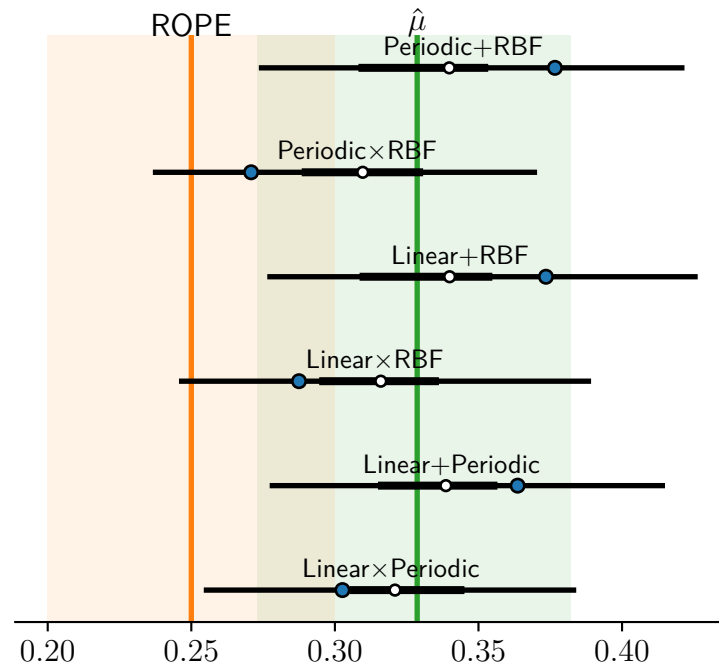
Figure 6.6: Per-rule median estimates (solid white circles) in Experiment 1, 5% and 95% HPD intervals. In green the overall participant estimate μ and its 95% HPD. In orange the ROPE around chance level $(1/4)$. The solid blue circles correspond to the observed proportions.

out detecting the compositional rule. A stronger form of one-shot generalization would entail generalizing the compositional rule to new combinations of kernels. This would amount to perceiving the compositional operation underlying the example rule and participants being able to apply it to a new pattern pair. We assessed this ability in Experiment 2.

6.3   Experiment 2: Generalizing a Composition

In Experiment 2, we tested whether participants generalize the compositional rule from the test set to a composition involving a new component. This would amount to inferring the compositional rule underlying the example and applying it to a new set of patterns, thus providing strong evidence for compositional

mechanisms in function learning.

6.3.1 Participants

We recruited 50 participants ($M_{\text{age}} = 30.5$, $SD = 7.04$, 19 female, 31 male) through Amazon's Mechanical Turk web service. Participants had to have more than 50 approved tasks with an approval rate of 95% or higher. The experiment took about 6 minutes on average, and participants received $0.30 for their participation.

6.3.2 Design and Procedure

The design and procedure were the same as in Experiment 1 with one difference: in the test set, one of the base kernels was replaced with the remaining kernel. For example, if the training set consisted of the RBF and linear kernel, the test set consisted of the RBF and periodic, or linear and periodic kernel.

6.3.3 Results

In contrast to Experiment 1, proportions for patterns corresponding to single kernels were higher than the true pattern (177 out of 500, 35%, $\hat{p} = 0.35$, $\text{HPD}_{95} = [0.29, 0.4]$), see Figure 6.7.

More importantly, participants did not select the sales pattern generated by the true composition (164 out of 500, 33%) credibly more often than the alternative (159 out of 500, 32%). Similarly, estimated proportions for the true pattern ($\hat{p} = 0.32$, $\text{HPD}_{95} = [0.27, 0.38]$) did strongly overlap with the alternative ($\hat{p} = 0.32$, $\text{HPD}_{95} = [0.27, 0.36]$). There was therefore no evidence for a difference between the two compositions.

While all participants exhibited proportion estimates larger than chance, none of those estimates were credibly different from the ROPE (all 50 participant

(a) Posterior $\hat{p}$ for each option.

(b) Posterior $\hat{p}$ for the true function.
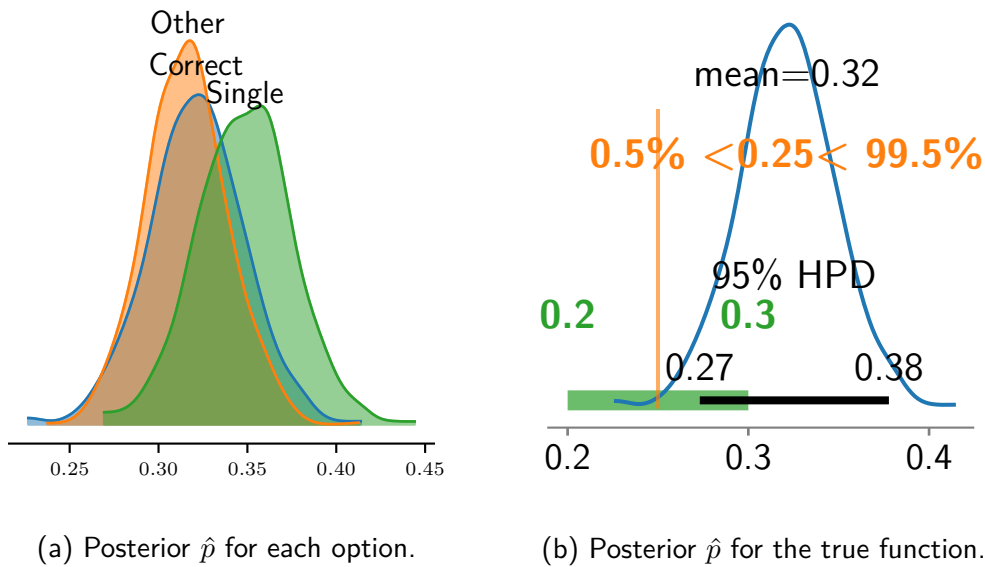
Figure 6.7: In contrast to Experiment 1, participants did not select the true pattern over the alternative rule or the constituent patterns (Figure a). The estimated proportion $\hat{p}$ overlapped significantly with the the random-chance ROPE (Figure b).

proportion estimates overlapped with the ROPE, $M_{overlap} = 32\%$, $SD_{overlap} = 13\%$. For the observed proportions, see Figure 6.8; for estimated proportions $\hat{p}$, see Figure 6.9.

Participants selected the correct choice slightly higher than chance rates for additive compositions, see Figure 6.10. However, all HPD_{95} intervals overlapped with the ROPE, therefore there was no credible difference between the ROPE and the estimates. Multiplicative compositions were selected less frequently, and again all HPD_{95} intervals overlapped with the ROPE.

Interim Discussion

Experiment 1 showed that participants generally preferred compositional rules and could distinguish the true composition from the alternative. However, in Experiment 2, we did not find that participants could generalize these compositions to novel combinations.
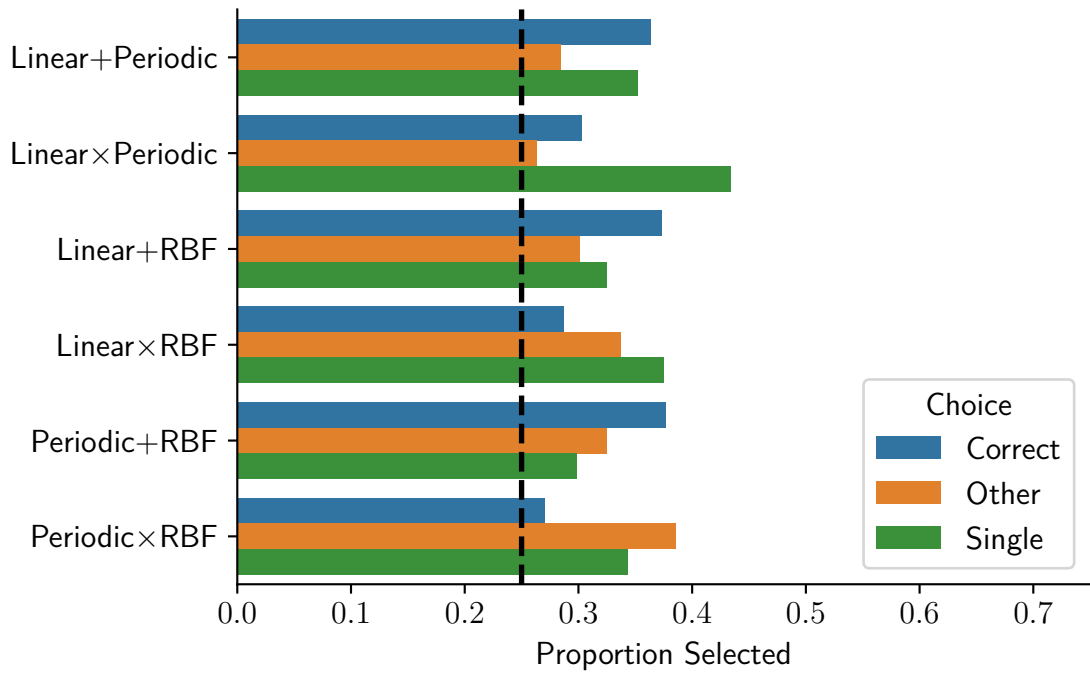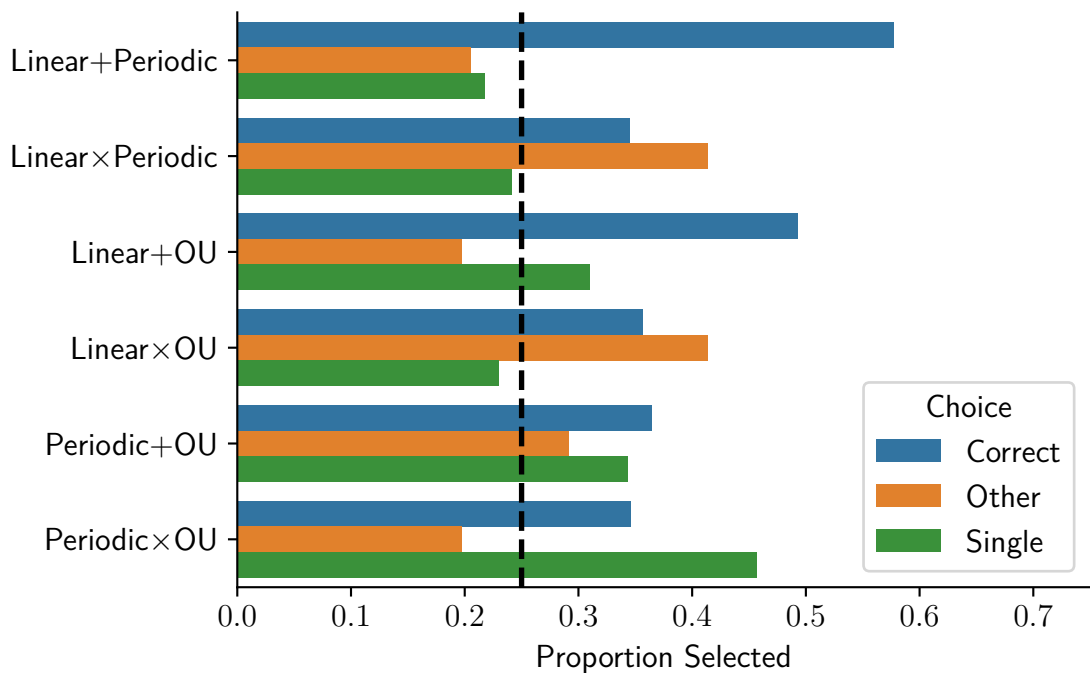
Figure 6.8: Choice proportions for each of the six underlying rules and chance level (dashed line) in Experiment 2.

However, the kernels we selected might have been sub-optimal for the experimental design, as their compositions can result in patterns visually indistinguishable from the alternative or the constituents. This explanation is especially suggestive for multiplications of RBF and periodic since RBF can exhibit semi-periodic patterns. Thus, participants might not have been able to detect any composition at all for conditions where periodic and RBF were the example rule (RBF×Linear and Periodic×Linear). Also, the choice pattern RBF×Linear might not have been salient enough for participants to distinguish it from its alternative RBF×Linear. Thus even if participants inferred a multiplicative pattern for conditions with Periodic×Linear or RBF×Linear as example rules, they might have been at chance level selecting between the Periodic×RBF and Periodic+RBF patterns.

In Experiment 3, we tested if a more salient kernel in the context of the training set would allow rule generalization. We replaced the RBF kernel with

Figure 6.9: Per-participant median estimates (solid circles) for Experiment 2, 5% and 95% HPD intervals. In green, the overall participant estimate μ and its 95% HPD interval. In orange, the ROPE around chance level $1/4$.

the Ornstein-Uhlenbeck kernel (OU). Samples from an OU kernel are less smooth than samples from the RBF kernel and, therefore, might be easier to distinguish visually from periodic samples. For a comparison of samples from an OU, RBF, and periodic kernel, see Figure 6.11.

6.4   Experiment 3: Generalizing Distinguishable Compositions

6.4.1   Participants

We recruited 50 participants ($M_{\text{age}} = 32.54$, $SD_{\text{age}} = 9.58$, 20 female, 30 male) through Amazon's Mechanical Turk web service. Participants had to have more than 50 approved tasks with an approval rate of 95% or higher. The experiment

Figure 6.10: Per-rule median estimates in Experiment 2 (solid white circles), 5% and 95% HPD intervals. In green, the overall participant estimate and its 95% HPD. In orange, the ROPE around chance level $(1/4)$. The solid blue circles correspond to the observed proportions.

took 6 minutes on average, and participants were paid $0.30 for their participation.

6.4.2 Design and Procedure

We replaced the RBF kernel with an OU kernel. Otherwise, the design and procedure were the same as in Experiment 2.

6.4.3 Results

As in Experiment 1, participants selected the correct pattern (204 out of 500, 41%, $\hat{p} = .4, HPD_{95} = [0.33, 0.46]$) at higher rates than the single patterns (150 out of 500, 30%, $\hat{p} = .28, HPD_{95} = [0.2, 0.33]$). The correct pattern was selected at

(a) Linear+OU   (b) Linear×OU

(c) Periodic+OU   (d) Periodic×OU
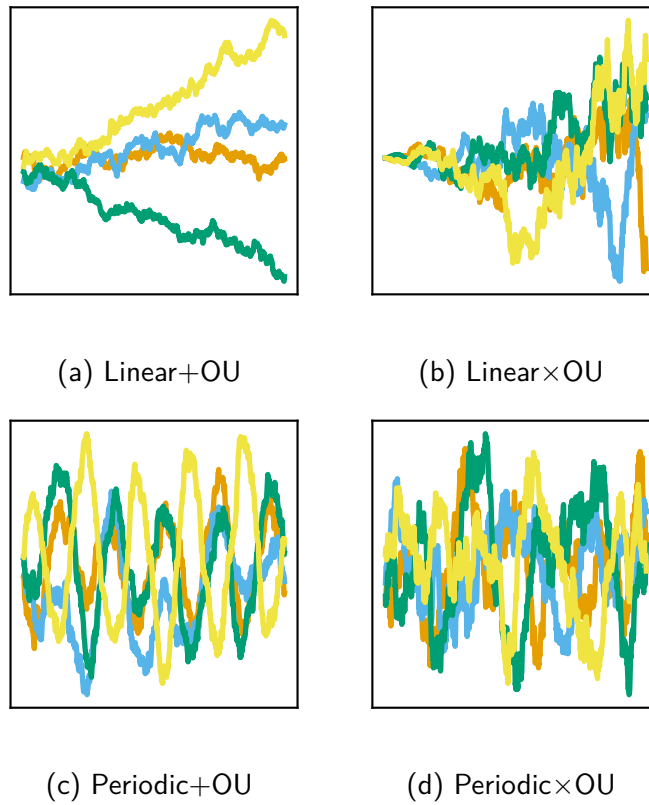
Figure 6.11: Samples resulting from adding linear and OU kernels exhibit a linear trend with additive random-walk patterns. Multiplication of OU and linear generates samples that exhibit random-walks with increasing variance. For compositions of periodic and OU, patterns are difficult to distinguish. Periodic+OU samples correspond to a periodic pattern added to a mean-reverting random walk. Periodic×OU samples correspond to a periodic pattern stretched by a random walk.

credibly higher rates than the alternative (146 out of 500, 29%, $\hat{p} = .28, HPD_{95} = [0.23, 0.33]$), see Figure 6.12.

(a)

(b)
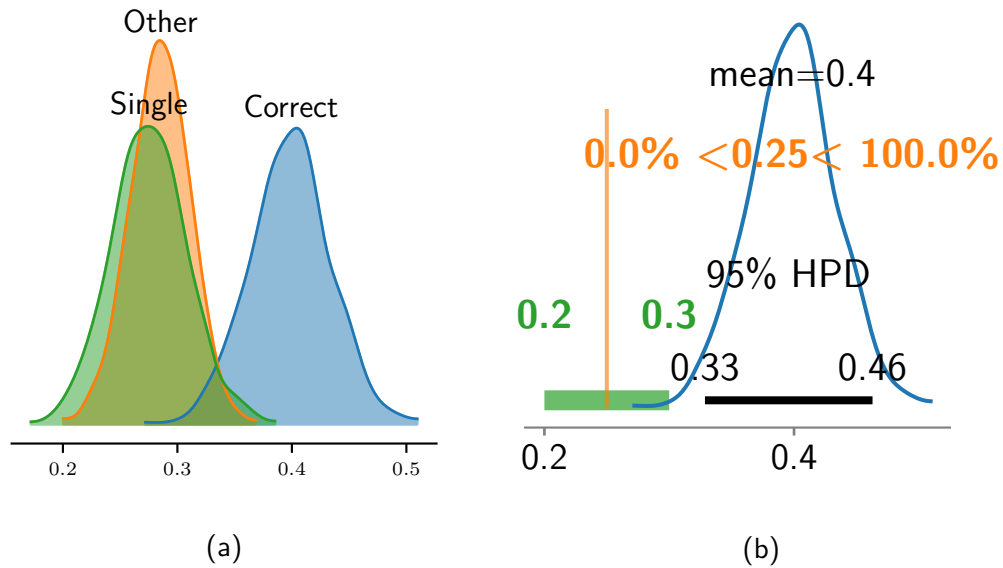
Figure 6.12: In Experiment 3, participants did select the true pattern over the alternative rule or the constituent patterns.

All participants exhibited proportion estimates larger than chance, and nine out of 50 participants had proportion estimates for the correct composition credibly above the ROPE. The other participants were undecided and often had low overlap with the ROPE ($M_{overlap} = 19\%$, $SD_{overlap} = 15\%$). For all participants' estimates, see Figure 6.13.

Additive compositions were chosen at highest rates, with Linear+Periodic chosen at highest proportions (45 out of 78, 58%, $\hat{p} = .52, HPD_{95} = [0.40, 0.63]$). Linear+OU was selected at similar high rates (35 out of 71, 49%, $\hat{p} = .46, HPD_{95} = [0.36, 0.57]$). Finally, the true composition was selected frequently for Cos+OU (35 out of 96, 36%, $\hat{p} = .38, HPD_{95} = [0.29, 0.46]$), however, the alternative was also selected often.

The multiplicative compositions resulted in lower proportions of correct patterns selected, and proportions were also lower than chance-level and the alternative compositions. While Linear×OU was selected relatively frequently (31
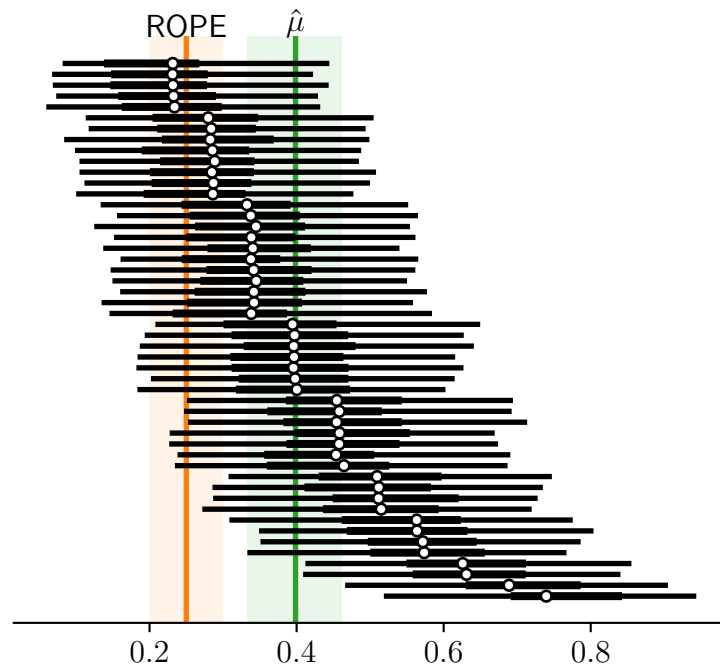
Figure 6.13: Per-participant median estimates (solid circles) for Experiment 3,5% and 95% HPD intervals. In green, the overall participant estimate μ and its 95% HPD interval. In orange, the ROPE around chance level $1/4$.

out of 87, 36%, $\hat{p} = .37, HPD_{95} = [0.28, 0.46]$), the alternative, Linear+OU, was selected more often (36 out of 87, 41%). Similarly, Linear×Periodic was selected above chance (39 out of 87, 34%, $\hat{p} = .36, HPD_{95} = [0.28, 0.45]$), but the alternative was selected at higher rates (36 out of 87, 41%). Cos×OU was not selected credibly above chance (28 out of 81, 35%, $\hat{p} = .36, HPD_{95} = [0.27, 0.46]$). Instead, participants preferred to select the constituents Periodic and OU. For the proportions selected for each composition, see Figure 6.14 for estimated proportions, see Figure 6.15.

Figure 6.14: Choice proportions for each of the six underlying rules and chance level (dashed line) in Experiment 3.

Experiment 3 suggests that one of the reasons participants in Experiment 2 did not generalize the pattern was that the patterns were not visually distinct enough. Substituting the RBF kernel with a more salient OU kernel, we found that participants were able to generalize the compositional rules for additive compositions. However, for multiplicative rules, they generally preferred the additive alternative. Finally, for Periodic×OU, they selected the constituents, potentially
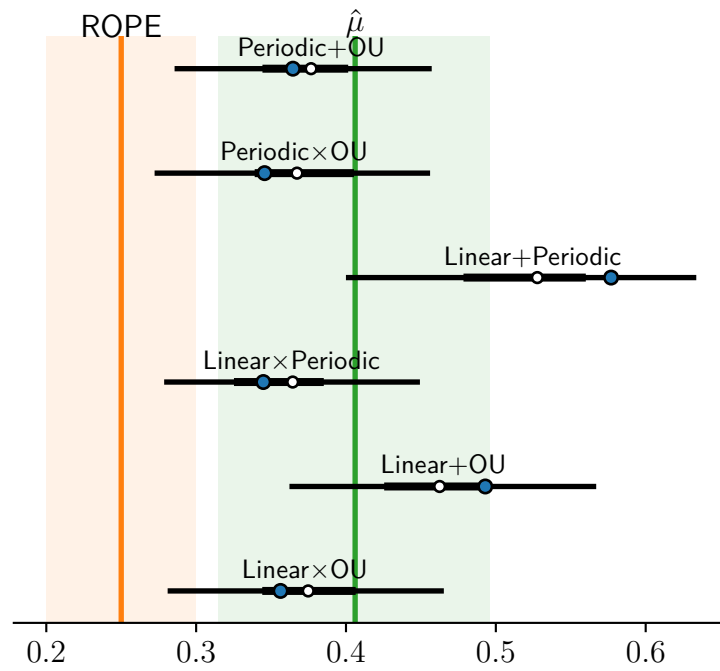
Figure 6.15: Per-rule median estimates in Experiment 3 (solid white circles), 5% and 95% HPD intervals. In green, the overall participant estimate and its 95% HPD. In orange, the ROPE around chance level $(1/4)$. The solid blue circles correspond to the observed proportions.

because Periodic×OU was not distinguishable enough from its constituent patterns.

6.5 Experiment 4: Alternative Explanations

Experiments 1-3 provided some evidence that participants could detect and generalize the compositional structure of functions. However, this ability was limited to additive compositions and depended on the distinctiveness of the kernels.

Given this interdependence, one natural alternative explanation is that participants simply matched the offspring in the training set to the presented samples. For example, surface features in the offspring pattern, such as local variance or monotonicity, could be sufficient to allow participants to guess the right pattern. This behavior would invalidate our hypothesis that participants engaged in rule-learning, since they would not detect the compositional rule, nor apply the composition to a new pair of stimuli. Experiment 4 assessed whether the true composition could be detected solely based on features of the offspring in the training set.

6.5.1 Participants

We recruited 50 participants ($M_{\text{age}} = 32.83$, $SD_{\text{age}} = 10.63$, 20 female, 30 male) through Amazon's Mechanical Turk web service. Participants had to have more than 50 approved tasks with an approval rate of 95% or higher. The experiment took about 5 minutes on average, and participants received \$0.30 for their participation.

6.5.2 Design and Procedure

The design and procedure were the same as in Experiment 3 with one crucial difference: participants did not see any samples of the base kernels. Thus only

the composition in the training set and no samples in the test set were shown. Participants were instructed to imagine the rules by which the sample in the training set was produced and then pick the pattern that they thought was most likely created by applying the same rule.

6.5.3   Results

Participants did not choose the true composition (141 out of 500, 28%, $\hat{p} = 0.28$, $\text{HPD}_{95} = [0.23, 0.32]$) more frequently than the alternative ($\hat{p} = 0.23$, $\text{HPD}_{95} = [0.19, 0.26]$). For the overall estimates, see Figure 6.16. Participants instead chose

(a)   (b)
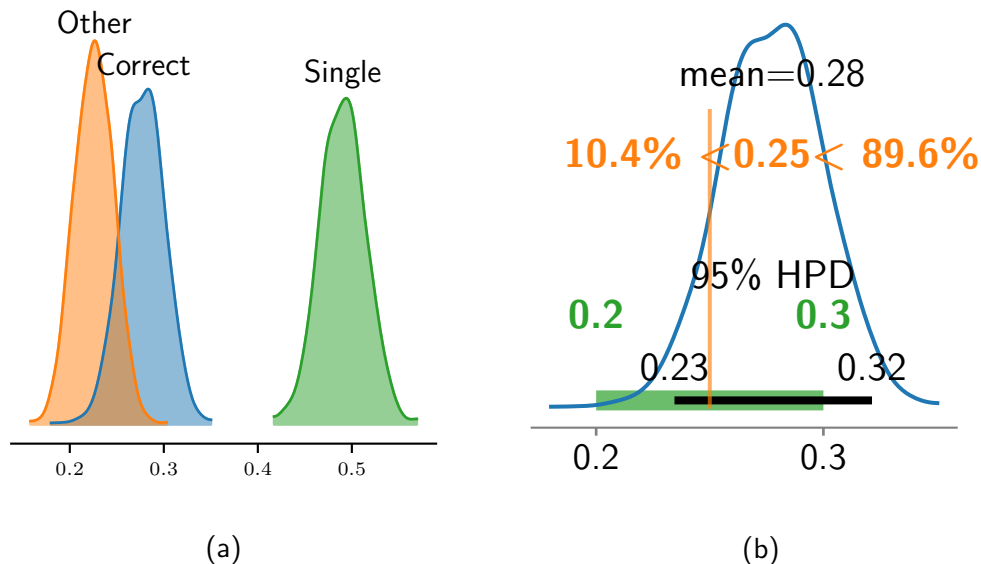
Figure 6.16: Without the presentation of the constituent patterns, participants did not select the correct pattern significantly more than the alternative.

the single components (245 out of 500, 49%, $\hat{p} = 0.49$, $\text{HPD}_{95} = [0.45, 0.54]$).

While all participants produced correct proportion estimates numerically higher than chance, none of the posterior estimates were credibly different from chance (all undecided, $M_{overlap} = 58\%$, $SD_{overlap} = 8\%$), see Figure 6.17.

Only Linear×OU was selected often, but did overlap with the ROPE (30 out of 87, 34%, $\hat{p} = 0.3$, $\text{HPD}_{95} = [0.24, 0.39]$). All other compositions were not
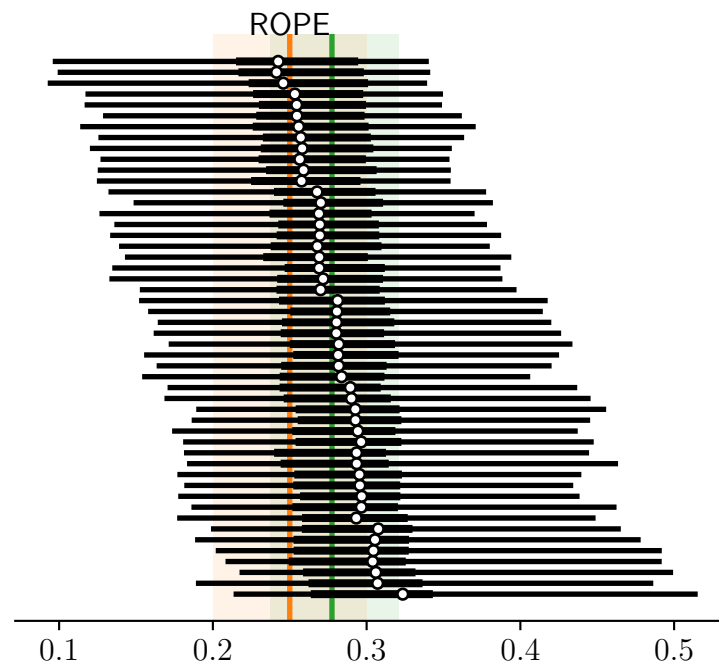
Figure 6.17: Per-participant median estimates (solid circles) for Experiment 4, 5% and 95% HPD intervals. In green, the overall participant estimates and its 95% HPD interval. In orange, the ROPE around chance level $1/4$.

selected at rates credibly above chance.
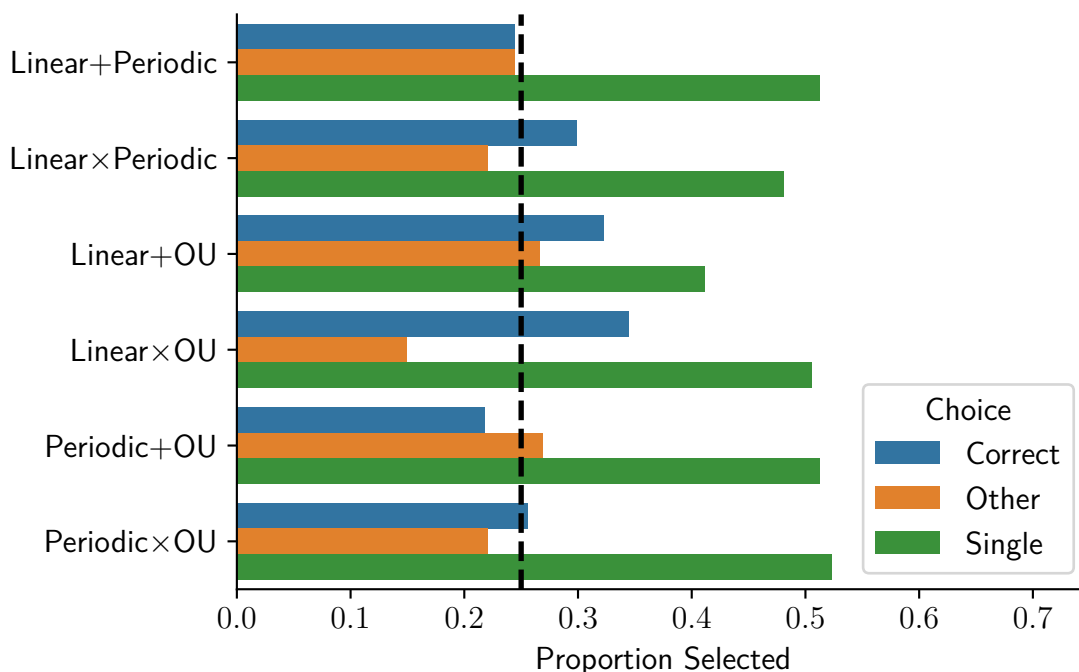
Figure 6.18: Choice proportions for each of the six underlying rules and chance level (dashed line) in Experiment 4.

6.6   Discussion and Conclusion

We explored the human ability to discover and generalize compositional rules in the domain of function learning. Experiment 1 showed that people could distinguish compositions from simpler generating functions, as well as alternative compositions. The second experiment assessed generalization to a new composition. In this more complicated version, participants did not succeed at reliably identifying the correct rule. However, when we replaced the RBF kernel with the more distinctive OU kernel, we found that participants could infer the true compositional rule for additive compositions.

Moreover, when the training set did not contain the individual components, participants were unable to identify the true composition, suggesting that the
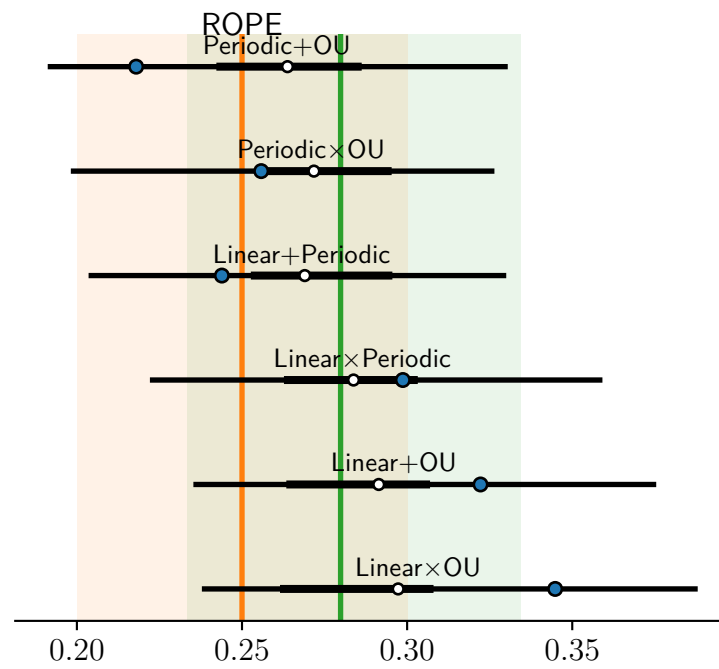
Figure 6.19: Per-rule median estimates in Experiment 4 (solid white circles), 5% and 95% HPD intervals. In green, the overall participant estimate and its 95% HPD. In orange, the ROPE around chance level $(1/4)$. The solid blue circles correspond to the observed proportions.

generalization in Experiments 1 and 2 rested on understanding how patterns combined and not merely on surface-level features. We, therefore, conclude that people can recognize compositional structure in functions and, for additive and visually sufficiently distinct patterns, perform one-shot generalization.

The inability of participants to generalize multiplicative patterns poses an interesting question – are humans biased against inferring these compositions, or are the particular compositions we picked not identifiable? Research in rule-learning and causal reasoning has shown that human learners heavily favor particular compositions. In rule-learning, research starting from the foundational work of Bruner et al. (1956) and Shepard et al. (1961) has shown that some logical forms are easier to learn. Adults expect conjunctive rules and learn those faster than disjunctive rules (Alfonso-Reese et al., 2002; Bourne, 1970; Salatas and Bourne, 1974). In contrast, in causal reasoning, adults are biased towards causes that independently bring about their effects. They expect disjunctive causes and learn these relationships more quickly (Lucas and Griffiths, 2010). Interestingly, this seems to be a bias acquired through development, as children are more flexible than adults and more rapidly infer conjunctive rules (Lucas et al., 2014; Gopnik et al., 2017).

Our results suggest that function learning mimics the biases observed in the rule-learning literature, as additive compositions correspond to conjunctive statements. This preference could be the result of tracking real-world statistics (Griffiths and Tenenbaum, 2006). If real-world sales patterns are more likely to be composed additively, a bias towards additivity might be adaptive. Indeed, when Quiroga et al. (2018) assessed participants' intuitive priors over structures, most of the likely compositions were combined additively and contained a linear component. However, research in rule-learning and causal learning has highlighted the critical role of contextual information and stimulus features. Research in causal learning has shown that participants form expectations about how causes

interact to bring about an effect. For example, Waldmann (2007) showed that participants expected combinations of liquids to affect heart-rate additively if the effect was the result of the liquids' taste. If heart rate instead depended on the liquids' strength, participants assumed that the average of the two liquids was the form of the relationship. These results suggest that participants disfavored multiplicative combinations in the context of our cover story because they expected biological traits to interact additively.

Finally, complex interactions between constituent stimuli and their compositions might have skewed our results. In rule-learning, characteristics of the stimuli have been suggested to bias participants towards conjunctive or disjunctive forms, with integrable stimuli producing conjunctive biases and separable attributes resulting in disjunctive inferences (Bourne Jr, 1979; Reznick et al., 1978, for an alternative explanation, see Ketchum and Bourne Jr (1980)). Thus, the visual characteristics of the kernels may result in different expectations of composition and decomposition. For instance, while linear or periodic samples might be treated as atomic constituents and combine easily, RBF or OU samples might be treated as more complex, potentially decomposable structures. Future research should investigate these issues in detail to uncover the set of characteristically human kernels underlying human function generalization.

In the next chapter, we will continue exploring what sorts of compositional structures can be learned and how people decide between extrapolating based on a complex, compositional function or simpler alternatives.

Chapter 7

Transferring Function Compositions

In the previous chapter, we saw that people could detect compositions and generalize additive combinations to new pairs of functions. However, instead of extrapolating, participants only had to choose between (possibly unconvincing) options, and our experimental design always implied that features were combined, which suggested compositions. Thus it is unclear whether people can use the inferred compositional structure to extrapolate. Here we expand our analysis of Chapter 6, first by asking participants to extrapolate, and second by evaluating what kind of expectations people form when repeatedly faced with complex compositional functions.

Repeated exposure to a pattern changes one's expectations about future patterns. If the pattern is perceived as governed by a simple function, for example, a linear function, future expectations will be biased to infer linear relationships. If instead, the pattern is a composition of functions, future inferences will be biased toward that composition. However, the representation of the composition and its encoding in memory will affect what kind of biases are produced. One possibility is that the whole composition is chunked and remembered as a non-decomposable unit. For example, many time series exhibit cyclical patterns, such as seasonal changes and daily day-night cycles. Frequently observing these

patterns co-occurring, one forms an expectation that phenomena in the same domain share the same structure. Learning this composition of two cyclical patterns as a chunk would only allow the inference of that particular composition, and one could not infer the yearly structure without inferring the daily cycle. Alternatively, both the constituents and the compositional operation could be encoded in a language-like representation. Analogous to Goodman et al. (2008), hypothesized functions could be generated and learned as compositions of simpler constituent functions. The hypothesis space that people consider when learning functions is produced by a probabilistic language of thought, and hypotheses are compositional expressions generated from a grammar over hypotheses (for an overview, see Piantadosi and Jacobs, 2016). After seeing the doubly-periodic patterns, one would form an expectation of periodic patterns and additive compositions occurring. Given a new situation, one would not necessarily infer the additive composition of the two periodic patterns, but favor additive compositions, and the two periodic relationships.

To see how both hypotheses predict different generalization behavior, consider one last example. After learning about the doubly-periodic pattern and frequently observing it in time series, a new pattern is encountered. The pattern exhibits a clear seasonal pattern but no daily cycle. If exposure to the doubly-cyclic pattern resulted in the storage of a non-decomposable chunk, one would expect the new pattern to be doubly periodic. If the chunk cannot account for the new pattern, default inductive priors would guide extrapolation, and most likely, extrapolation would be linear. These results would be consistent with general simplicity biases in cognition and explanation (Pothos and Chater, 2002; Blanchard et al., 2018) and previous results in function learning by Wilson et al. (2015). In contrast, if one learns *how* the pattern was composed, we would predict that extrapolation is based on an additive composition of the seasonal pattern and another function, perhaps an a priori salient linear function. That is because

learning about the doubly periodic pattern resulted in biases towards additivity (of any two functions), and biases toward both the seasonal and daily cycles.

7.1 Experiments

In this chapter, we explored these predictions. Concretely, we examined how repeated exposure to either a constituent of a compositional pattern or repeated exposure to the compositional pattern reflected in subsequent extrapolations. Our experiments adopted the same general design as in Chapter 5. However, while in previous experiments, participants received training consistent with one particular function in Chapter 3, or variations of the same function in Chapter 5, here the training exhibited repeated compositional structure. Figure 7.1 shows the experimental setup.

We presented participants with two compositional extrapolation patterns (the training set), with samples generated from additive compositions of three kernels (linear, periodic, OU). The training set's compositional structures were either two repetitions of the same composition (A+B, A+B), or had one shared, overlapping component (A+B, A+C). Afterward, they saw a new pattern (the test set) and again had to extrapolate. They were told that all three sets belonged to the same underlying function. The test was consistent with one of the repeated constituents in the training set (for example, for the overlapping condition: A+B, A+C $\rightarrow$ A). We will call the conditions according to the shared structure in training repeated and non-repeated (or overlapping) and further differentiate according to the underlying transfer pattern into linear and periodic training sets[1].

For example, for linear training sets, participants in the repeated training would see two patterns generated by a linear and periodic function ($2\times$Lin+Cos). In the non-repeated training, participants would instead see one pattern gener-

---

[1]We require this further differentiation, as one pattern Lin+Cos repeated had two corresponding transfer sets (linear and periodic).
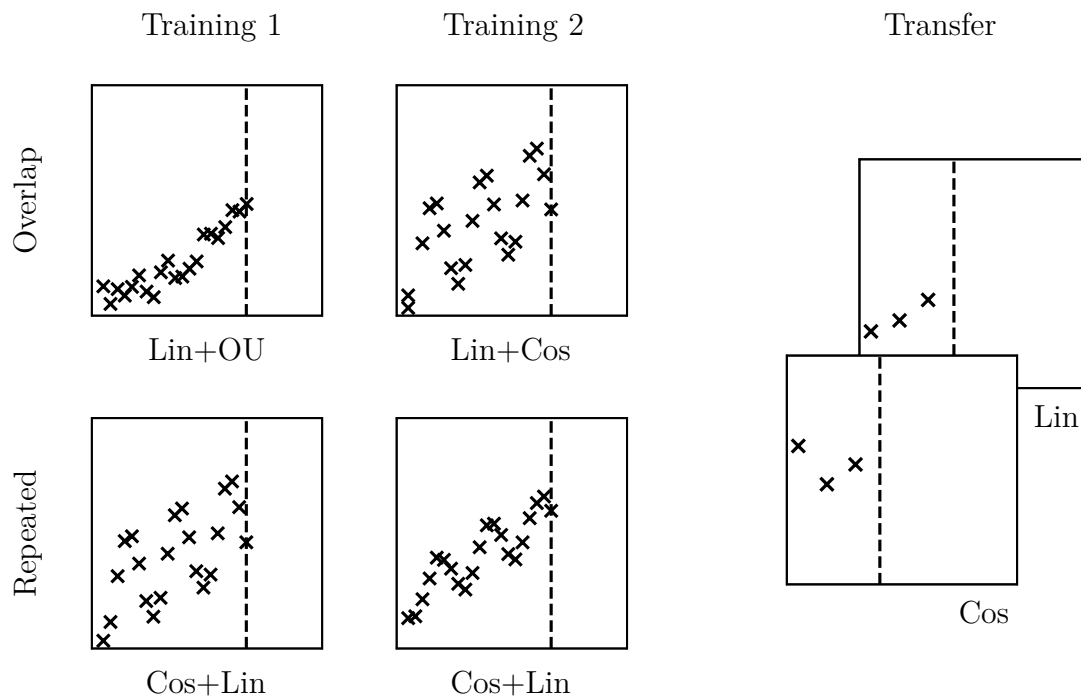
Figure 7.1: Participants received two training sets and subsequently had to extrapolate in the transfer set. The training sets were samples from compositional functions. Across the two training sets, the compositional structure overlapped in one constituent of the composition, or the whole composition was repeated. For example, participants in the overlap condition could have first received data sampled from Lin+OU and then Lin+Cos (with the constituent Lin overlapping). In the repeated condition, participants would be presented with two realizations of the same composition, here Cos+Lin. In the transfer block, participants received a third set of data. The data in the transfer block was not indicative of the compositional pattern, but only reflected one overlapping constituent.

ated from a linear and a periodic function and one pattern generated by a linear and an OU function. The full set of experimental conditions was $2\times$Lin+Cos, Lin+Cos$\rightarrow$Lin+OU, and Lin+Cos$\rightarrow$Lin+OU, for linear transfer, and $2\times$Cos+Lin, $2\times$Cos+OU, and Cos+Lin$\rightarrow$Cos+OU, for periodic transfer sets. Note that the transfer set is denoted by the leading condition name.

We hypothesized that for repeated training sets, participants would form an expectation for the composition. Then, when faced with the transfer set, many participants should prioritize the instruction that all patterns belonged to the same function and extrapolate as in the training set. Thus they would pick a more complex hypothesis over the simpler constituent. In contrast, we expected that participants would select the simpler, repeated function that was also suggested by the test set for overlapping training. To contrast our predictions with participants' extrapolations in the absence of training, we also tested two control conditions in which no training was provided. If participants formed strong biases for compositional or shared constituents, this would provide evidence for the idea that people represent much like a probabilistic language of thought.

7.1.1 Participants

We recruited a total of 402 participants ($M_{\text{age}} = 32.18, SD_{\text{age}} = 10.39$; 180 female, 222 male) on Amazon Mechanical Turk. Participants had to have completed more than 50 approved tasks with an approval rate of 95% or higher. In total, 302 participants completed the experimental conditions and 100 were in the control conditions. Participants in the experimental conditions received \$0.85 for participation and took less then 10 minutes ($M = 8.39, SD = 7.01$) to complete the experiment. The control condition took about 4 minutes to complete ($M = 3.30$, $SD = 3.47$) and participants received \$0.50. Participants were randomly assigned to the experimental and control conditions ($N_{\text{Lin+Cos}\rightarrow\text{Lin+OU}} = N_{2\times\text{Lin+OU}} = 51$, all other conditions $N = 50$).

7.1.2   Procedure

As in the experiments in Chapter 5, participants were instructed that they would
learn the relationship between two substances, substance x and substance y.
They were told that all patterns followed the same regularity and that they had
to predict the relationship for ten new values. As in Chapter 5, they received a
visual aid depicting the training setup.

In the control conditions, participants were only instructed that they would be
presented with a relationship between two substances and that they would have
to first extrapolate for a new pattern. Then they had to choose the most likely
relationship from a set of six patterns. In contrast to our previous experiments,
we adopted a within-participant design. Thus the same participants performed
both extrapolation and forced-choice tasks.

7.1.2.1   Training Phase

In both training blocks, the extrapolation task was presented in the form of a scat-
ter plot. Participants were presented with data and extrapolated the value of the
substance by selecting the height of the y-axis. Participants were then shown the
actual value as feedback for one second. If their choice deviated by ± 0.025 from
the actual value, they had to readjust their selection. During this re-selection,
the actual value was visible. Training blocks were presented in randomized order.

7.1.2.2   Test Phase

After training, participants were reminded that the next pattern followed the
same relationship. Then they were presented with either the linear or periodic
three-point pattern and had to extrapolate for 30 points without feedback.

7.1.2.3 Choice Phase

Once participants had submitted the 30 extrapolation values, they proceeded to the forced-choice task. Participants were instructed that they would be presented with a pattern of three points that belonged to the same relationship as in the training phase. They then had to choose from six patterns, the one they deemed the most likely relationship for the 3-point pattern. The six patterns were scatter plots corresponding to one conditional sample given the three basic kernels and their additive compositions (see Section 7.1.3). Options were presented in randomized order, and the presented function realization was counterbalanced. After the test phase, participants completed a short demographic survey, were debriefed and compensated. For screenshots of the experimental stimuli and instructions, see Appendix D.

7.1.3 Materials

The functions were sampled from three constituent GPs. We chose the same three kernel types and mean functions as in Chapter 5, but this time composed these by adding kernels. These materials allowed us to express more complex generating mechanisms. The three resulting additive kernels were $Lin+Cos$, $Lin+OU$, and $Cos+OU$. For the hyperparameters for those kernels, see Table 7.1.

7.1.3.1 Training Sets

We generated training data by sampling three sets of 30 points in the range 0.05–0.95. We resampled if the data did not fall in the presentation range $[0,1]$, or if the samples were visually not indicative of the GP[2].

The first 20 points were provided as evidence. For the remaining 10 points, participants had to extrapolate the target value. Participants received one realiza-

---

[2]For instance, samples from a OU kernel could produce trends or semi-periodic patterns for the experiments' range of presentation.

Table 7.1: Kernels and kernel parameters (variance σ, lengthscale λ, intercept β_0, and slope β_1) generating the training data. For all models, we set the noise variance $\sigma = 0.01$.

	σ	λ	β_0	β_1
Lin	0.018	–	0.2	0.5
Cos	0.03	0.035	0.5	–
OU	0.05	1	0.5	–

tion of the three sets. The order of block presentations was also counterbalanced. For the three realizations of the training data, see Figure 7.2.

(a) Linear + Periodic   (b) Linear + OU   (c) Periodic + OU
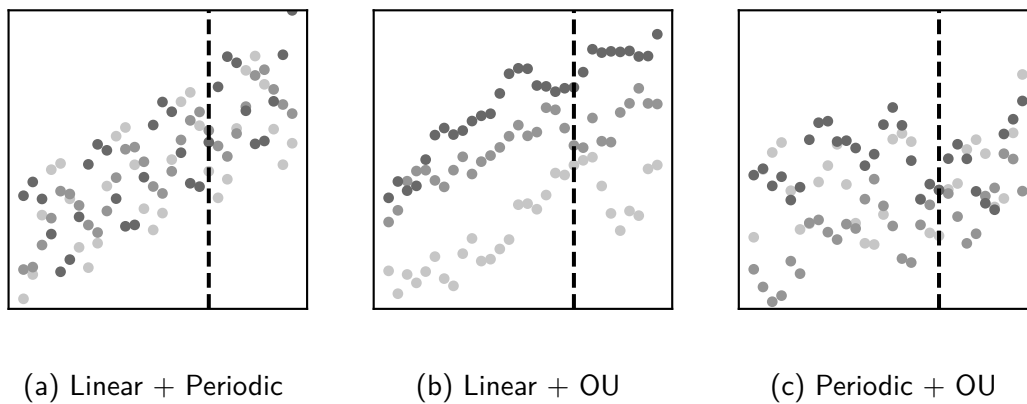
Figure 7.2: Training data used in the experimental conditions (three realizations of each function). For each condition, there were two sets of points to be learned. Participants received the first 20 points and had to extrapolate for the 10 remaining points. The dashed line is the cutoff between presented evidence and training.

7.1.3.2   Transfer Set

The data in the transfer set consisted of three points, $x = \{0.05, 0.17, 0.30\}$, and three sets of corresponding target values in all conditions:

$$y_{\text{Lin}} = [\{0.25, 0.30, 0.39\}, \{0.24, 0.28, 0.35\}, \{0.22, 0.26, 0.30\}],$$

$$y_{\text{Cos}} = [\{0.60, 0.44, 0.52\}, \{0.39, 0.60, 0.40\}, \{0.66, 0.35, 0.63\}].$$

Participants had to extrapolate 20 points in the range 0.67–0.95.

7.1.3.3   Forced-Choice Options

The samples for forced-choice were generated from the three compositional GPs ($Lin + Cos$, $Cos + OU$, $Cos + OU$), as well as the three constituent kernels (Lin, Cos, OU). We conditioned these GPs on the three extrapolation points (see Section 7.1.3.2) and generated three samples for the range 0.67–0.95. Again, we resampled if the samples fell out of the presentation range or were visually not representative of the implied function.

Participants received one of those samples at random for each of the six options. Options matched the participants' conditions: Participants who received the linear transfer set had forced-choice options conditioned on the linear points. Participants who received periodic data in the transfer set had options corresponding to those points. For the samples presented for linear transfer sets, see Figure 7.3; for periodic transfer set options, see Figure 7.4.

7.2   Results

7.2.1   Training Errors

We calculated *MAE*s for extrapolations in the training blocks for submissions before the participants had received feedback. Error was lowest in the linear

(a) Cond. linear: Linear   (b) Cond. linear: Cos   (c) Cond. linear: OU

(d) Cond. linear: Linear+Cos (e) Cond. linear: Linear+OU   (f) Cond. linear: Cos+OU
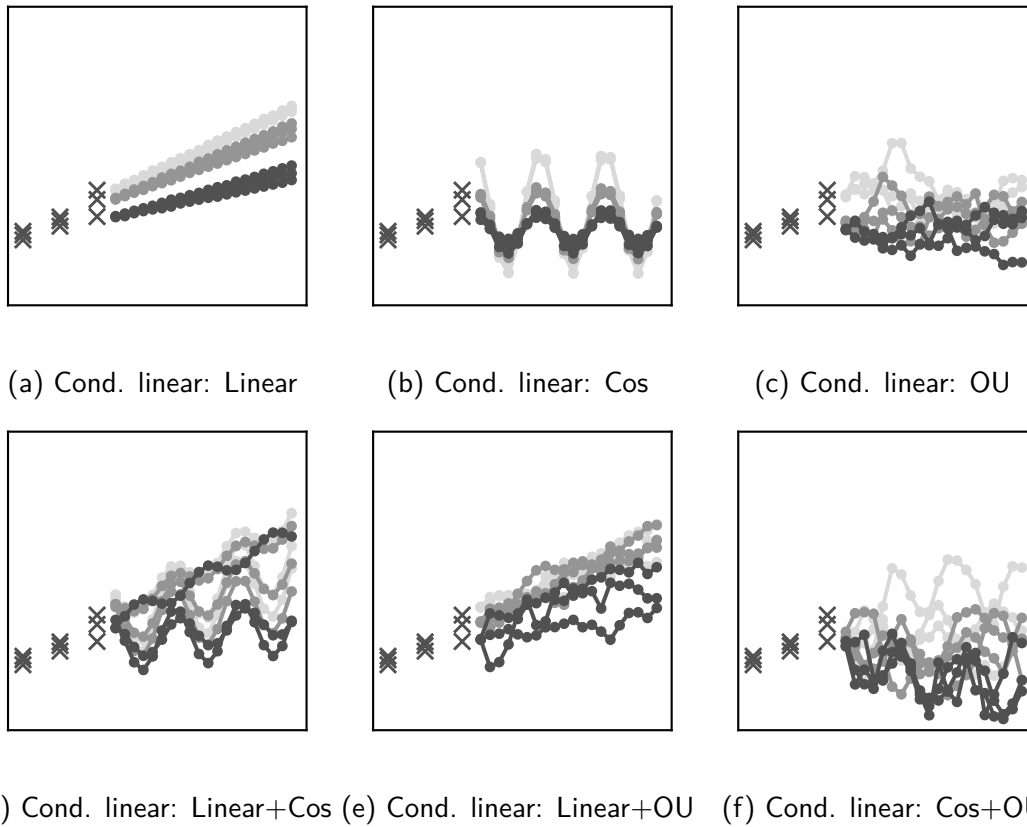
Figure 7.3: The six options presented in the forced-choice block for the linear transfer set conditions. Participants received one sample (from three realizations) of of each option at random.

(a) Cond. Cos: Lin (b) Cond. Cos: Cos (c) Cond. Cos: OU

(d) Cond. Cos: Cos+Lin (e) Cond. Cos: Lin+OU (f) Cond. Cos: Cos+OU
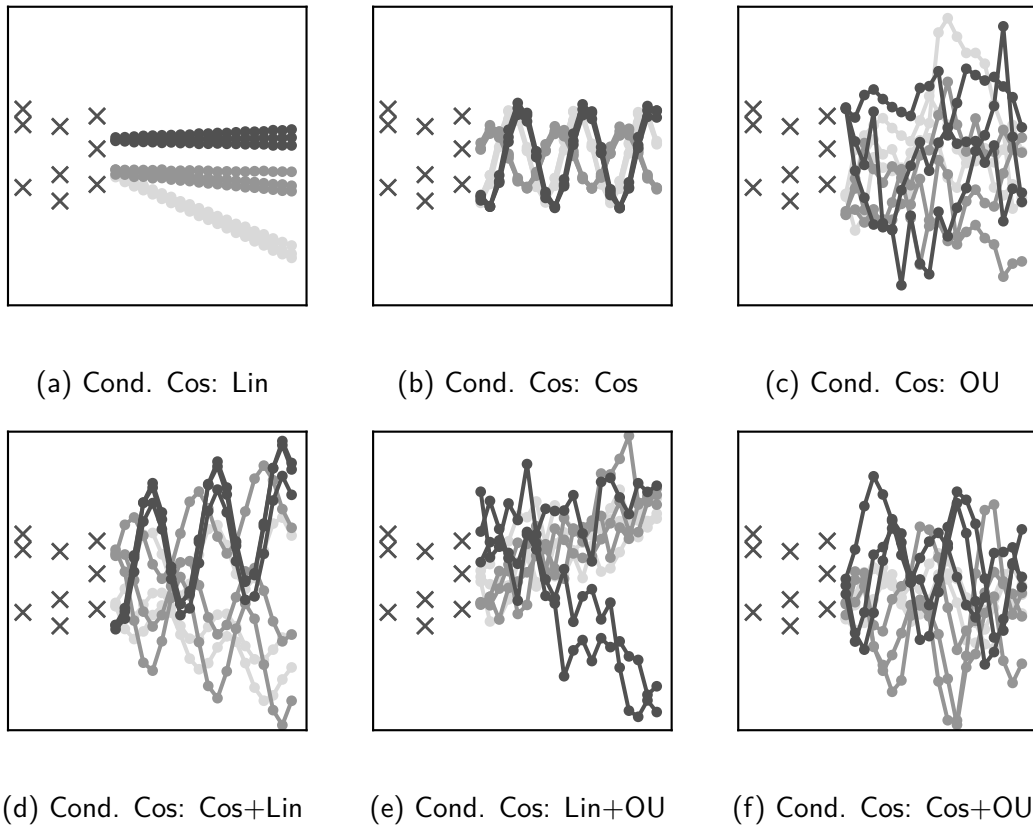
Figure 7.4: The six options presented in the forced-choice block for the periodic transfer set. Participants received one sample (from three realizations) of each option at random.

conditions with 2×Lin+OU exhibiting the lowest errors ($M = 0.08, SD = 0.05$), followed by Lin+Cos→Lin+OU ($M = 0.09$, $SD = 0.03$).

Error for 2×Lin+Cos was higher ($M = 0.11$, $SD = 0.05$). We discuss errors for both 2×Lin+Cos and 2×Cos+Lin aggregated across both function conditions (linear vs periodic), since during the two training blocks both conditions were identical. Both remaining periodic conditions exhibited similarly high errors ($M_{\text{Cos+Lin}\rightarrow\text{Cos+OU}} = 0.11$, $SD_{\text{Cos+Lin}\rightarrow\text{Cos+OU}} = 0.03$, $M_{2\times\text{Cos+OU}} = 0.11$, $SD_{2\times\text{Cos+OU}} = 0.02$).
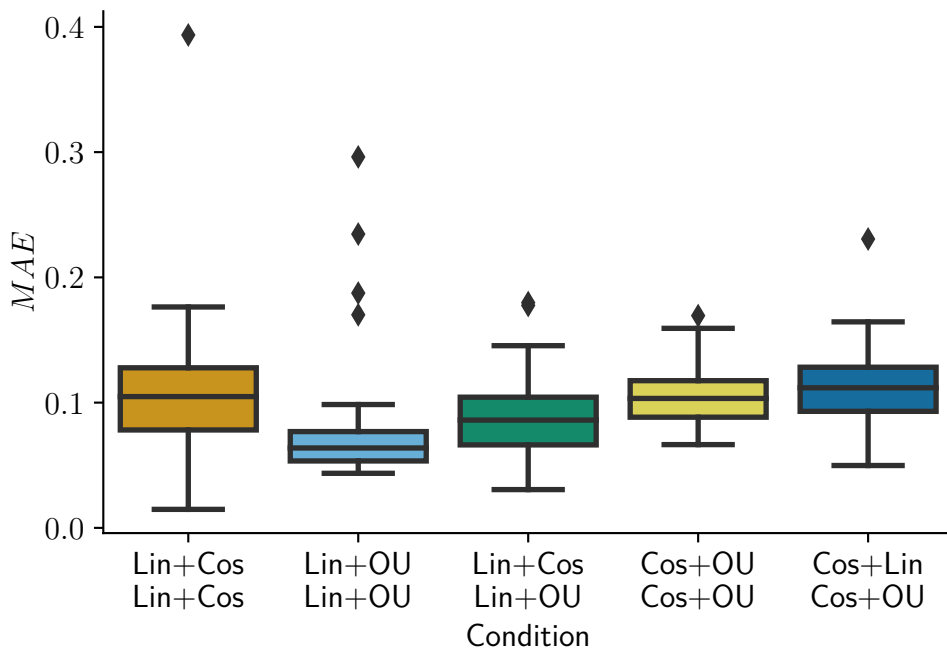
Figure 7.5: Averaged errors across the two training blocks. Errors were similar across conditions and only 2×Lin+OU was somewhat lower. Boxplots display first, second (median) and third quartiles. Whiskers show the 1–5 interquartile range (*IQR*).

7.2.1.1 Error Decay

We did not expect a strong reduction in error across the two blocks since we presented only two training sets, and those blocks contained functions that are known to be difficult to learn (cyclical and noisy). Furthermore, two conditions

did not share the same functional form across training, which should lead to high error and no decrease across training blocks. As in Chapter 5, we evaluated how well several Bayesian models captured the change in *MAE*s across blocks. Consistent with the results in section 5, we found that a hierarchical log-normal, $log(\text{error}) \sim \mathcal{N}(\mu, \sigma)$ model fit the data best. For a discussion of our model results and model comparisons, see Appendix F.2.1.
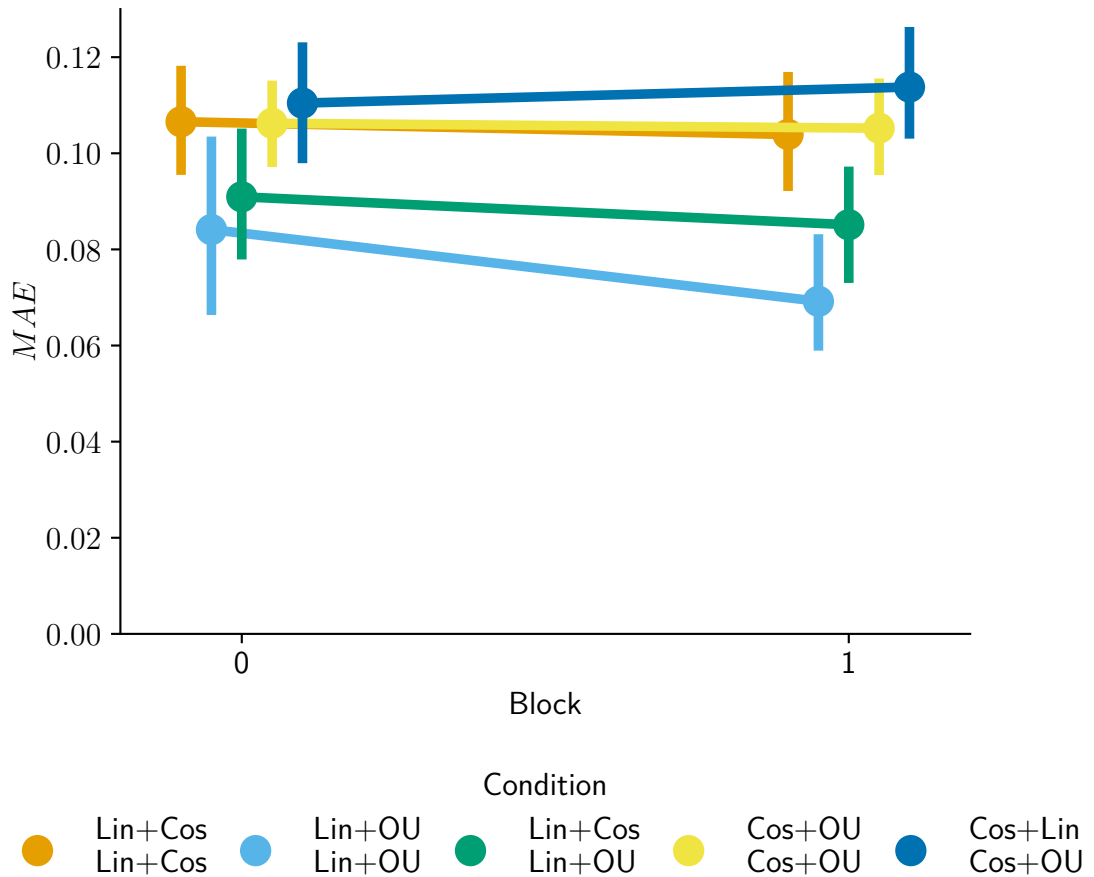
Figure 7.6: Mean absolute error across training blocks and condition. Error bars display 95% bootstrapped confidence intervals.

Estimates for the overall condition difficulty (intercepts) were fairly narrow, around 0.1, reflecting that the conditions were homogeneous in terms of their overall error (HPD_{95} between 0.05–0.13). However, intercept estimates were broad and for all conditions HPD_{95} spanned positive and negative ranges, reflecting the

Table 7.2: Group-level estimated means $\hat{M}$, for intercepts, β_0, and slopes, β_0, as well as 95% highest-posterior density intervals estimated via MCMC for the log-normal model.

	$\hat{M}_{\beta_0}$	$\mathrm{HPD}_{95}\beta_0$	$\hat{M}_{\beta_1}$	$\mathrm{HPD}_{95}\beta_1$
2×Lin+Cos	0.09	[0.08, 0.11]	-0.06	[-0.22, 0.12]
2×Lin+OU	0.07	[0.05, 0.09]	-0.10	[-0.34, 0.12]
Lin+Cos→Lin+OU	0.08	[0.06, 0.10]	-0.04	[-0.28, 0.23]
2×Cos+OU	0.10	[0.08, 0.13]	0.08	[-0.22, 0.38]
Cos+Lin→Cos+OU	0.10	[0.08, 0.13]	-0.02	[-0.26, 0.26]
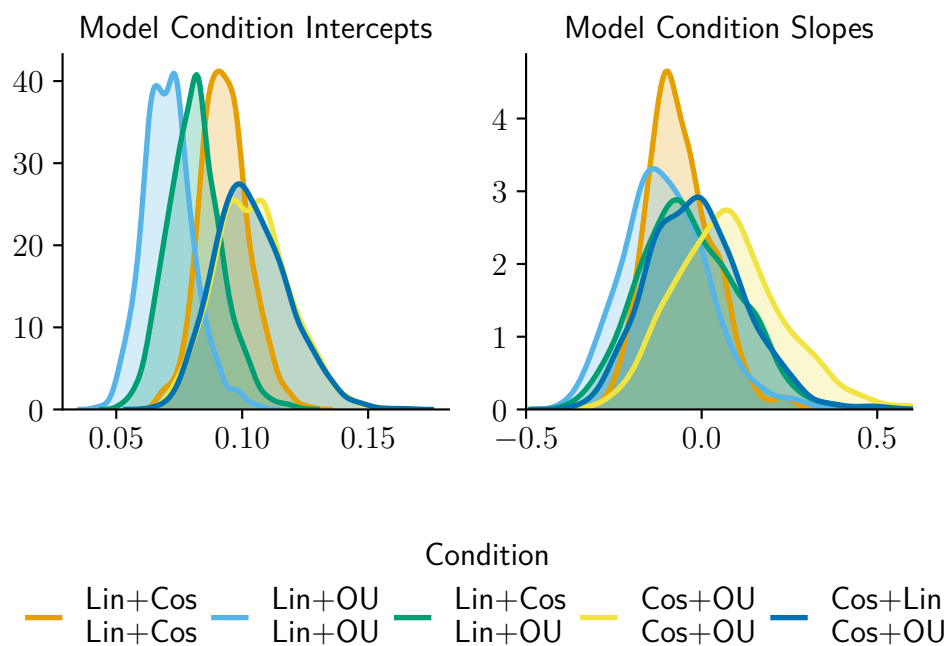
Figure 7.7: Group-level estimates of error intercepts and slopes estimated via the hierarchical Bayesian model.

high inter-participant variability. While some participants managed to reduce their errors across training blocks, even though the same function realization was never repeated, for others, the error stayed constant or even increased. For per-participant errors and model fits, see Figure F.11.

7.2.2 Extrapolations

In both control conditions, participants' extrapolation patterns suggested that the extrapolations were based on the training data presented. In the linear control condition, extrapolations were generally consistent with a positive linear trend, similar in slope to the three points presented. For the periodic control condition, most extrapolations suggested stationary, high-noise or possibly periodic patterns. For extrapolations in the two control conditions, see Figure 7.8.
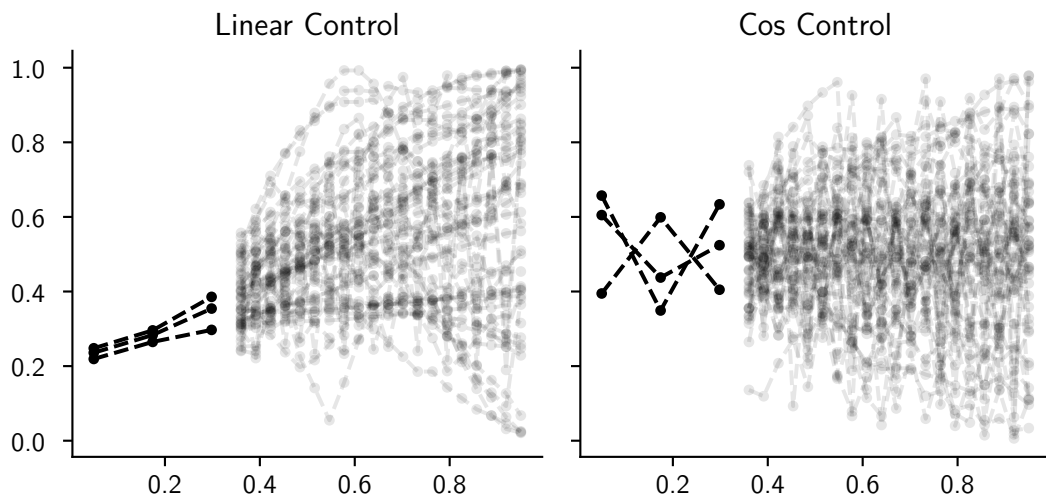
Figure 7.8: Extrapolations in the two control conditions generally resembled the three points presented in the transfer sets.

For experimental conditions, extrapolation patterns were more difficult to characterize. In general, a large proportion of extrapolations in the linear conditions suggested positive trends and often periodic or noisy additive structure. For extrapolations in the two training sets and the transfer set for the Linear

conditions, see Figure 7.9.
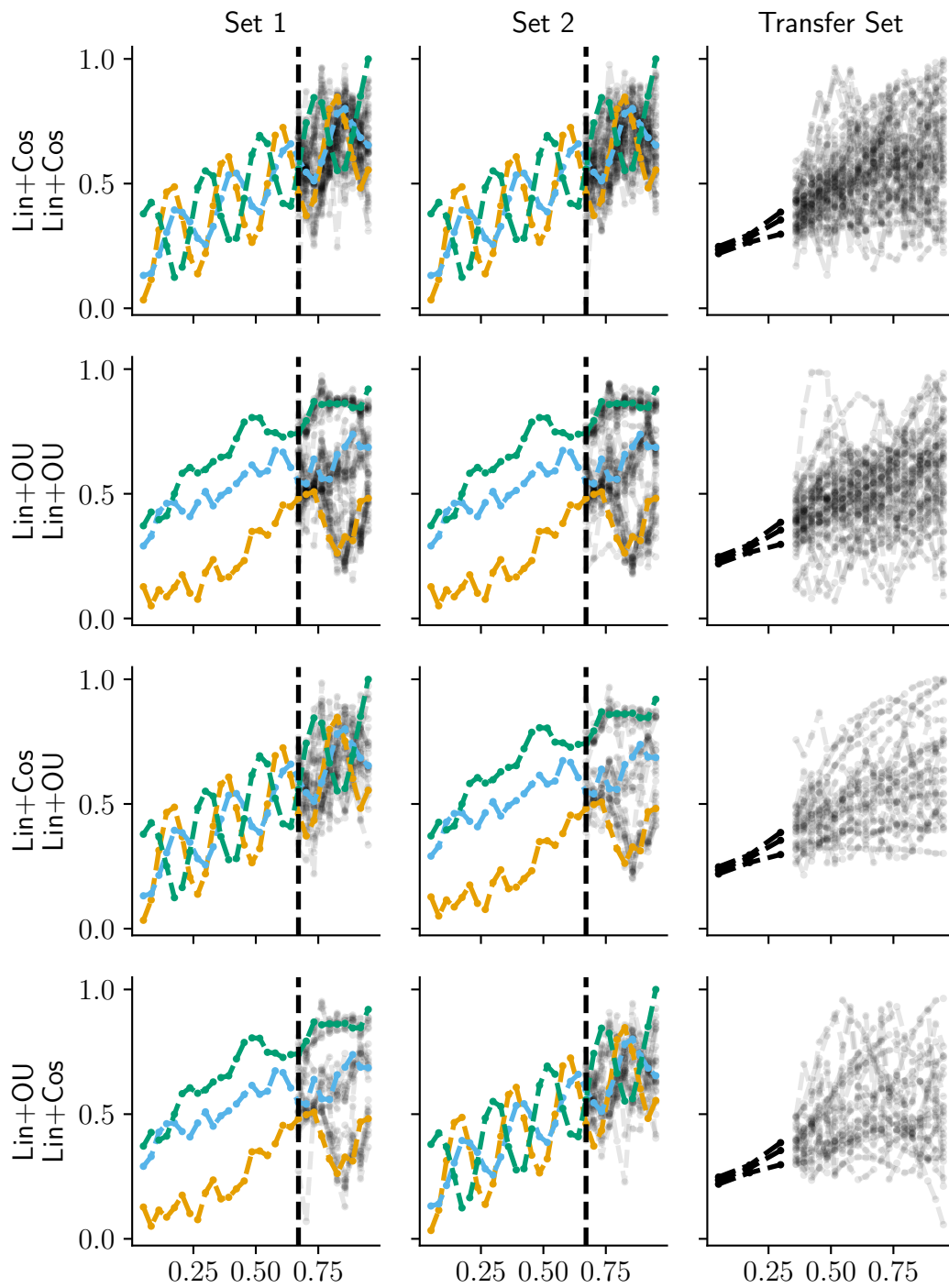
Figure 7.9: Extrapolations in the linear experimental conditions. Extrapolations in the transfer set generally exhibited positive slopes and, in many cases, additive periodic or high-noise structure.

For periodic conditions, extrapolations often suggested periodic or high-noise patterns. For 2×Cos+Lin conditions, patterns suggested more positive trends compared to the other conditions. For extrapolations in the two training sets and the transfer set for the cosine conditions, see Figure 7.10. While visual inspection suggested that many participants produced extrapolations consistent with the compositions or constituents of the functions learned in the training sets, the data is generally highly idiosyncratic and noisy. Thus, in the next section, we attempted a more systematic classification of the extrapolation patterns.

7.2.2.1 Recovering Function Types from Extrapolations

As in the previous chapters, we evaluated which generating functions accounted best for participants' extrapolation patterns. We ran *MLE* for each individual participant and each generating GP. Since our generating functions included kernel compositions, the number of free parameters and their interactions made the estimation task considerably more demanding. To aid inference and discourage degenerate and non-kernel-specific solutions we constrained the hyperparameter ranges[3]. This was necessary, since for instance, too-high *Cos* lengthscales in combination with too-low variance would amount to practically linear functions. On the other hand, too-short lengthscales can resemble extremely wiggly extrapolations, which should be better captured by an OU kernel.

We ran up to 400 optimization runs, adaptively increasing the number of optimizations for functions, including the periodic kernel. Optimization runs were increased to allow the periodic functions to achieve comparably good fits[4]. We then used the type of generating GP with the highest likelihood to predict each participants' condition. For a confusion matrix for all conditions and their

---

[3]$\lambda_{\mathrm{Cos}} \in [0.02, 0.05]$, $\sigma_{\mathrm{Cos}} \in [0.02, 02]$, $\lambda_{\mathrm{OU}} \in [0.02, 02]$, $\sigma_{\mathrm{OU}} \in [0.015, 0.2]$.

[4]Since we used the pure periodic kernel to fit the data, the model was not able to accommodate small deviations in frequency or amplitude (apart from the shared additive Gaussian noise, σ_{noise}). As a result, *MLE* struggled to find optimal parameter settings, as those were generally concentrated in very narrow regions of the parameter space and surrounding parametrizations exhibited extremely low likelihood scores.
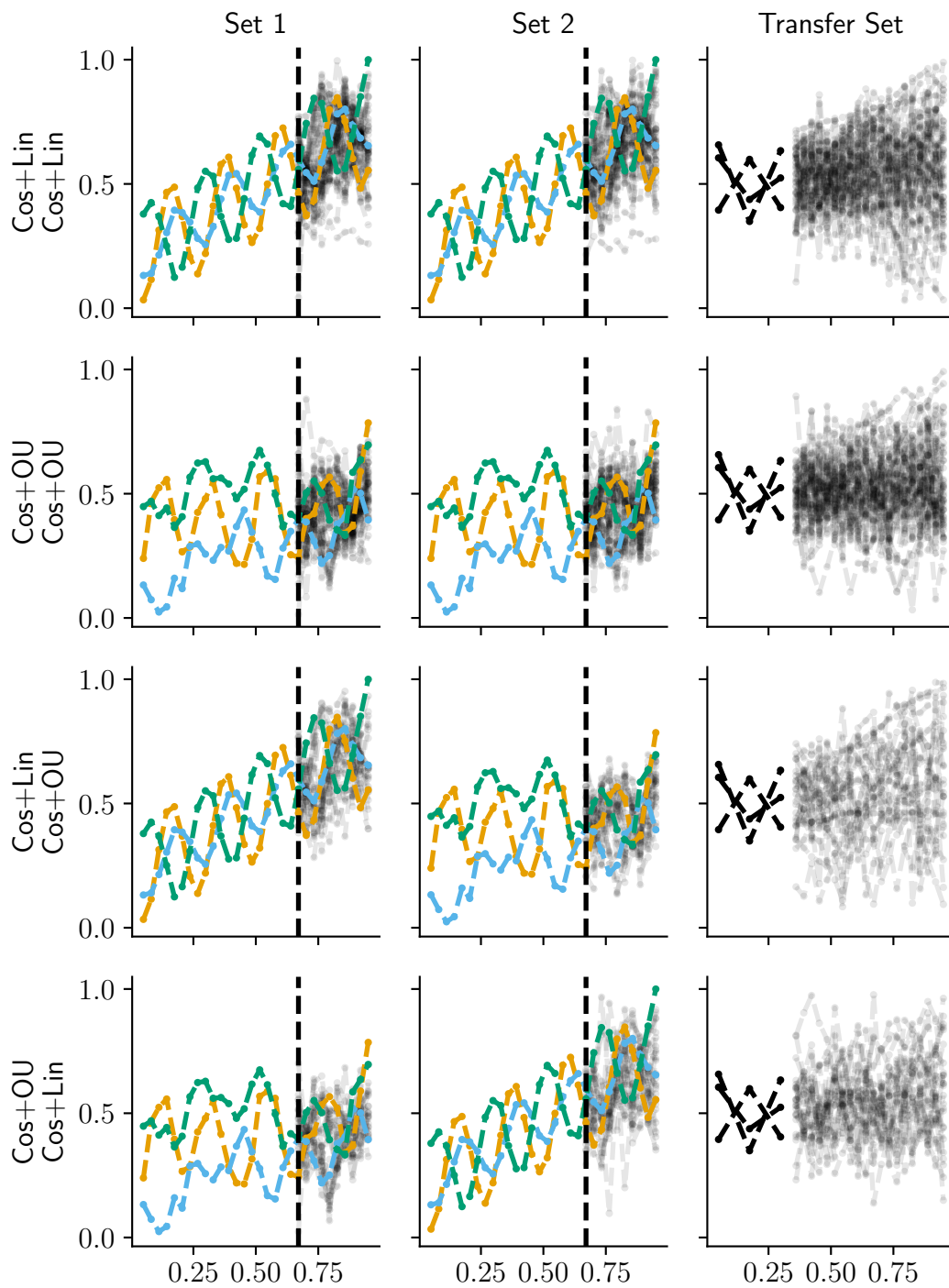
Figure 7.10: Extrapolations in the periodic conditions. In contrast to the linear conditions, participants' extrapolations exhibited less pronounced positive trends and more periodic or high-noise patterns.

best-fitting *MLE* functions, see Figure 7.11; for the five best-fitting extrapolations in each condition, see Figure 7.12 and Figure 7.13.
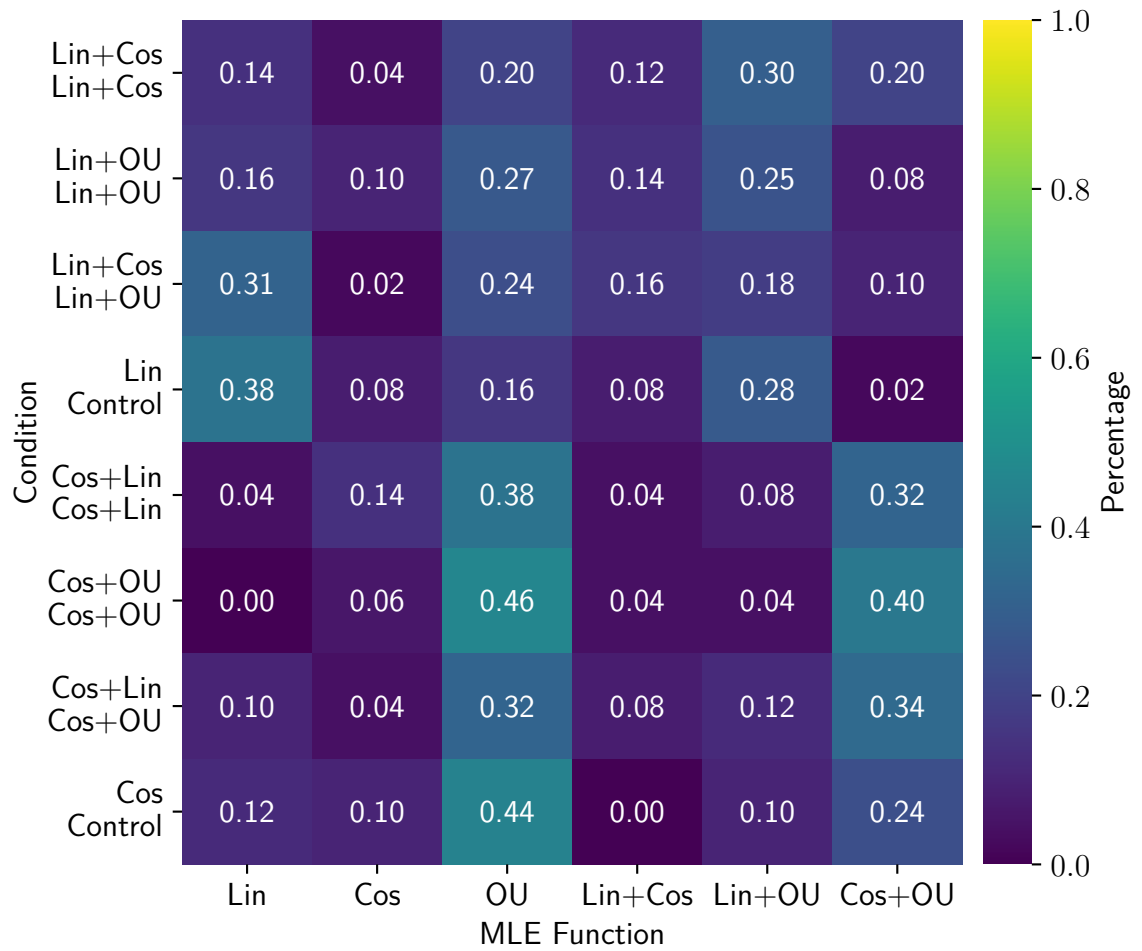
Figure 7.11: The *MLE* classification did heavily favor OU and Cos+OU in the periodic conditions. For linear functions, the control condition and Lin+Cos→Lin+OU were mostly classified as linear, whereas both 2×Lin+Cos was classified mostly as Lin+Cos→Lin+OU and 2×Lin+OU was mostly classified as OU.

For the linear control condition, the majority of extrapolations were best captured by a linear function (19 out of 50, 38%, $p < .001^5$). However Lin+OU was also assigned frequently (14 out of 50, 28%, $p < .05$). In the 2×Lin+Cos conditions, the majority of participants were assigned Lin+OU (15 out of 50, 30%, $p < .05$). For 2×Lin+OU the majority of participants' extrapolations were

---

[5]We again report one-sided, exact Binomial tests testing against chance (1/6).
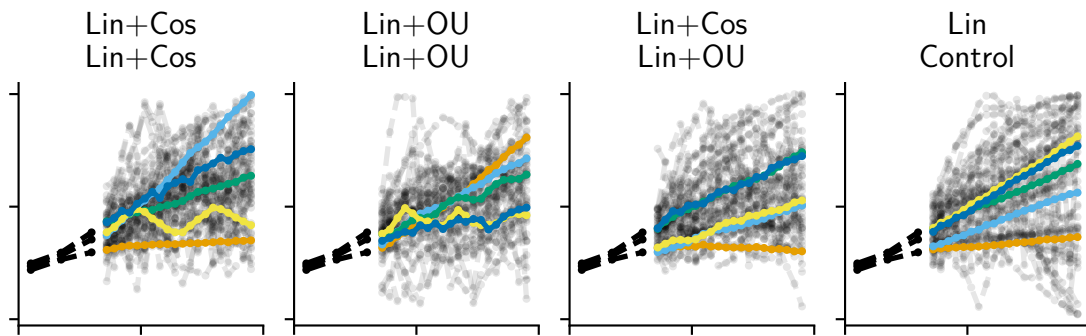
Figure 7.12: The five extrapolations with the highest likelihood scores in the four linear conditions.
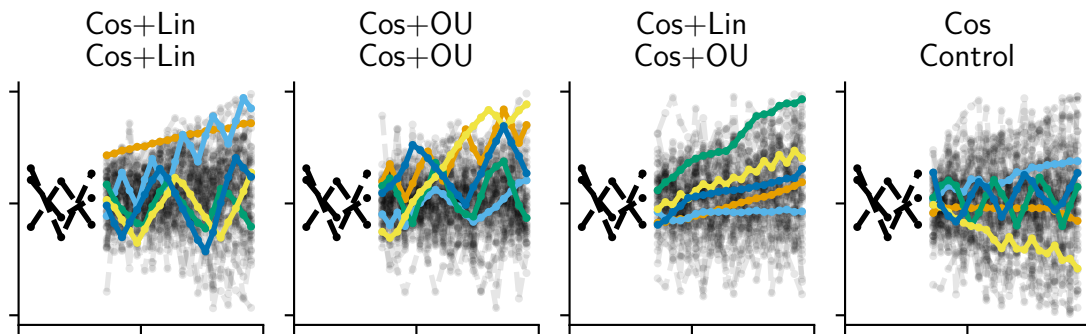
Figure 7.13: The five extrapolations with the highest likelihood scores in the four periodic conditions.

best accounted for by OU (14 out of 51, 27%, $p < .05$). For Lin+Cos→Lin+OU, extrapolations were mostly accounted for by linear functions (16 out of 51, 31%, $p < .05$).

For the experimental linear conditions, our procedure did not fully capture our expectations — for 2×Lin+Cos, practically no Lin+Cos functions were assigned. On the other hand, for 2×Lin+OU conditions, the true composition Lin+OU and both constituents were assigned relatively often. Similarly, in line with our hypothesis, for Lin+Cos→Lin+OU, both constituent functions often fit participants' extrapolations best. However, it is debatable whether this is due to the participants' extrapolations genuinely exhibiting these patterns or just because of an advantage of the OU kernel (and additive combinations involving OU) in capturing the idiosyncratic, but no-OU patterns.

The difficulty of accounting for the participants' extrapolation patterns based on the true generating functions was even more apparent in the periodic conditions. In the control condition, most extrapolations were assigned to the OU function (22 out of 50, 44%, $p < .001$). Similarly, for 2×Cos+Lin the majority of the extrapolations were assigned to OU (19 out of 50, 38%, $p < .001$), followed by Cos+OU (16 out of 50, 32%, $p < .001$). For 2×Cos+OU, again OU was assigned most often (23 out of 50, 46%, $p < .001$), followed by Cos+OU (20 out of 50, 40%, $p < .001$). For Cos+Lin→Cos+OU, OU was assigned most often (17 out of 50, 34%, $p < .001$), followed by OU (16 out of 50, 32%, $p < .001$).

One explanation for the large proportion of OU estimates is that across experimental and control conditions, participants' extrapolations resembled the rugged, autocorrelated OU-features, either because OU is highly salient and a priori favored, or because the inferred function was distorted due to the experimental setup and as a result exhibited rugged features. Visual inspection of the extrapolation patterns, especially in the Lin+Cos→Lin+OU and Cos+Lin→Cos+OU sets, suggested that some participants did provide such idiosyncratic and noisy

extrapolations. However, overall this explanation is not entirely convincing since participants did not favor OU in the control condition. Instead, participants' extrapolations were better accounted for by a linear function.

An alternative explanation is that the functions we fitted to the participant extrapolations were ill-equipped to account for the idiosyncratic extrapolations. While the pure periodic and linear GPs can produce stereotypical patterns that are ideal for generating experimental materials, these functions might be too rigid to account for the highly variable human data. Both the linear and pure cosine kernels have clear parametric analogs, and thus are ill-equipped to capture small deviations from the expected function patterns. In contrast, the OU kernel is more flexible and can capture the highly idiosyncratic human extrapolations.

While some of the participants' extrapolations exhibited rugged and autocor-related OU-like patterns, we favor this second explanation. Many of the extrapolations assigned to OU were fairly regular, stereotypical up-down, cyclic patterns, see Figures 7.16 and 7.17. Furthermore, when we ranked the assigned functions by their *MLE* score, less than 10% of the top 25% of scores corresponded to functions involving OU (OU: 1 out of 101, 1%; Lin+OU: 3 out of 101, 3%). Instead, linear and periodic combinations featured frequently (Lin: 54 out of 101, 53%; Lin+Cos: 25 out of 101, 25%; Cos: 18 out of 101, 18%). In contrast, for the bottom 25% of likelihood scores, OU and functions involving OU were the only functions assigned (OU: 43 out of 101, 43%; Cos+OU: 38 out of 101, 38%; Lin+OU: 20 out of 101, 20%); see also Figures 7.14 and 7.15 for the lowest-likelihood *MLE* extrapolations. For all extrapolations associated with a function type, see Figures 7.16 and 7.17.

7.2.3   Choices

We again contrasted the proportion estimates of a Dirichlet-Multinomial model against random choice (1/6) and the proportion selected in the control condi-
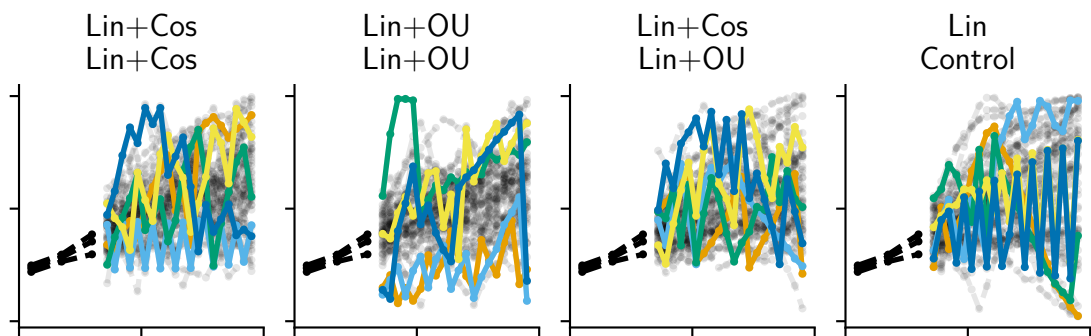
Figure 7.14: The five extrapolations with the lowest *MLE* scores in the four linear conditions.
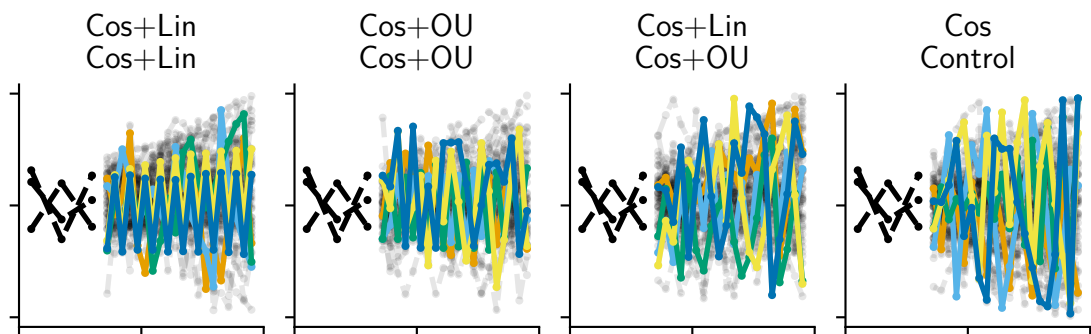
Figure 7.15: The five extrapolations with the lowest *MLE* scores in the four periodic conditions.
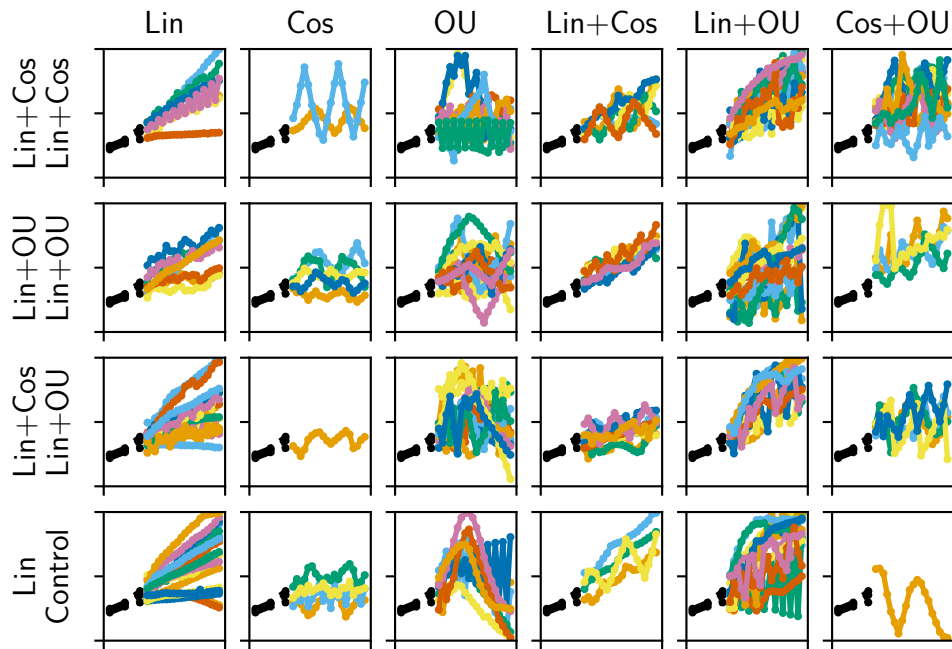
Figure 7.16: Extrapolations in the linear conditions assigned to the individual functions via MLE.
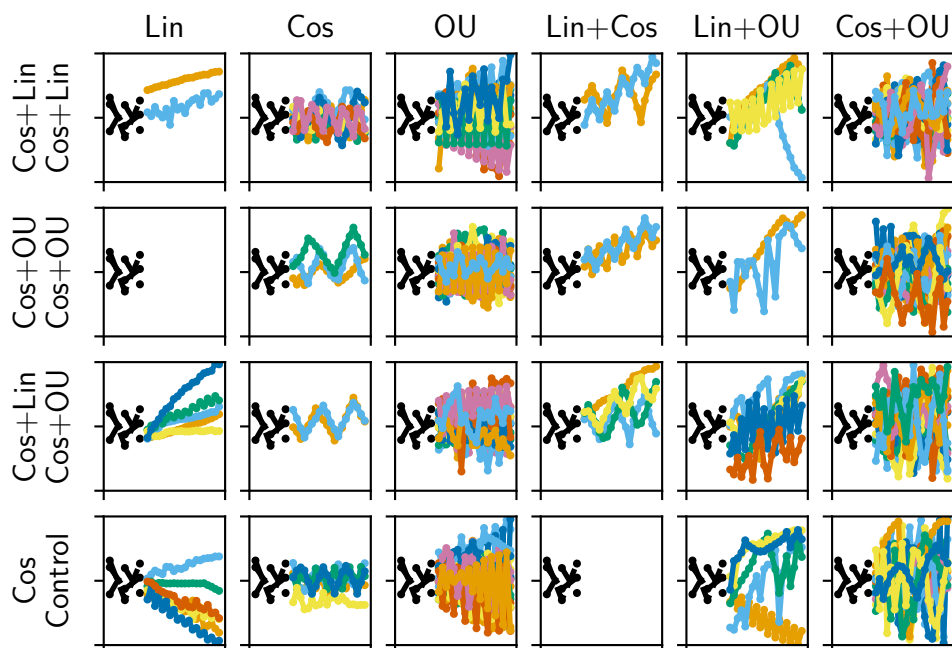
Figure 7.17: Extrapolations in the periodic conditions assigned to the individual functions via MLE.

tion[6]. For the linear transfer conditions, only in Lin+Cos→Lin+OU were proportions credibly higher than chance. Participants selected linear above chance in Lin+Cos→Lin+OU (16 out of 51, 31% and > 99% of the proportion estimates were larger than chance, $\hat{p}_{Linear>1/6} > .99$), but the estimated proportion was not higher than the corresponding control proportion.

For periodic conditions, proportions for Cos were higher than chance in the 2×Cos+Lin condition (20 out of 50, 40%, $\hat{p}_{Cos>1/6} > .99$), but not higher than the corresponding control proportion ($p_{control} = .38$, $\hat{p}_{Cos>.38} = .46$). Lin+Cos was selected higher than chance in the Cos+Lin→Cos+OU condition (14 out of 50, 28%, $\hat{p}_{Lin+Cos>1/6} > .97$). This proportion was higher than the corresponding control proportion ($p_{control} = .16$, $\hat{p}_{Lin+Cos>.16} = .86$). For estimated parameters and a contrast to the control conditions, see Figures 7.16 and 7.17. For all estimated proportions, see Figure 7.18, and Tables F.3 and F.4.

These results do not support our hypothesis. While for linear transfer sets proportions were comparable to control for Lin+Cos→Lin+OU, repeated conditions did not result in higher proportions of the intended compositional patterns. For periodic transfer sets, none of the proportions larger than chance corresponded to our hypothesis. For Cos+Lin→Cos+OU, participants selected Cos at lower rates than control, whereas for repeated conditions, participants did not select the intended composition at rates higher than chance.

Finally, we evaluated if the forced-choices corresponded to the function assigned to the extrapolations. If the assigned options corresponded to the choices performed by the participants, that would suggest that our classification scheme captured their inferred function. However, only in the linear control condition (19 out of 50, 38%, $p < .001$) and 2×Cos+OU (14 out of 50, 28%, $p < .05$) did choices correspond to the extrapolation classification. These results allow for two interpretations. One possibility is that participants did not choose consistent

---

[6]We also calculated exact Binomial tests which were consistent with the Bayesian estimates.
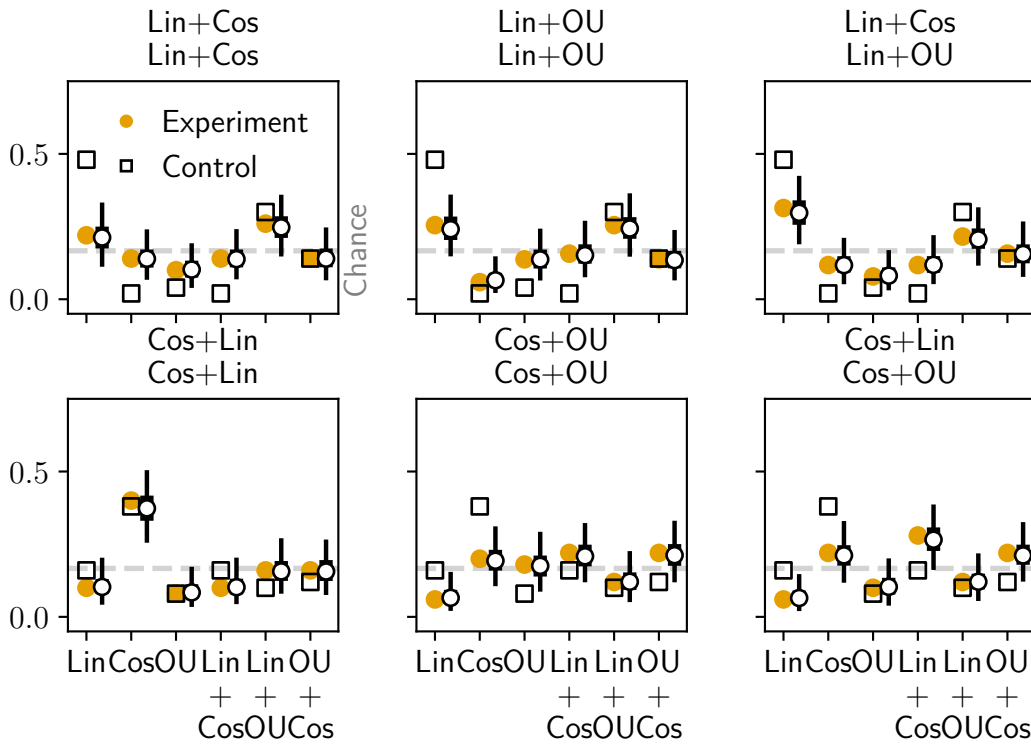
Figure 7.18: Proportion of options chosen in the three linear and periodic conditions (round marks) and control (square marks). Linear and Linear+OU were selected at rates higher than chance $(1/6)$ for linear transfer sets. Similarly, Cos was selected at rates higher than chance for $2\times$Cos+Lin, and Lin+Cos was selected at rates higher than chance for Cos+Lin$\rightarrow$Cos+OU. However, proportion estimates for the options selected in the experimental conditions (round marks with 5–95% *IQR*) revealed that these preferences were only larger than the corresponding control proportions for Lin+Cos.

with their previous extrapolation, either due to inattention or fatigue, or because none of the options were deemed similar to the intended function. Alternatively, these results provide further evidence that the classification scheme adopted did not faithfully capture participants' extrapolations and did not correspond to their choices. For per-choice accuracy scores, see Figure 7.19.
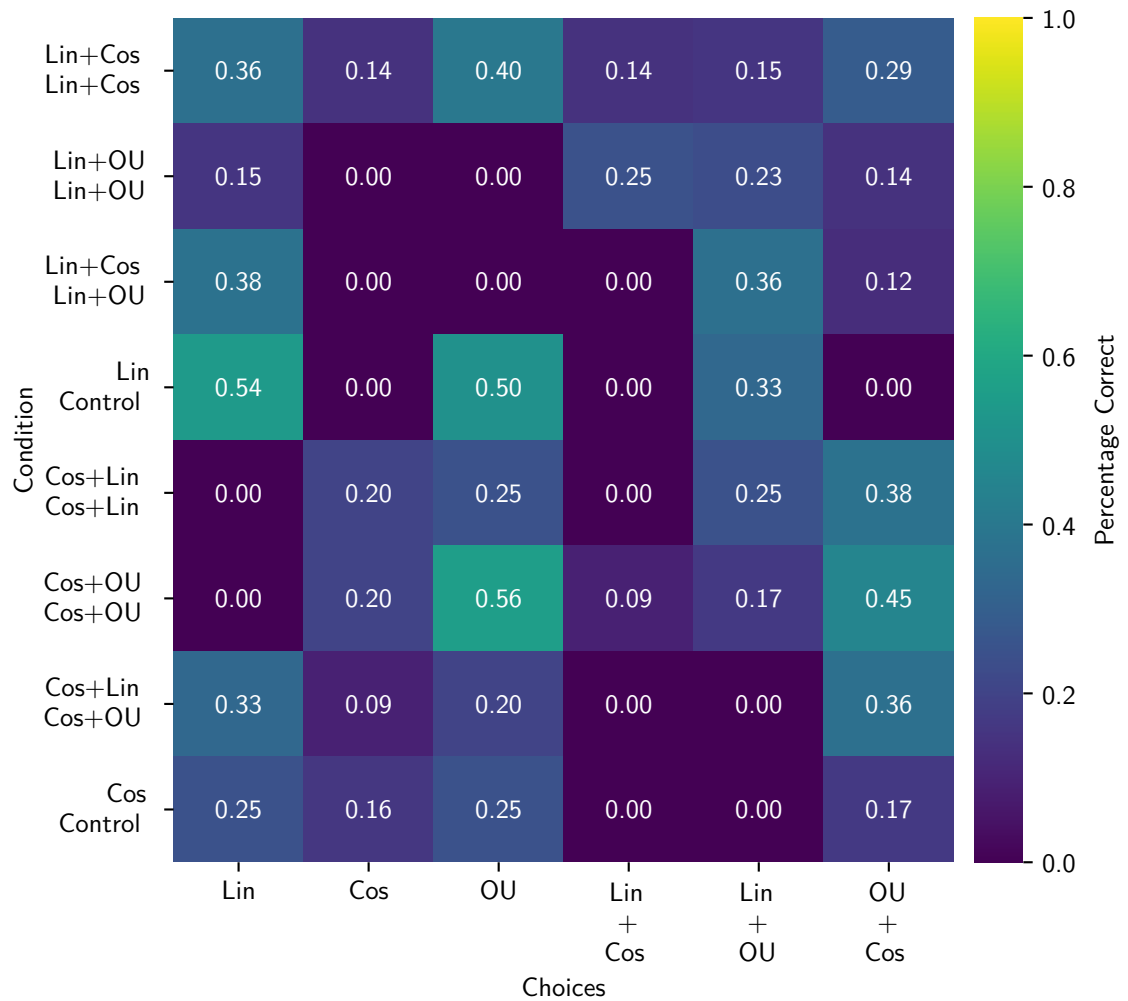
Figure 7.19: Accuracy of the classification scheme in matching the participants' choice. Only in the linear control condition did our method predict participants' choices.

7.3   Discussion

We hypothesized that participants preferred the simpler functions in the non-repeated conditions and extrapolated according to the structure of the training data in the repeated conditions. However, our data do not support any of these hypotheses. Thus, we did not find convincing evidence for compositional transfer or our more specific hypothesis of training repetition and generalization.

While some extrapolation patterns visually resembled the training, our classification of the extrapolations could not verify this impression. Many participants were better accounted for by OU-functions instead of the hypothesized functions. Results in the linear training conditions were more promising than in the periodic conditions, with control and Lin+Cos→Lin+OU conditions providing some evidence for our hypothesis. However, for both repeated training-set conditions, the majority of patterns were best described by OU or OU-compositions.

We hypothesize that this outcome is the result of the set of kernels we compared and the parameter estimation method. First, we used the same GPs to generate the training data and subsequently classify our participants. This approach was favored over adopting more complex and flexible GPs to better correspond to our previous analysis (Chapter 5) and to provide a close correspondence between the presented material and the computational method. To better capture the idiosyncratic human extrapolations, future work should instead adopt more flexible kernels. Given the GPs we contrasted, the overly rigid GPs were severely disfavored to account for these patterns compared to the OU GPs. Additionally, while Chapter 6 suggested that OU kernels produced more salient patterns than alternatives, the results in the linear control condition put into doubt whether participants realized the structure in the OU samples or if they treated them as noisy linear functions.

Second, our results raise a general concern – fitting idiosyncratic participant extrapolations poses fundamental challenges for function learning research. Fu-

ture research should attempt to improve on these results by adopting regularized or fully Bayesian models. However, even with regularization, two fundamental issues complicate analysis: participants' extrapolations often do not comply with standard assumptions about residual noise, and the *human kernel*[7] (Wilson et al., 2015) is unlikely to be in the researcher's set of candidate kernels. Given these sources of model misspecification and low amounts of data, minor differences in likelihoods or posterior distributions cannot reliably be interpreted as a meaningful inference of the human kernels or inferred functions. This problem is exacerbated by the fact that, unlike rule-learning models or language models where compositionality is often adopted, many of the commonly postulated human kernels can fit any pattern.

Thus, future research should attempt to better understand what human kernels are, what noise assumptions characterize human extrapolations, and how extrapolations relate to the inferred functions of the participants.

While our results did not provide much evidence for our hypothesis, we still think that encouraging conclusions can be derived. First, visually many samples exhibited the structures in the training set. These highly structured patterns are a motivating indicator that, even with very little training, participants can infer complex patterns. Our results, in combination with previous results by Schulz et al. (2017) and the results in Chapter 6, suggest that abstract, compositional learning mechanisms could provide the basis for flexible and far-ranging generalization mechanisms. Uncovering these mechanisms would resolve fundamental issues in function learning, and help resolve questions that have been outstanding since the first function learning experiments (Carroll, 1963; Brehmer, 1974).

In the final chapter, I summarize the contributions of this thesis and return to the question about how we can uncover human kernels and generalization mechanisms, both computationally and experimentally.

---

[7]Or, for that matter, individual, highly idiosyncratic *human kernels*.

Chapter 8

Conclusion

I have argued through a series of experiments and computational models that hypothesis spaces over abstract functions underpin human function generalization. Furthermore, these spaces are continuously refined and adapted to the tasks at hand. Before discussing the implications of these results and proposing future avenues for research, the next section will briefly summarize the main contributions of the thesis.

8.1   Contributions

Inductive biases are flexible and adaptive   The results across all chapters challenge the idea of an unspecific bias for linearity. In Chapter 3, we saw that participants extrapolated according to quadratic or periodic relationships, even when the data had to be remembered implicitly. In both Chapters 5 and 7, participants often extrapolated in non-linear fashion, even if the transfer data was linear. Thus, these results suggest that previously reported biases for linearity are too strong. However, these results do not necessarily contradict previous research. Instead, they suggest that given contextual information, such as previous training tasks, and in less memory-taxing experimental setups, participants can overcome default preferences.

People track high-level features of functions   Throughout this thesis, we have seen that people represent high-level features of functions even when details of these functions are not tracked precisely. These abstract features can amount to the abstract type of function (Chapters 3 and 5), variance (Chapter 4), or compositional structure (Chapters 6 and 7). These results contrast previous accounts of function learning as parameter estimation. These one-size-fits-all models propose that when faced with a function learning task, participants learn the details of that function and not information about the its type. Instead, we propose that tracking these high-level features provides reusable abstractions that allow far-ranging transfer.

People can transfer abstract information about the function learned   This thesis provides first experimental evidence that people can use high-level features about previous functions in subsequent tasks. People can transfer the information learned in flexible ways, in Chapter 5 by applying the abstract function type, or, in Chapters 6 and 7, by applying knowledge of how functions combine, to new situations. These results expand our understanding of how people excel at far-ranging extrapolations. Previous research has treated function learning as a domain-general process. These models cannot account for the flexible transfer of knowledge observed in our experiments. We suggest that domain-specific learning allows more flexible learning and stronger generalizations.

People can perceive, recognize, and transfer additive compositional structures   In Chapters 6 and 7, we saw that people can perceive, recognize, and transfer compositional functional structure. While people were able to infer an abstract rule from as few as one presentation, we also saw that they did so predominantly for additive compositions. These results, in combination with earlier results by Schulz et al. (2017), suggest that people can perceive deep latent structure in patterns, form abstract compositional hypothesis spaces, and, in some

cases, apply this knowledge in order to extrapolate. These results provide important insights into the structure and origin of the hypothesis space of functions. Previous approaches had to postulate an ever-expanding, but ultimately narrow, set of candidate functions to account for the flexibility of human extrapolations. Instead, our results highlight that a small set of broadly applicable functions, combined with compositional principles, can produce very flexible and complex hypotheses.

8.2 Open Questions and Implications

The results of this thesis raise several important questions and suggest future experiments. First, experiments in this thesis always explicitly instructed participants that the relationships belonged to the same underlying pattern. Thus, it is critical to determine if our results generalize to more realistic setups, in which no explicit instruction about the underlying function is given. Second, the analysis in this thesis has highlighted the technical and experimental difficulties encountered when attempting to uncover human kernels and their parametrizations. Third, while this thesis has expanded our understanding of learning at the level of the hypothesis space, less is known about how and when individual subtypes of functions, such as positive and negative linear functions, are learned. Finally, these results highlight the importance of adopting a wide range of methods and experimental paradigms to uncover the basis of human generalization. I will discuss these open questions and suggest possible experimental and methodological approaches in the next paragraphs before briefly stating the implications of this work for human generalization research.

Implicit learning of shared generative functions All our experiments explicitly instructed participants to treat patterns as "underlying the same structure". However, in the real world, this information is rarely directly available and, in-

stead, people have to learn shared structure implicitly. Future research should thus focus on how and when people infer these structural similarities. These experiments will likely require considerably more training data and also have to take into account how context-relevant features of the domain are introduced and displayed experimentally.

Difficulties uncovering human inductive biases   A second important issue regards the efforts of reverse-engineering the inductive biases underlying human generalizations. Previous work has focused either on parametric, rule-like forms (Carroll, 1963; Brehmer, 1971), associative mechanisms (McDaniel and Busemeyer, 2005), or has approached the task from a computational level (Lucas et al., 2015). Given these theoretical commitments, previous research has then inferred parameters of these rules (Brehmer, 1974), or hybrid mechanisms (Busemeyer et al., 1997), or suggested human kernels (Wilson et al., 2015). From a computational perspective, Gaussian processes seemed especially promising, as they can encompass both the flexibility of associative learning and, via the mean function, the parametric commitments of rule-learning (Lucas et al., 2015; Schulz et al., 2017).

However, the results in this thesis suggest that in many cases, Gaussian processes exhibited similar problems as previous rule-based or associative approaches. First, introducing deterministic mean functions as a surrogate for participants' far-ranging extrapolations introduces the same issues as previous rule-based approaches: the resulting extrapolations are mainly dictated by the parametric function and are often too rigid. Second, the flexibility of the kernels might not correspond to the way humans extrapolate. For instance, the variance of a linear kernel does not express the firm commitments to intercepts and slopes that many human extrapolations exhibit. These issues arise both when fitting models to participant data, as well as when creating experimental materials. Third, the kernels

discussed in this thesis, and in previous work, often cannot capture the human patterns without considerable additional constraints or regularization. For example, RBF or OU kernels capture smooth or rough patterns and can fit any data. Thus, only in conjunction with parameter biases are these kernels informative for human generalization.

In addition to the question about kernels and kernel parameters, the data collected in this thesis highlights that, echoing the results in other areas of psychological research (Gilden et al., 1995), standard noise assumptions do not necessarily match the data. Participants might occasionally produce extrapolations corresponding to the maxima of the target scale or may even invert their implied functions (as seen in Chapter 3), resulting in complex residual distributions. Estimating model parameters as characteristic of human inductive biases with models that exhibit mismatched noise assumptions then risks being biased, as the kernel hyperparameters have to capture these idiosyncrasies. Two possible approaches to this problem – one experimental and one computational– seem promising.

First, instead of drawing or submitting individual points, participants could submit several function extrapolations, thus producing average or aggregate per-participant patterns. Also, experiments could include uncertainty estimates for individual data points to better understand the range of values entailed by the participants' behaviors.

A second approach would focus on the computational process modeling the behavior. First, more flexible heavy-tailed processes could be fitted instead of Gaussian processes (for instance, Student's-T processes). Heavy-tailed processes would allow the model to disregard extreme outliers and produce a more robust estimate of the participants' regressions. However, these processes make computations costly, as they do not have the favorable analytical properties of GPs. Furthermore, these processes would not allow a straightforward compositional approach, as compositions of non-Gaussian processes do not necessarily produce

valid stochastic processes.

Overall, these results highlight that to characterize human function extrapolations fully, we require human kernels, human kernel parameters, and human noise distributions. Therefore, using GPs to capture human generalization is not a worry-free computational approach, capable of characterizing all features of human extrapolations. Without a thorough grasp of kernel types, parameters, and noise, we lose our ability to reliably interpret these models as characteristic features of human inferences. Echoing the sentiment of MacKay (2003), one might ask if by replacing previous rule or associative models with GPs, there is a risk of "throwing out the baby with the bathwater".

Learning function subtypes   Previous research has highlighted the importance of particular parametric forms in function extrapolation, such as the positive matched linear function. However, it has also highlighted the inverse of this function, the negative linear, as highly salient. A critical theoretical question is how these salient subtypes (in this case, linear functions) are learned and maintained. For instance, how does a repeated presentation of a particular function affect the extrapolation of alternative functions belonging to that function type?

One possibility is that these kernel types are learned analogous to nonparametric models in machine learning. Repeated exposure to data amounts to the introduction of new kinds of functions. New instances close enough to that type will be associated with it, whereas situations that are different are assigned to alternative functions. These questions could be informed by new experimental tasks, such as expanding on the work in Chapter 4 by obtaining graded similarity judgments between function realizations for functions of different types.

Similar models and experiments could be explored regarding compositional learning: Are new compositions cached as entities for reuse, or do people have to assemble them continually? One exciting prospect would approach this compu-

tationally from a language learning perspective, for example, by applying ideas from nonparametric grammar learning (Johnson et al., 2007).

Finally, to truly uncover how these hypothesis spaces are learned in the first place, future research should focus on developmental experiments. Developmental research has shown that many general inductive biases appear early in infancy (for an overview, see Xu, 2019) and young children can exhibit different inductive biases than adults (Lucas et al., 2014). Obtaining such experimental results would be vital to uncovering human kernels and the origins and mechanisms of how new functions can be learned and adapted.

Uncovering the basis of human generalization   Previous research and the experiments presented in this thesis have highlighted the importance of testing human extrapolation patterns to reveal human inductive biases. However, this work has also highlighted the difficulty of capturing these extrapolations, given that human extrapolations are variable and complex. These results suggest two valuable insights for future research.

First, future work should focus closely on individual-level data and develop computational models capable of capturing the richness of human extrapolations. Second, as suggested in Chapter 4, complementary experimental approaches should be adopted to reveal human generalization. This includes the further development of experiments, such as iterated learning (Kalish et al., 2007), memory-effects (Schulz et al., 2017), or Markov chain Monte Carlo with people (Chapter 4). However, this should also include closer analysis of the *psychological* hypothesis space, for instance, adopting experimental setups such as multidimensional scaling, or categorization tasks for function patterns. These approaches would also allow us to further test the results of this thesis, namely that abstract features of functions are preferred representations.

One experimental test of our results regards the influence of training order

on the inferred abstract functions. Our results in Chapter 3 suggest that people are surprisingly good at detecting high-level features even when data is not readily available, and less capable at remembering the exact details of the function. These results suggest that, similar to results in categorization (Mathy and Feldman, 2009), manipulating the presentation order to facilitate or hinder inferences about the type of underlying function, would result in a strong preference for the suggested function. Previous work in function learning has only focused on systematically-increasing sequences or random sequences (Byun, 1995) and quadratic functions (Kwantes et al., 2012). Systematic presentations resulted in lower training error (Byun, 1995), and manipulation of the implied steepness of subsequent training points resulted in steeper linear extrapolations compared to orders that emphasized shallowness (Kwantes et al., 2012).

Future research should elaborate on these results and examine if other features apart from shallowness and steepness, such as sampling rate for periodic functions, affect what kind of function is inferred.

Implications

Overall, the results presented in this thesis expand our understanding of how humans generalize. When participants were instructed that patterns followed the same underlying structure, they adapted their prior expectations and performed far-ranging generalizations. These generalizations followed flexible, highly structured, and often compositional inductive biases.

The thesis also highlights that function learning entails abstract, systematic, and compositional, as well as graded and flexible learning processes. Bridging statistical and symbolic learning, function learning provides an ideal field of study for the representational underpinnings of human learning, generalization, and transfer.

Appendix A

Gaussian Processes

Throughout this thesis, we have encountered the task of fitting intricate regression patterns. At the same time, we usually needed to express these functions in terms of the abstract principles that underlie those patterns. Gaussian processes allow us to express features of functions, such as smoothness, periodicity, or roughness, without having to commit to a particular parametric form. At the same time, GPs can include parametric mean functions that capture long-range extrapolation trends. Finally, combining individual GPs, for example, by addition, results in complex, compositional models.

This chapter gives a brief overview of GPs for regression tasks, focusing mainly on how GPs allow us to express abstract functional features through the kernel and its hyperparameters and compose GPs through addition and multiplication. For a more thorough introduction to GPs, see Rasmussen and Williams (2006); for an up-to-date tutorial on GPs as a modeling tool in cognitive science, see Schulz et al. (2018).

A.1   What are Gaussian Processes?

A Gaussian process is a collection of random variables, of which any finite subset has a joint Gaussian distribution (Rasmussen and Williams, 2006). A Gaussian

process specifies a distribution over functions $f(x) \sim GP(\mu, k)$, where $\mu(x) = E[f(x)]$ and k is the covariance function $k(x, x') = cov(f(x), f(x'))$.

The kernel defines how much values of x depend on the other values x' and specifies a similarity measure over x. Therefore the characteristics of a Gaussian process model crucially rests on the choice of the kernel and its parametrization. For a selection of kernel functions, see Figure A.1. For samples from GPs with those kernel functions, see Figure A.2.
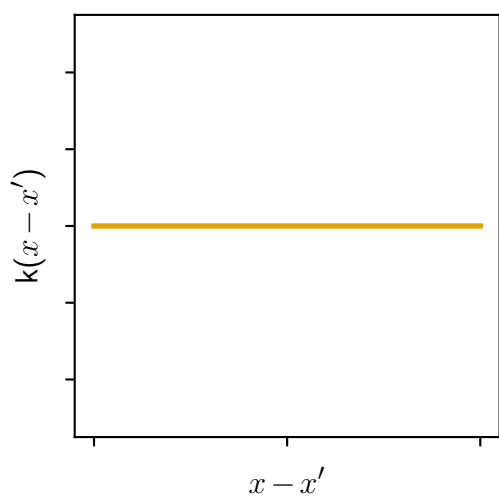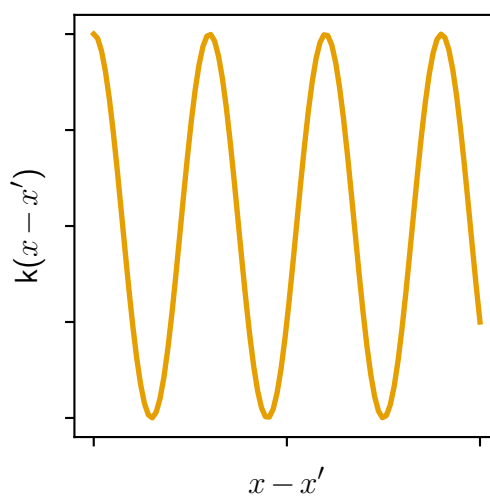
(a) The linear kernel.   (b) The cosine kernel.
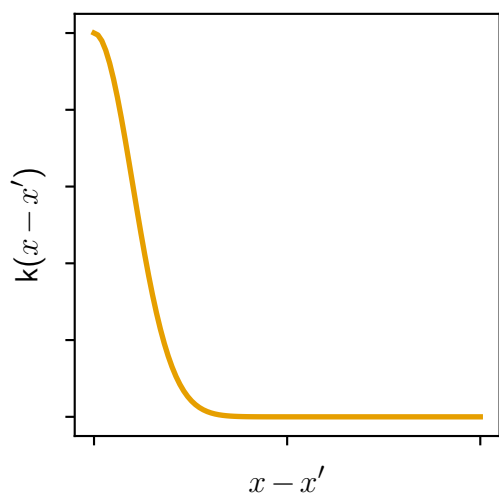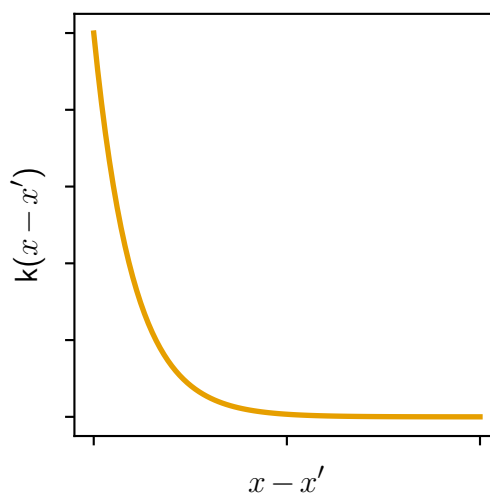
(c) The radial basis function kernel.   (d) The Ornstein-Uhlenbeck kernel

Figure A.1: The kernel as a function of distance $x - x'$.

Many kernels have been used in statistics and machine learning, one of the most widely used being the squared exponential basis kernel:

$$k_{RBF}(x, x') = \exp(-0.5||x - x'||^2).$$

Hyperparameters can be added, for example a length-scale parameter l resulting in:

$$k_{RBF}(x, x') = \exp(-0.5||x - x'||^2/l^2).$$
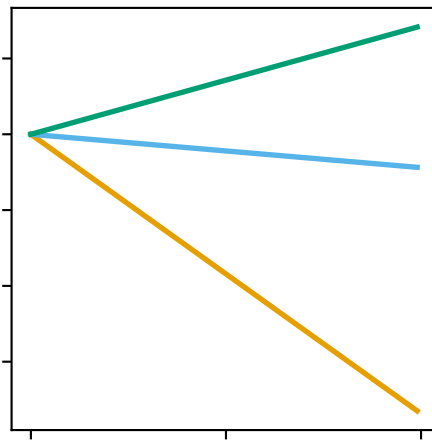
(a) Samples from a linear-kernel GP.
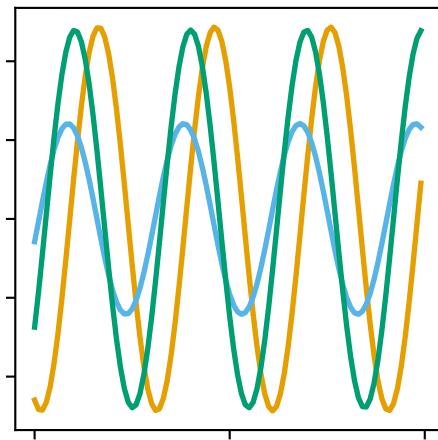
(b) Samples from a cosine-kernel GP.
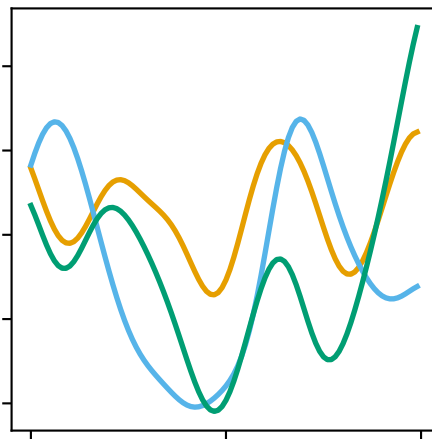
(c) Samples from an RBF-kernel GP.
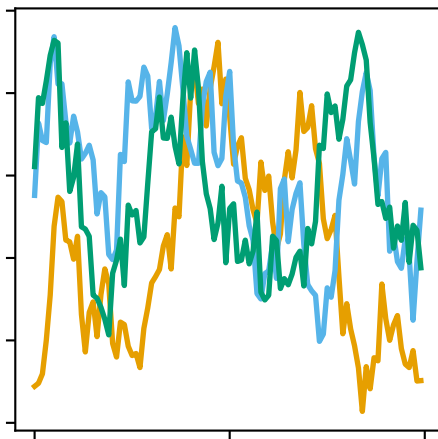
(d) Samples from an OU-kernel GP.

Figure A.2: Three samples from Gaussian processes with different kernels.

A.2 Sampling from the Prior and Conditioning on Data

We can sample from the GP at any collection of points X by evaluating the kernel function $K(X,X)$ and sampling $f(X)$ from a multivariate Gaussian distribution $\mathcal{N}(\mu(X), K(X,X))$. For samples from the priors of an RBF kernel with long and short lengthscales, see Figures A.3a and A.3c.

More useful for practical applications is calculating the posterior distribution of f given some observations. For GPs, this amounts to conditioning the joint Gaussian prior distribution on the observations. Crucially, for fixed hyperparameters, this conditional distribution can be obtained analytically. For posterior means and posterior samples for the GPS with the RBF kernel with long and short lengthscales, see Figures A.3b and A.3d. Given the commitment to a particular kernel or a composition of kernels, fitting a Gaussian process to data amounts to learning the set of hyperparameters specified by the kernel. Varying those hyperparameters will result in very different extrapolations (see Figure A.3d).

A.3 Composing Gaussian Processes

Kernels can be combined by multiplication or addition, resulting in more complex kernel structures (Duvenaud et al., 2013). For example, adding a RBF kernel and a linear kernel results in smoothly varying periodic patterns:

$$k_{Linear+Periodic} = k_{Linear}(x, x') + k_{Periodic}(x, x')$$

For examples of the kernels obtained by addition and multiplication, see Figure A.5), for samples from those kernels, see Figure A.4.

Expressing kernel functions as compositions allows for intuitive interpretation of functional constituents. In this thesis, expressing the kernel as a composition
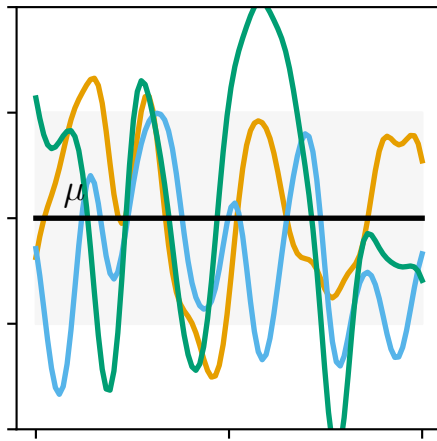
(a) Long-lengthscale prior samples.
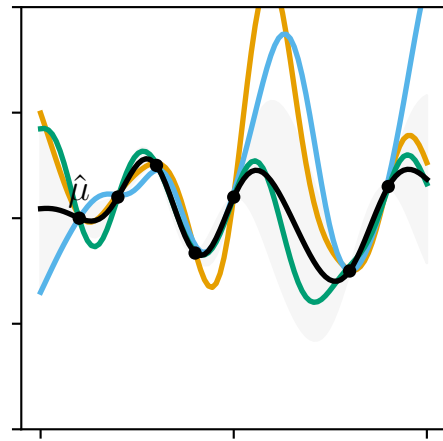
(b) Long-lengthscale posterior samples.
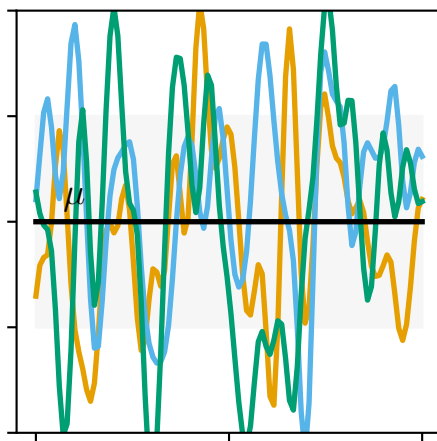
(c) Short-lengthscale prior samples.
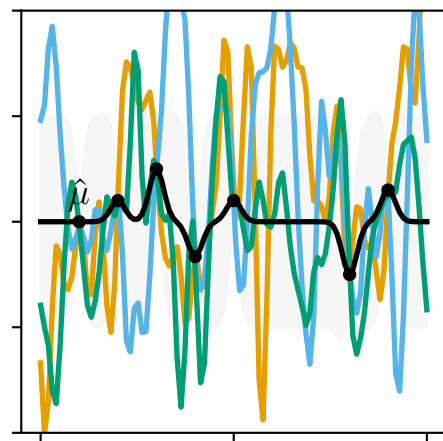
(d) Short-lengthscale posterior samples.

Figure A.3: Prior and posterior mean and $2SD$ (gray area), as well as samples from two RBF kernels.

of simple elements allows us to propose hypothesis spaces in which compositional elements can result combining simpler constituents.

(a) Linear+Cos kernel   (b) Linear+OU kernel   (c) Cos+OU kernel

(d) Linear×Linear kernel   (e) Linear×Cos kernel   (f) Cos×RBF kernel
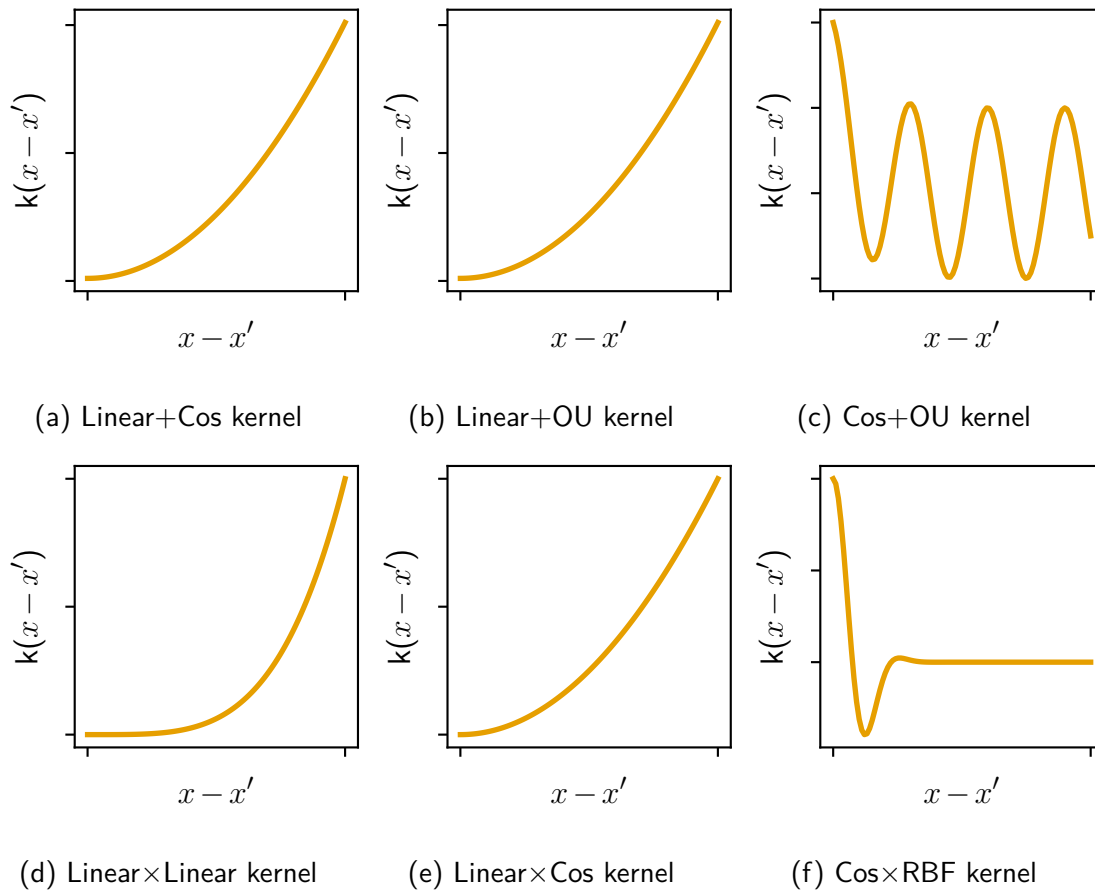
Figure A.4: Kernel functions obtained by addition (first row), or multiplication (second row).

(a) Linear+Cos   (b) Linear+OU   (c) Cos+OU

(d) Linear×Linear   (e) Linear×Cos   (f) Cos×RBF
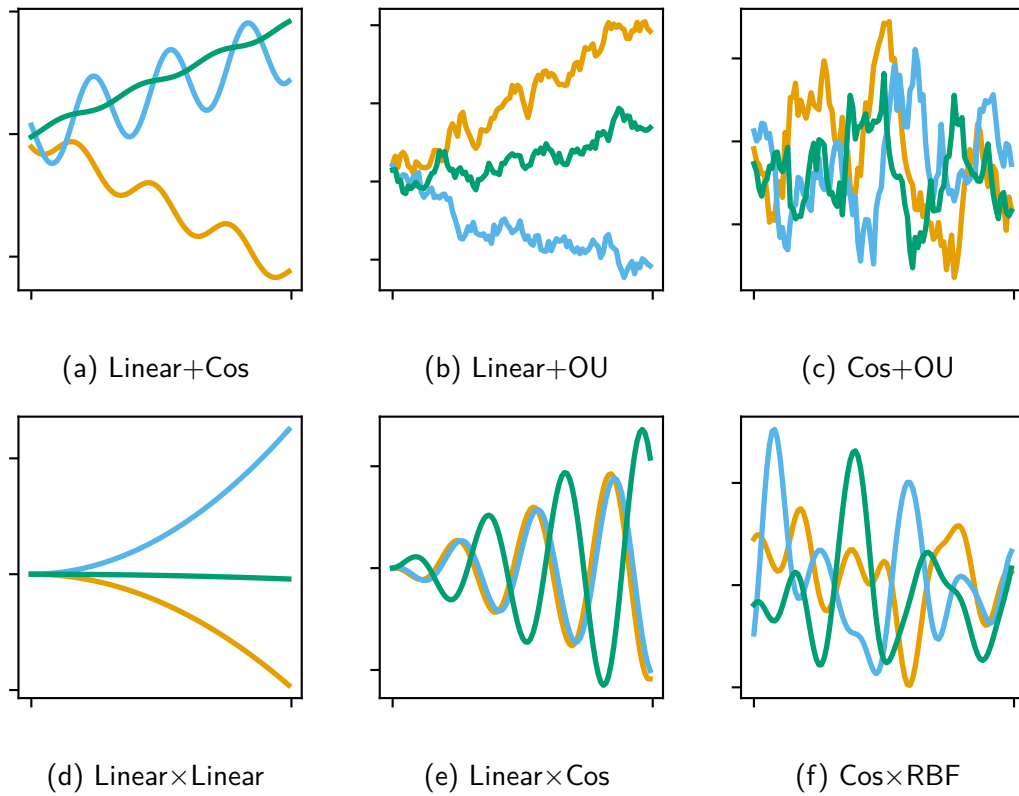
Figure A.5: Three samples from Gaussian processes obtained by adding (first row) or multiplying (second row).

Appendix B

Function Representation and Generalization
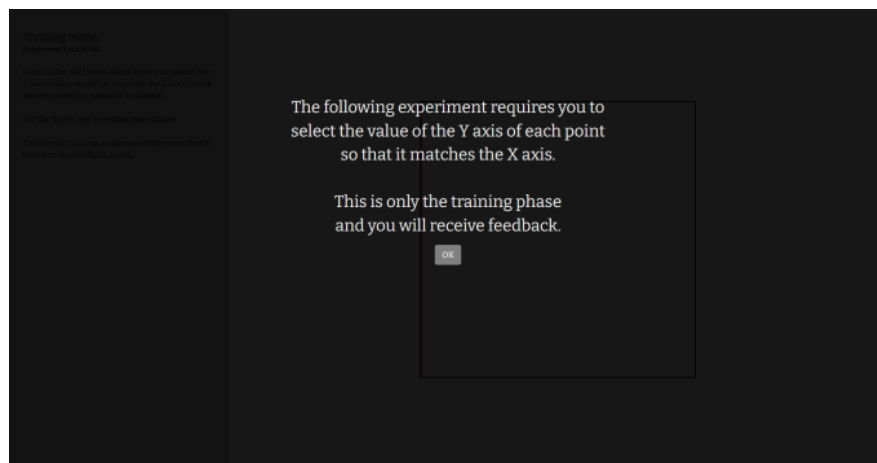
Figure B.1: The first screen of the experimental block in the scatter plot conditions introduced participants to the general task.
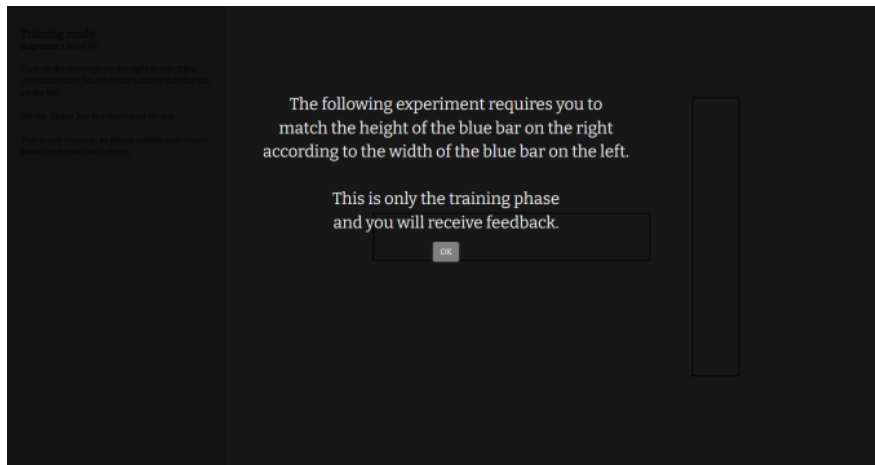
Figure B.2: The first screen of the experimental block in the Bar condition.
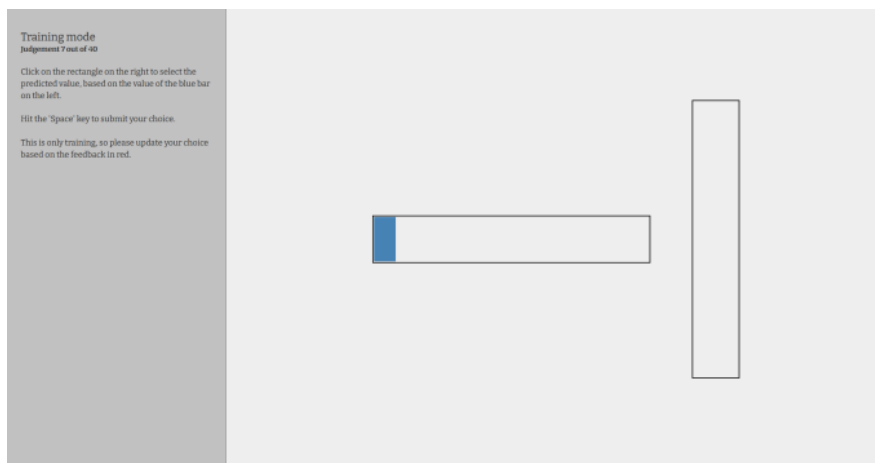
Figure B.3: In the Bar condition, participants had to predict the vertical bar's value on the right, given the left horizontal bar's extent.
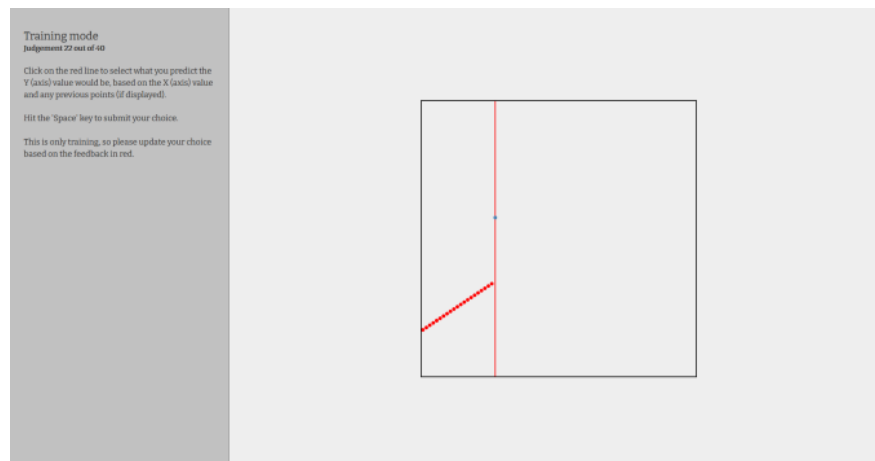
Figure B.4: In the scatter plot conditions, participants had to predict substance y on the y-axis.
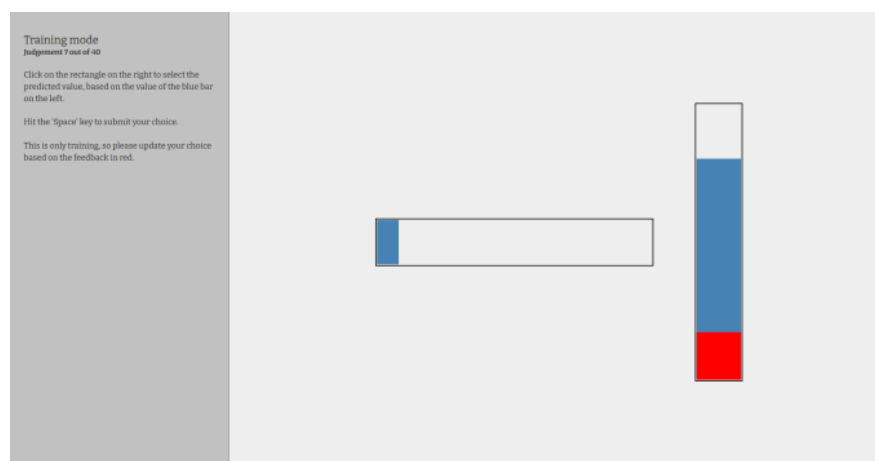
Figure B.5: During the training phase, participants received feedback. In the Bar condition, the true value was presented as a red bar.
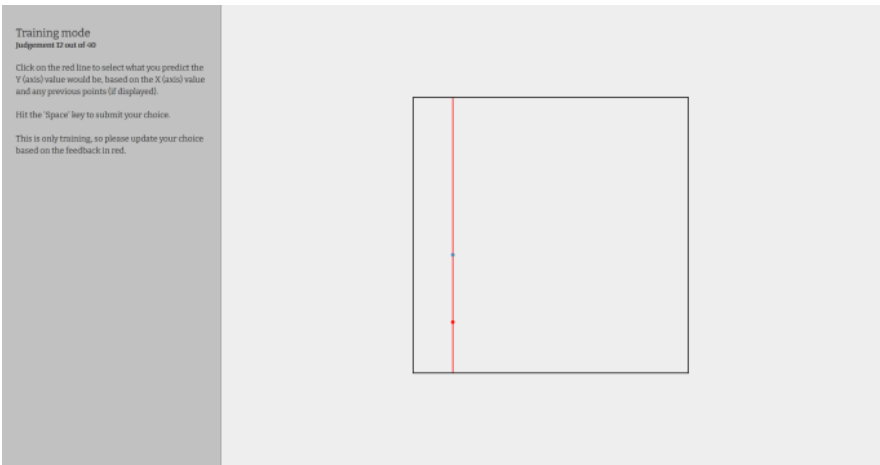
Figure B.6: During the training phase, participants received feedback in the form of a red dot displaying the true value.

Appendix C

A Distributional Space of Functions
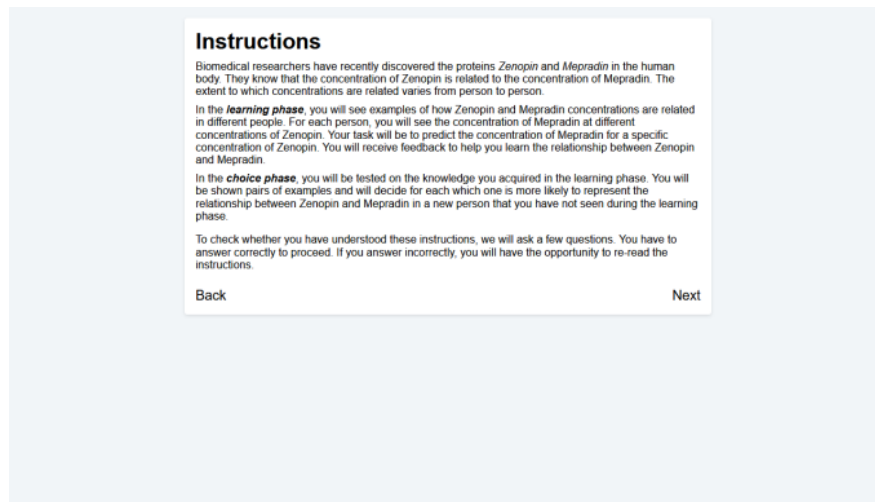
Figure C.1: The first instruction screen of the experiment introduced the general task.

Figure C.2: After the instruction screen, participants had to pass a comprehension check.

Figure C.3: After the instruction screens, the learning phase was explained.
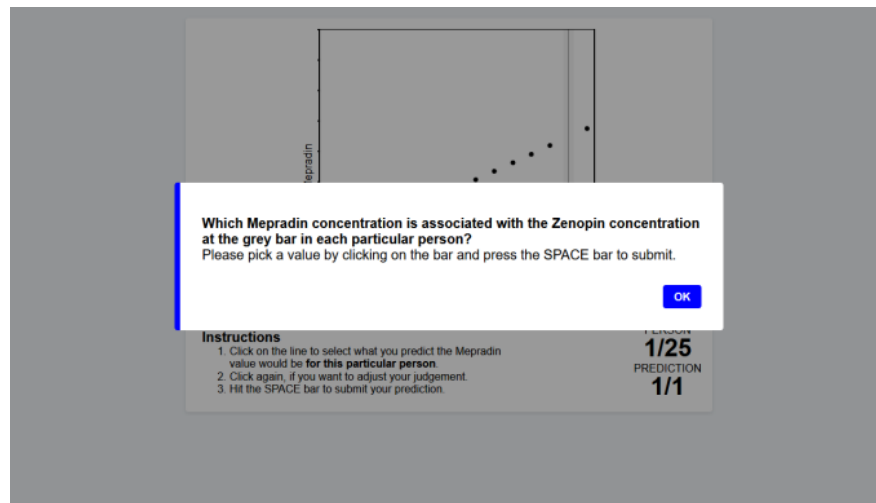
Figure C.4: At the beginning of the learning phase, participants received additional instruction.
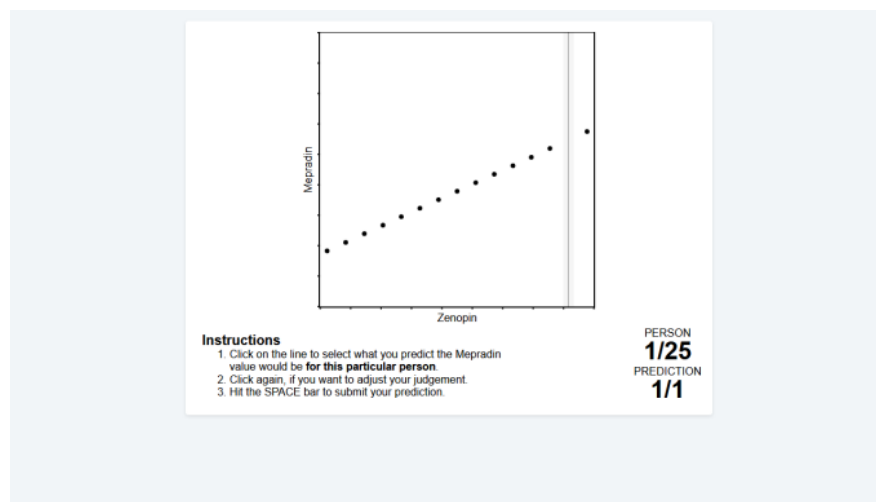
Figure C.5: Then, participants had to predict the value for one value of each pattern for 25 patterns.
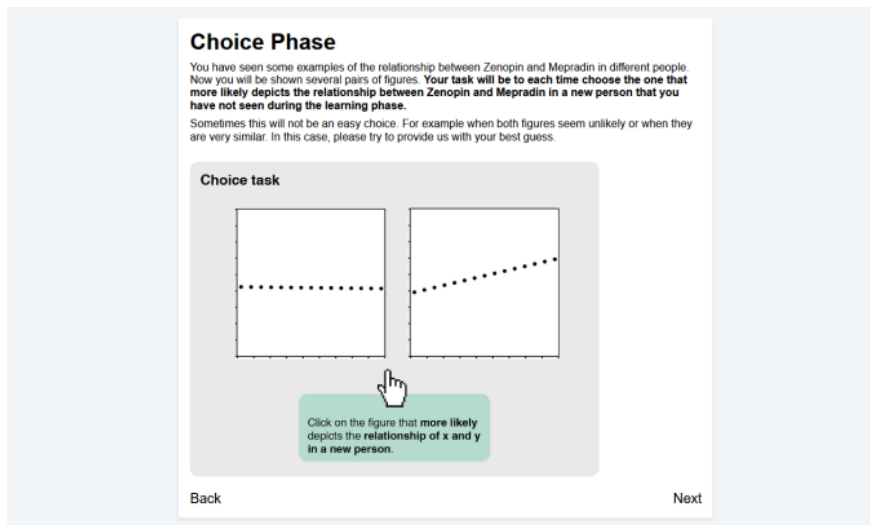
Figure C.6: After the learning phase, participants received further instructions for the main experimental section.
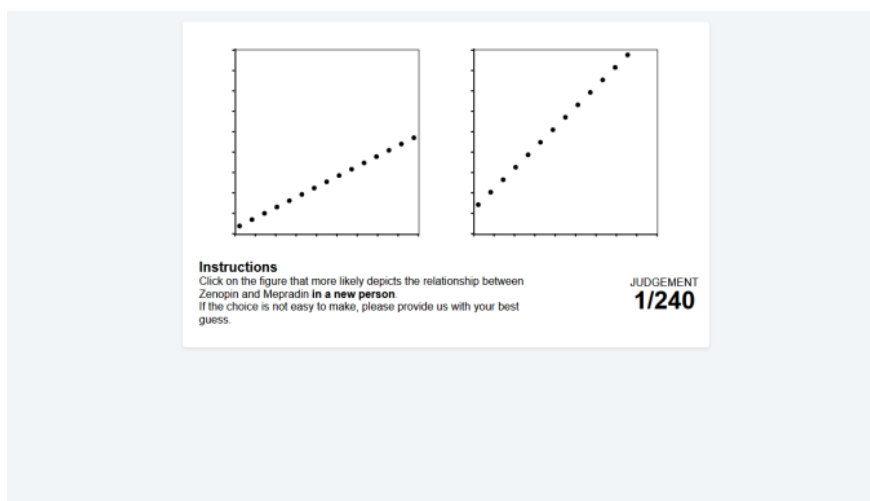
Figure C.7: In the main part of the experiment, participants had to choose one of the two patterns for 240 trials.

Appendix D

Transferring Functions and Parametrizations

D.1 Experiment

Welcome!

The purpose of the study is to investigate how people learn relationships between quantities.
Please complete the study in one sitting without taking breaks.

In rare cases (usually 1-5%) we may reject HIT submissions where people are not trying or paying attention.

Next
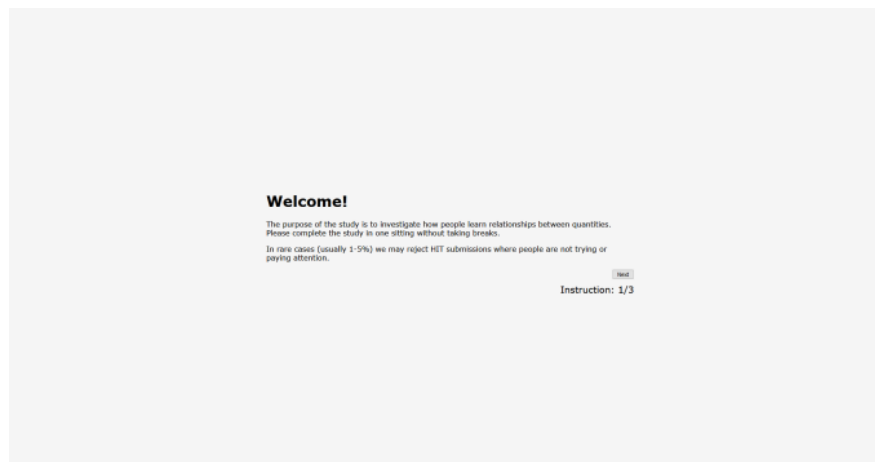
Instruction: 1/3

Figure D.1: The main instruction screen introduced the participants to the task.
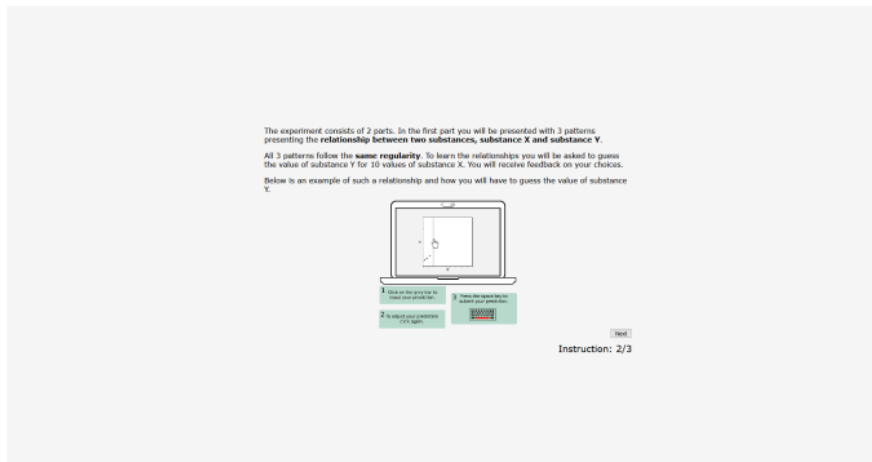
Figure D.2: After the general introduction, participants were shown an example of the experimental screen.

Figure D.3: Finally, participants were introduced to the experimental structure.
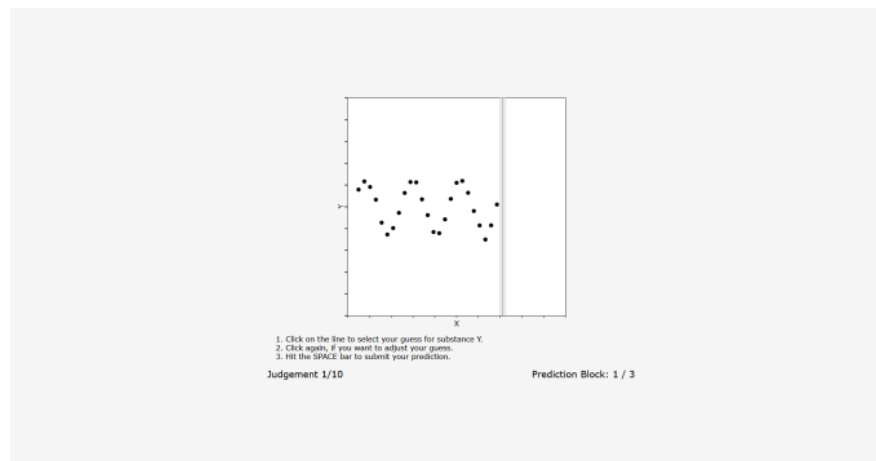
Figure D.4: Then, participants had to learn the relationships in the patterns over two training blocks.

Figure D.5: After training, the transfer task was introduced. Here, we display the transfer instructions for the forced-choice experiment.
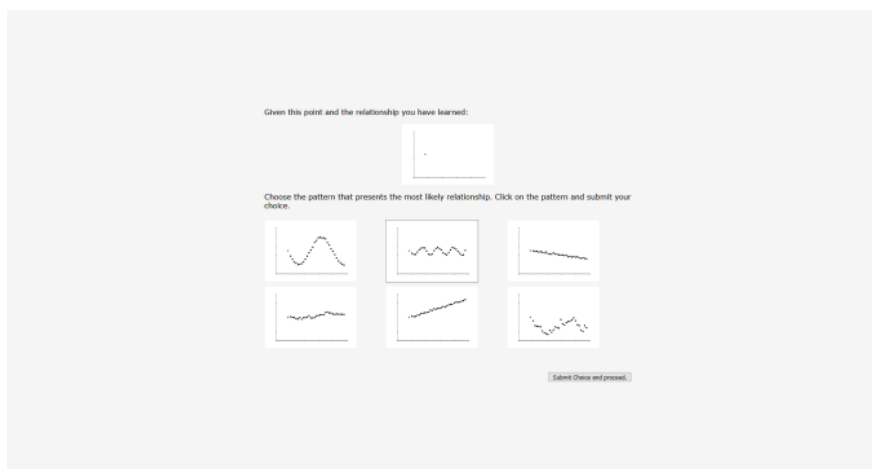
Figure D.6: In the forced-choice transfer block, participants had to select the most likely pattern from six candidates. In the extrapolation condition, participants performed an extrapolation task that followed the same design as the training blocks.

D.2 Error Models

We specified all models in PyMC3, (Salvatier et al., 2016) and obtained posterior distributions via three chains of NUTS Hamiltonian Monte Carlo sampling (Hoffman and Gelman, 2014). We chose sufficiently large tuning runs to obtain satisfactory convergence diagnostics (all $\hat{R} < 1.1$, ESS > 1000).

D.2.1 Log-normal Error Models

In addition to convergence diagnostics we were also interested in obtaining posterior predictive draws, i.e. predictions of the model after training. These draws provide us with a visual way of model criticism. Figure D.7 shows that our model captured the error distribution fairly well. Note that while the empirical error
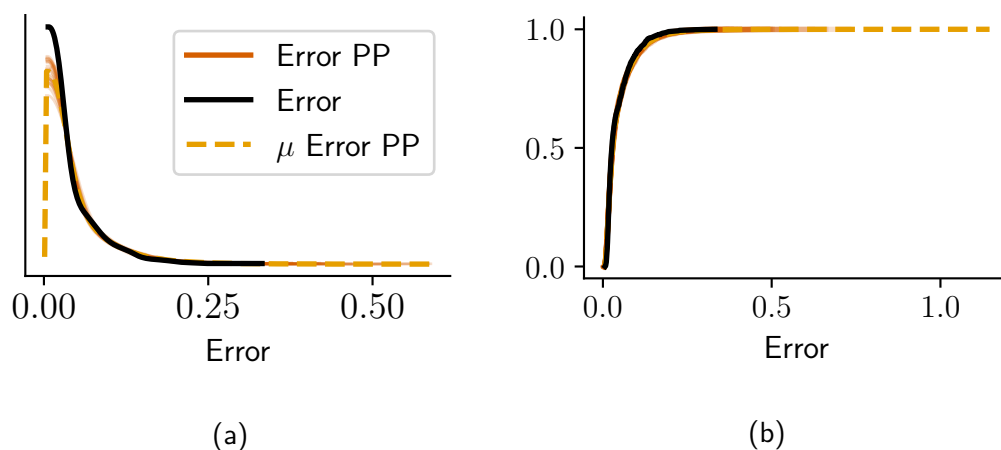
Figure D.7: Mean posterior predictive density (a) and cumulative density (b), (μ Error PP, dashed orange line) and 25 draws from the posterior (Error PP, solid orange line) for the log-normal hierarchical model. Our model captures the general shape of the empirical observations well (Error, black line).

distribution is upper-bound by 1, our model is not.

A further test of our model's ability to capture the empirically observed errors is to compare posterior predictive draws against per-participant error plots. Figure D.8 shows that participants differed considerably in their training error, with

some participants exhibiting non-decreasing error. Our model captured group- and individual errors reasonably well.
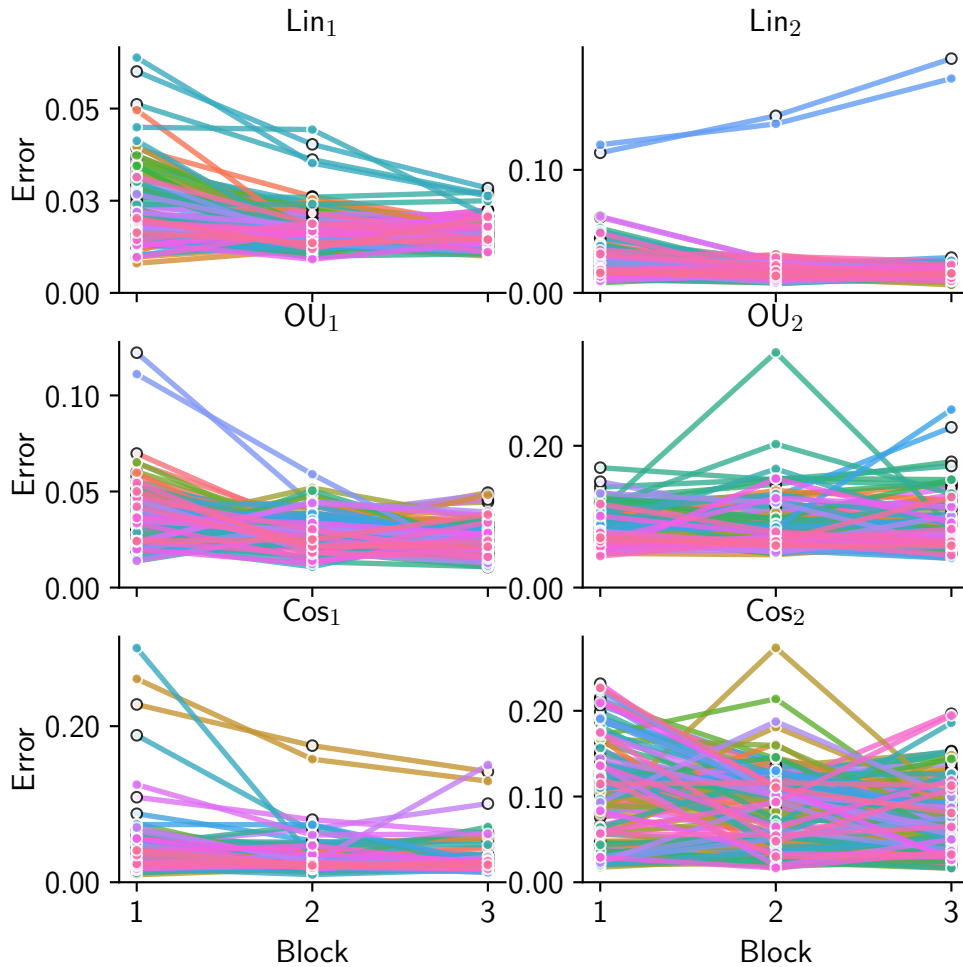
Figure D.8: Per-participant training errors (coloured markers and lines) and corresponding, per-participant mean of the posterior predictive distribution (white markers and coloured lines).

D.2.2 Exponential Decay Model

We fitted a hierarchical Bayesian exponential-decay model on the learning rates, with individual per-participant intercepts and slopes. Our model corresponds to the model in Kalish (2013) and expresses individual participants' errors as an exponential, strictly decreasing function of block. The hierarchical struc-

ture allowed us to capture both group-level variation and individual variation in terms of initial difficulty (intercept, k) and error decay (slope, d) as: error $\sim \mathcal{N}(\mu, \sigma)$, where $\mu = ke^{d \times block}$. Participants' s intercepts and slopes were pooled within their corresponding experimental condition c, with, $k_{sc} \sim \Gamma(\alpha_{ki}, \beta_k))$ and $d_{sc} \sim \Gamma(\alpha_{di}, \beta_d))$. Diverting from Kalish (2013) we use less dispersed hyperpriors, $\alpha, \beta, \sigma \sim \Gamma(0.01, 0.01)$.

We obtain group-level intercepts, $x_{kc} = \frac{a_{kc}}{b_k}$ and learning rates $x_{dc} = \frac{a_{dc}}{b_d}$. Group-level intercepts for both linear conditions, as well as the low-variance OU and the slow periodic were generally low, and error decayed across blocks, at a rate of ≈ 0.2.
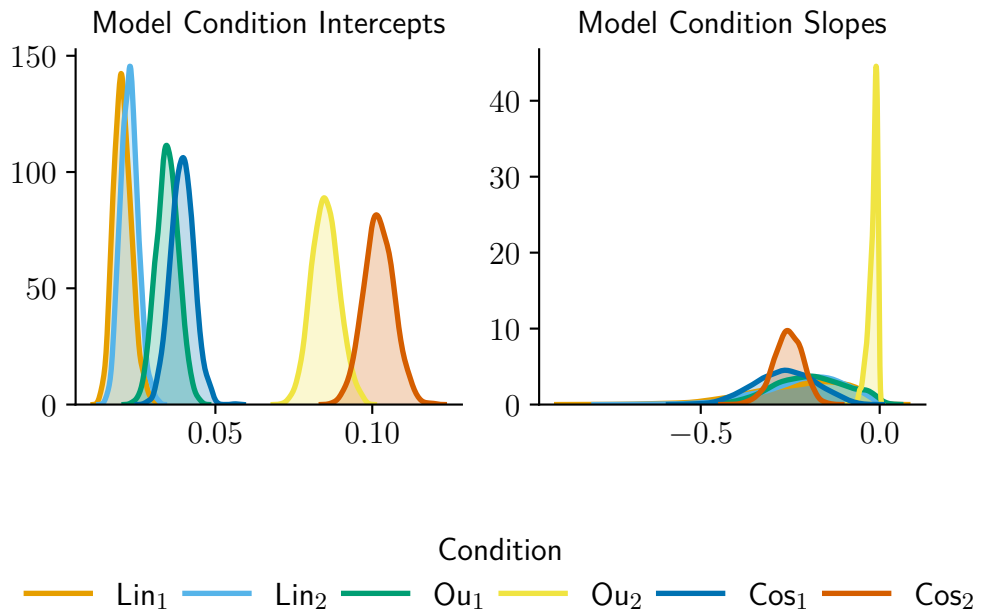
Figure D.9: Group-level estimates of error intercepts and slopes estimated via the hierarchical exponential-decay model. Both high-variance OU and the fast periodic condition exhibit large initial errors in contrast to the remaining conditions. While the error for the fast periodic condition decreases over blocks, the high-variance OU error remains high.

However, our posterior estimates for slopes exhibit relatively large uncertainty. This uncertainty stems directly from the model specification. Given that the

error is $\mu = ke^{d \times block}$, for values of k close to 0, d has only negligible influence on the overall error. In contrast, for high-variance OU, initial errors were high and did not change over training. Similar to high-variance OU, initial errors for fast periodic functions were high. However, these errors decreased significantly over training blocks, at a rate of ≈ 0.3. For group-level intercepts and slopes estimated by our model, see Figure D.9, for estimated parameters and highest posterior density intervals, see Table D.1.

Table D.1: Group-level estimated means $\hat{M}$ for intercepts, β_0 and slopes, β_0, as well as 95% highest-posterior density intervals estimated via MCMC for the exponential decay model.

	$\hat{M}_{\beta_0}$	$HPD_{95}\beta_0$	$\hat{M}_{\beta_1}$	$HPD_{95}\beta_1$
Lin$_1$	0.02	[0.02, 0.03]	0.23	[0.03, 0.48]
Lin$_2$	0.02	[0.02, 0.03]	0.21	[0.03, 0.41]
OU$_1$	0.03	[0.03, 0.04]	0.19	[0.01, 0.35]
OU$_2$	0.09	[0.08, 0.09]	0.02	[0.00, 0.04]
Cos$_1$	0.04	[0.03, 0.05]	0.27	[0.11, 0.43]
Cos$_2$	0.1	[0.09, 0.11]	0.26	[0.18, 0.34]

Close examination of the posterior predictive distributions, both for error, D.10 and individual, per-participant error, D.11 revealed that our estimates did sometimes not capture the data well, as some participants were not best described by monotonically decreasing functions. Finally, the normality assumption of the dependent variable resulted in our model underestimating the strong concentration of the empirical data.
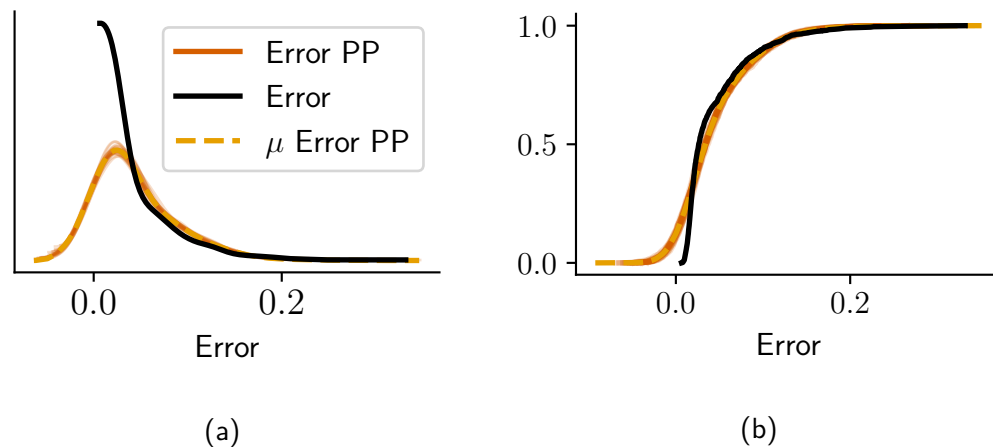
Figure D.10: Mean posterior predictive density (a) and cumulative density (b, μ Error PP, dashed orange line) and 25 draws from the posterior (Error PP, solid orange line) for the exponential-decay hierarchical model. The model captures the empirical observations fairly well (Error, black line), but the overall shape is less strongly concentrated.

D.2.3 Model Comparisons

We have seen that the posterior predictive for the log-normal was better aligned with the empirical data than the exponential-decay model. To further quantitatively evaluate the models' predictive accuracy, we evaluated the out-of-sample log-likelihood for draws from the posterior predictive (Vehtari et al., 2017). We contrasted them via the widely applicable information criterion, WAIC Watanabe (2010).

In addition to the log-normal (LogPP) and the exponential decay model (Exponential), we also compared a log-normal model that did not have per-participant intercepts and slopes $log(error) \sim \beta0 + \beta1 \times block$, (Log) and a linear model $error \sim \beta0 + \beta1 \times block$ (Linear). We found that the per-participant intercept and slope log-normal model fitted our data best, see Table D.2.
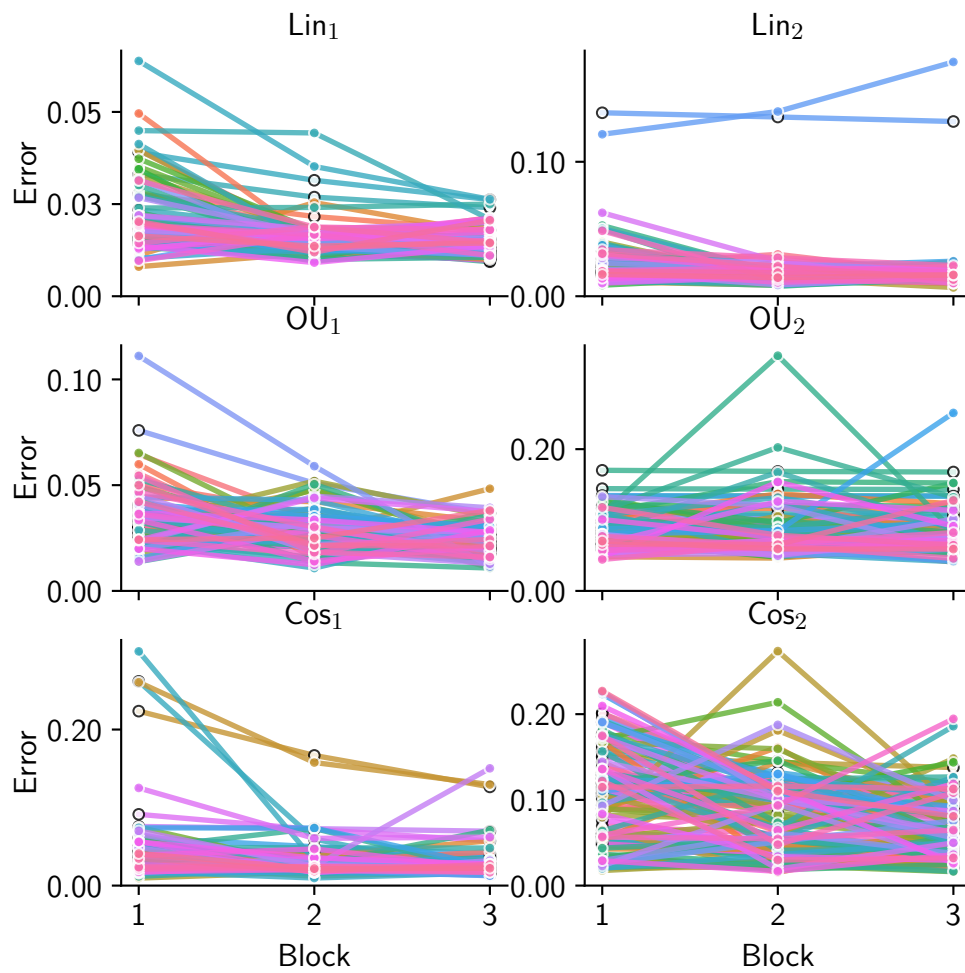
Figure D.11: Per-participant training errors (coloured markers and lines) and corresponding, per-participant mean of the posterior predictive distribution for the hierarchical exponential model (white markers and coloured lines).

	Rank	WAIC	p_{WAIC}	d_{WAIC}
LogPP	0	3222.61	450.03	0
Log	1	3122.29	13.91	100.32
Exponential	2	2620.82	333.14	601.78
Linear	3	2323.13	20.18	899.48

Table D.2: The four models compared and their rank. WAIC values correspond to posterior-predictive fits, where higher values reflect better fits, p_{WAIC} are estimated number of parameters, and d_{WAIC} is the relative difference for WAIC scores and the best-fitting (rank 0) model.

D.3 Choice Model

To estimate the proportions for each option, we modeled per-condition Dirichlet-Multinomial models. Each distribution of proportions was modeled as *choice* $\sim$ $Mult(n, \theta)$, where $\theta \sim Dirichlet(a)$ and $a = 1$. For the estimated mean proportions and HPDs, see Table D.3 and D.4.

Table D.3: Per-condition proportion estimates (mean $\hat{M}$ and 95% HPD intervals) obtained via the Dirichlet-Multinomial model for the 3-point experiment.

	Lin_1	Lin_2	OU_1	OU_2	Cos_1	Cos_2
$\hat{M}_{Lin1}$	0.55	0.05	0.14	0.09	0.09	0.09
HPD_{Lin1}	[0.36, 0.74]	[0, 0.13]	[0.02, 0.28]	[0, 0.21]	[0, 0.21]	[0, 0.21]
$\hat{M}_{Lin2}$	0.18	0.55	0.09	0.09	0.05	0.04
HPD_{Lin2}	[0.05, 0.33]	[0.36, 0.75]	[0, 0.21]	[0.01, 0.22]	[0, 0.14]	[0, 0.13]
$\hat{M}_{OU1}$	0.05	0.1	0.34	0.33	0.1	0.1
HPD_{OU1}	[0, 0.14]	[0, 0.22]	[0.14, 0.53]	[0.13, 0.52]	[0, 0.23]	[0.01, 0.21]
$\hat{M}_{OU2}$	0.09	0.09	0.09	0.36	0.14	0.23
HPD_{OU2}	[0, 0.2]	[0, 0.21]	[0, 0.22]	[0.18, 0.54]	[0.02, 0.28]	[0.06, 0.38]
$\hat{M}_{Cos1}$	0.09	0.05	0.14	0.05	0.53	0.14
HPD_{Cos1}	[0, 0.21]	[0, 0.13]	[0, 0.28]	[0, 0.13]	[0.32, 0.72]	[0.02, 0.3]
$\hat{M}_{Cos2}$	0.14	0.1	0.05	0.1	0.19	0.43
HPD_{Cos2}	[0.02, 0.3]	[0, 0.21]	[0, 0.13]	[0, 0.22]	[0.05, 0.36]	[0.24, 0.64]

Table D.4: Per-condition proportion estimates (mean $\hat{M}$ and 95% HPD intervals) obtained via the Dirichlet-Multinomial model for the 1-point experiment.

	Lin_1	Lin_2	OU_1	OU_2	Cos_1	Cos_2
$\hat{M}_{Lin1}$	0.55	0.05	0.14	0.09	0.09	0.09
HPD_{Lin1}	[0.36, 0.74]	[0, 0.13]	[0.02, 0.28]	[0, 0.21]	[0, 0.21]	[0, 0.21]
$\hat{M}_{Lin2}$	0.18	0.55	0.09	0.09	0.05	0.04
HPD_{Lin2}	[0.05, 0.33]	[0.36, 0.75]	[0, 0.21]	[0.01, 0.22]	[0, 0.14]	[0, 0.13]
$\hat{M}_{OU1}$	0.05	0.1	0.34	0.33	0.1	0.1
HPD_{OU1}	[0,14]	[0, 0.22]	[0.14, 0.53]	[0.13, 0.52]	[0, 0.23]	[0.01, 0.21]
$\hat{M}_{OU2}$	0.09	0.09	0.09	0.36	0.14	0.23
HPD_{OU2}	[0, 0.2]	[0, 0.21]	[0, 0.22]	[0.18, 0.54]	[0.02, 0.28]	[0.06, 0.38]
$\hat{M}_{Cos1}$	0.09	0.05	0.14	0.05	0.53	0.14
HPD_{Cos1}	[0, 0.21]	[0, 0.13]	[0.01, 0.28]	[0, 0.13]	[0.32, 0.72]	[0.02, 0.3]
$\hat{M}_{Cos2}$	0.14	0.1	0.05	0.1	0.19	0.43
HPD_{Cos2}	[0.02, 0.3]	[0, 0.21]	[0, 0.13]	[0, 0.22]	[0.05, 0.36]	[0.24, 0.64]

Appendix E

Generalizing Function Compositions

Figure E.1: The main instruction screen introduced the task.

Figure E.2: In the example trial, participants received additional information on the task.

Figure E.3: As in the main task, participants had to select the offspring from four alternatives in the example trial.

Figure E.4: The main task included a set of guidelines that participants could hide to allow for more space on the screen.
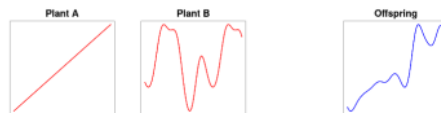
Figure E.5: As in the example trial, participants first received the example rule.
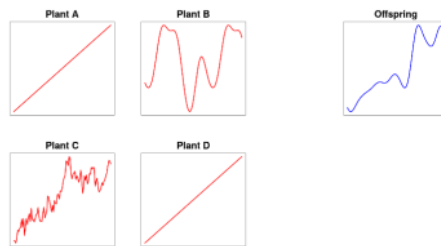
Figure E.6: Once they revealed the example rule, the parent plants of the test item were revealed.
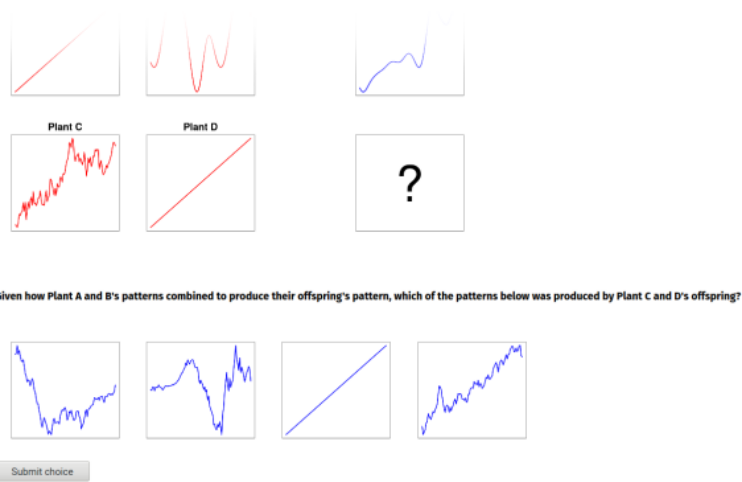
Figure E.7: Then, the four candidate patterns were shown, and the participant had to select the most likely offspring.

Appendix F

Transferring Function Compositions

F.1 Experiment
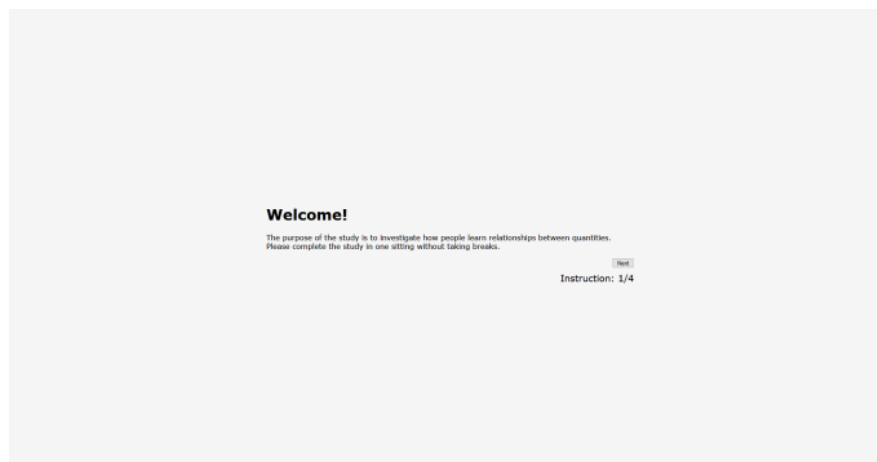
Figure F.1: The main instruction screen.
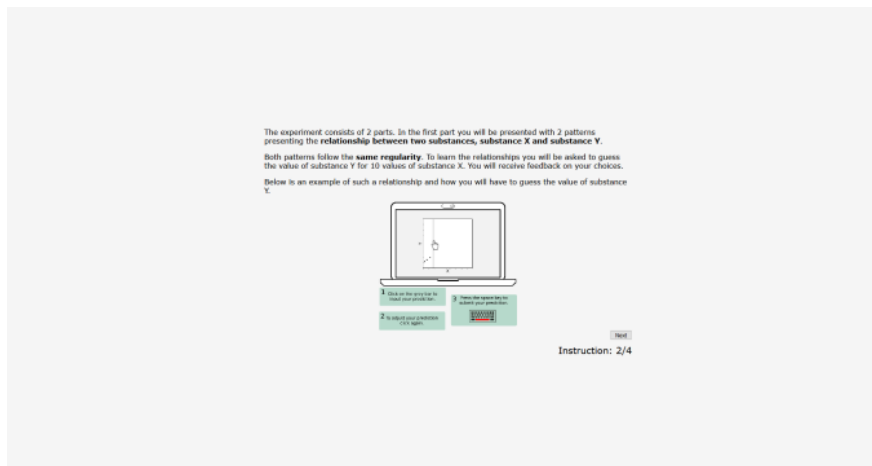
Figure F.2: After the general introduction, participants were shown an example of the experimental screen.

Figure F.3: Then, they were introduced to the experimental structure.

Figure F.4: The instructions also introduced the final, forced-choice task.
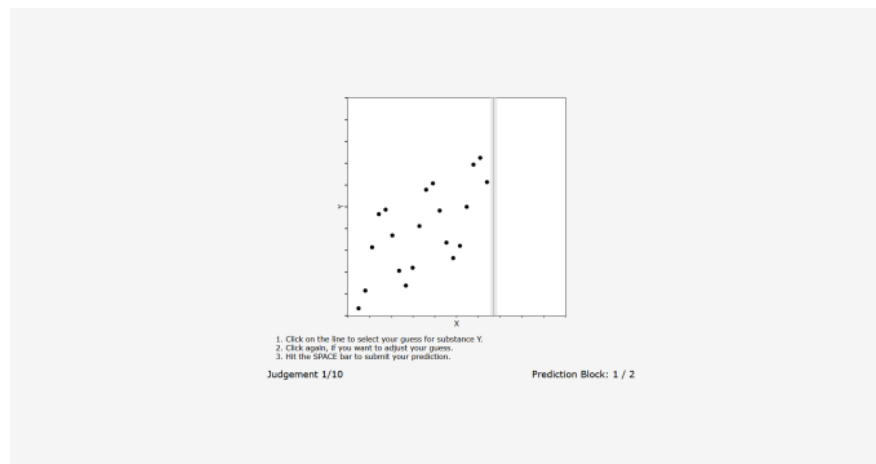
Figure F.5: In the two training blocks, participants had to predict the y values and received feedback.

Figure F.6: Before the transfer block, participants received additional instructions.
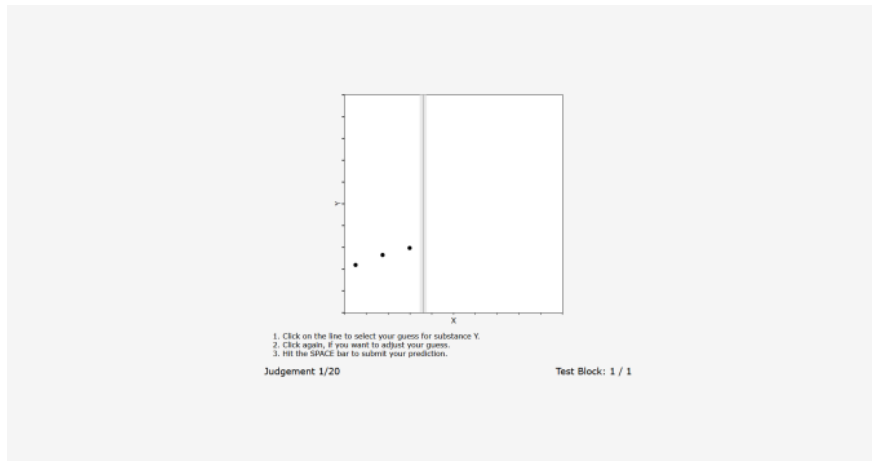
Figure F.7: In the transfer block, participants had to extrapolate the pattern given three points.

Figure F.8: After the transfer block, participants received instructions for the final forced-choice task.

Figure F.9: In the final block participants had to choose the most likely pattern for the transfer block.

F.2   Error Models

All models and analyses in this chapter were specified and fitted the same way as in Chapter 5 and Appendix D.

We compared the same models as in Chapter 5. For posterior predictive distributions for the log-normal model, see Figure F.10. For per-participant posterior predictive plots for the log-normal, see Figure F.10.
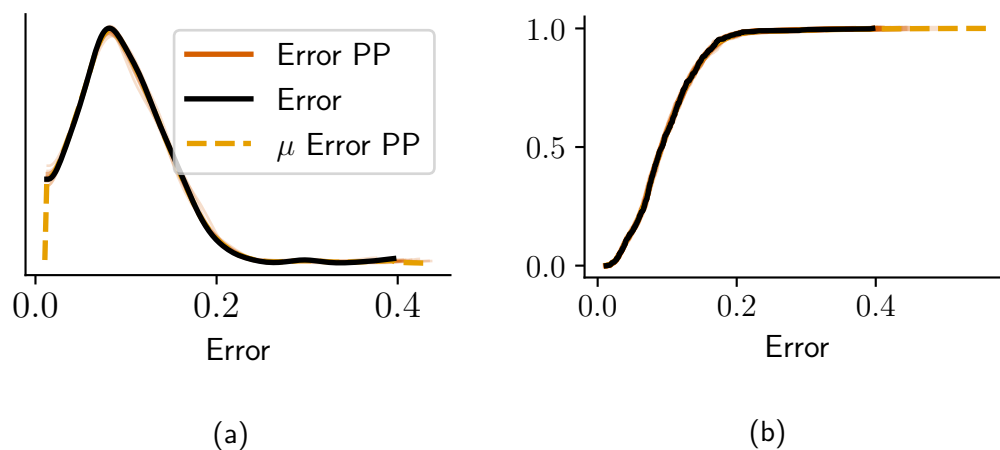
Figure F.10: Mean posterior predictive density (a) and cumulative density (b), (μ Error PP, dashed orange line) and 25 draws from the posterior (Error PP, solid orange line) for the log-normal hierarchical model for the composition experiment.

For group-level error estimates for the exponential model, see Figure F.12, and Table F.1. Per-participant posterior predictive plots for the exponential model are presented in Figure F.13.

F.2.1   Model Comparisons

As in the previous chapter, we compared the models via out-of-sample log-likelihoods and contrasted them via the WAIC. For a comparison of models, see Table F.2.
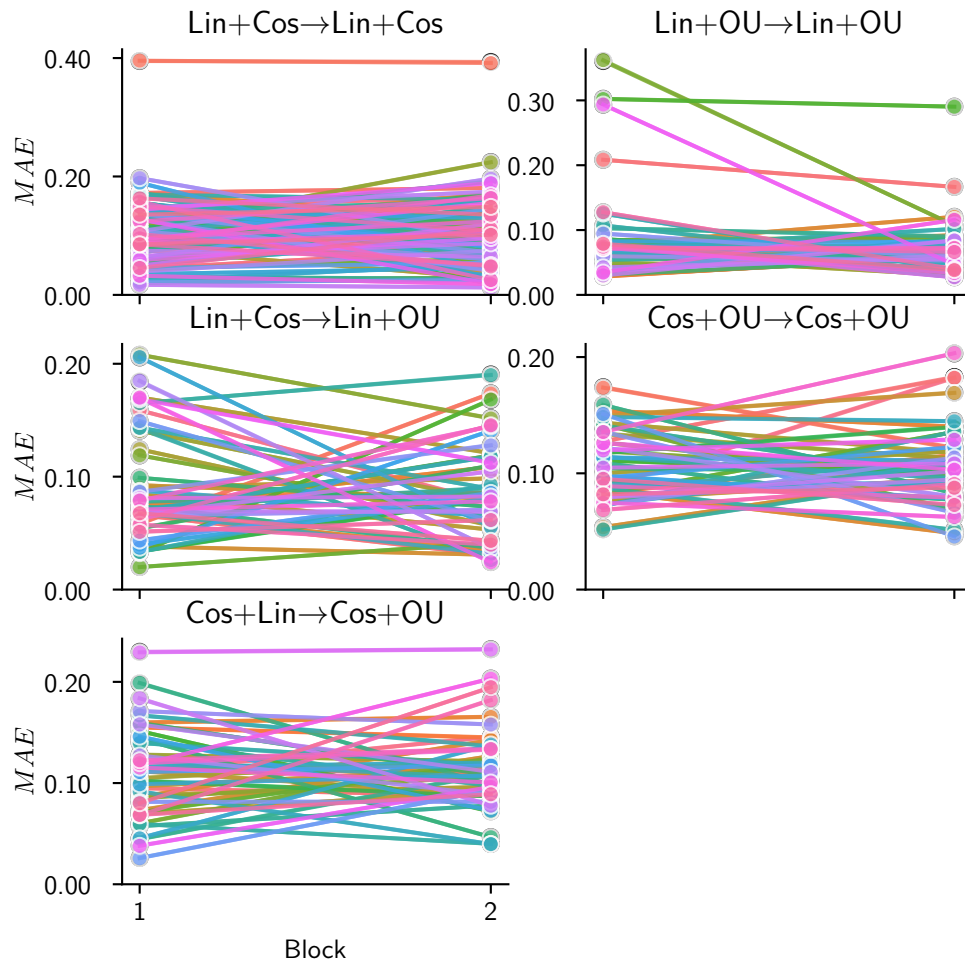
Figure F.11: Per-participant training errors (coloured markers and lines) and corresponding, per-participant mean of the posterior predictive distribution for the hierarchical exponential model (white markers and coloured lines).
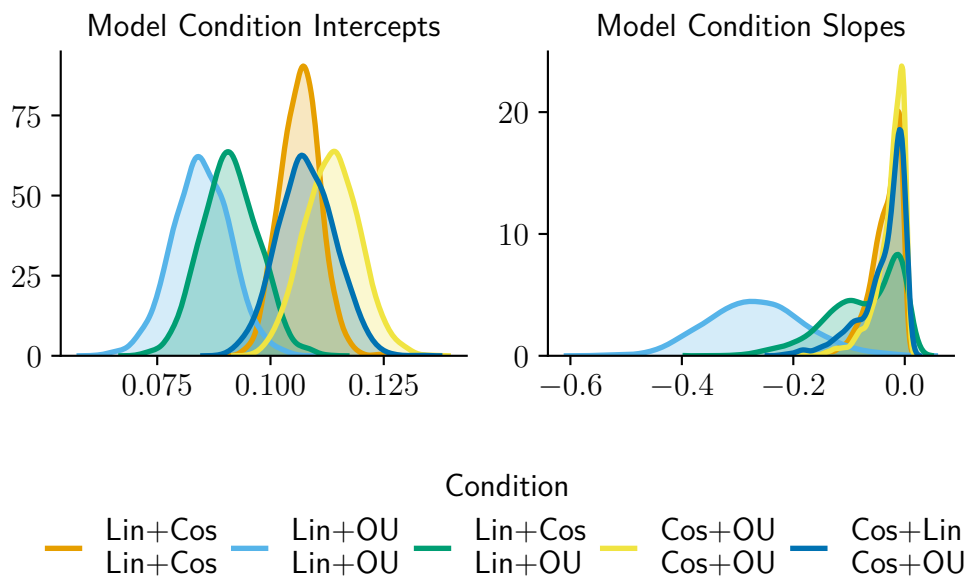
Figure F.12: Group-level estimates of error intercepts and slopes estimated via the hierarchical exponential-decay model.

(a)

(b)
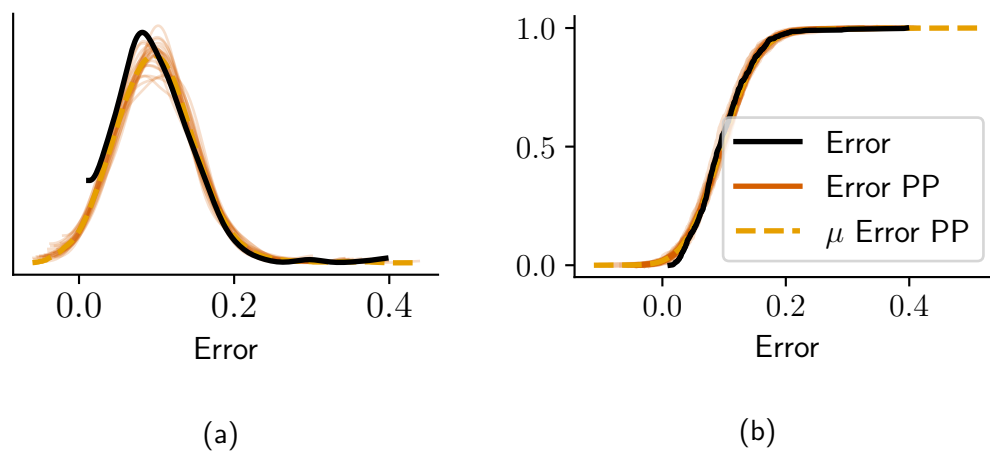
Figure F.13: Mean posterior predictive density (a) and cumulative density (b), (μ Error PP, dashed orange line) and 25 draws from the posterior (Error PP, solid orange line) for the exponential hierarchical model for the composition experiment.

|	$\hat{M}$	SD	$\mathrm{HPD}_{3\%}$	$\mathrm{HPD}_{97\%}$
$x_{k\mathrm{Lin}_1}$	0.02	0.00	0.02	0.03
$x_{k\mathrm{Lin}_2}$	0.02	0.00	0.02	0.03
$x_{k\mathrm{OU}_1}$	0.04	0.00	0.03	0.04
$x_{k\mathrm{OU}_2}$	0.09	0.00	0.08	0.09
$x_{k\mathrm{Cos}_1}$	0.04	0.00	0.03	0.05
$x_{k\mathrm{Cos}_2}$	0.10	0.01	0.09	0.11
$x_{d\mathrm{Lin}_1}$	0.23	0.13	0.03	0.46
$x_{d\mathrm{Lin}_2}$	0.21	0.11	0.03	0.40
$x_{d\mathrm{OU}_1}$	0.19	0.10	0.01	0.35
$x_{d\mathrm{OU}_2}$	0.02	0.01	0.00	0.04
$x_{d\mathrm{Cos}_1}$	0.27	0.08	0.11	0.42
$x_{d\mathrm{Cos}_2}$	0.26	0.04	0.18	0.33

Table F.1: Group-level estimated means $\hat{M}$ for intercepts, β_0 and slopes, β_1, as well as 95% highest-posterior densitz intervals estimated via MCMC for the exponential decay model.

	Rank	WAIC	$\mathrm{p_{WAIC}}$	$\mathrm{d_{WAIC}}$
LogNPP	0	1879.89	2346.33	0
Exponential	1	1045.21	163.08	834.68
Log	2	1024.86	10.49	855.03
Linear	3	971.50	13.48	908.39

Table F.2: The four models compared, their rank and WAIC score (higher values reflect better fits). $\mathrm{p_{WAIC}}$ are estimated number of parameters, and $\mathrm{d_{WAIC}}$ is the relative difference for WAIC scores and the best-fitting (rank 0) model.

F.3　Choice Model

As in Appendix D.3, we estimated the proportions for each option via a Dirichlet-Multinomial model. For the estimated mean proportions and HPDs, see Table F.3 and F.4.

$\hat{M}$	SD	HPD$_{2.5\%}$	HPD$_{97.5\%}$	Option	Condition
0.04	0.02	0.00	0.08	Cos	$2 \times Lin + Cos$
0.07	0.03	0.02	0.14	Cos	$2 \times Lin + Cos$
0.14	0.04	0.06	0.23	Lin+Cos	$2 \times Lin + Cos$
0.12	0.04	0.04	0.21	Lin+Cos	$2 \times Lin + Cos$
0.14	0.05	0.06	0.24	OU+Cos	$2 \times Lin + Cos$
0.14	0.04	0.06	0.23	OU+Cos	$2 \times Lin + Cos$
0.22	0.06	0.11	0.33	Lin	$2 \times Lin + OU$
0.30	0.06	0.18	0.42	Lin	$2 \times Lin + OU$
0.25	0.06	0.14	0.36	Lin+OU	$2 \times Lin + OU$
0.21	0.05	0.11	0.31	Lin+OU	$2 \times Lin + OU$
0.05	0.03	0.01	0.11	OU	$2 \times Lin + OU$
0.14	0.05	0.06	0.23	OU	$2 \times Lin + OU$
0.14	0.05	0.07	0.24	Cos	$Lin + Cos, Lin + OU$
0.12	0.04	0.05	0.21	Cos	$Lin + Cos, Lin + OU$
0.04	0.02	0.00	0.08	Lin+Cos	$Lin + Cos, Lin + OU$
0.16	0.05	0.06	0.26	Lin+Cos	$Lin + Cos, Lin + OU$
0.14	0.05	0.06	0.23	OU+Cos	$Lin + Cos, Lin + OU$
0.16	0.05	0.07	0.25	OU+Cos	$Lin + Cos, Lin + OU$
0.45	0.06	0.32	0.56	Lin	Lin Control
0.25	0.06	0.15	0.36	Lin	Lin Control
0.29	0.06	0.18	0.40	Lin+OU	Lin Control
0.25	0.06	0.14	0.35	Lin+OU	Lin Control
0.11	0.04	0.04	0.18	OU	Lin Control
0.09	0.04	0.02	0.16	OU	Lin Control

Table F.3: Proportion estimates (mean $\hat{M}$ and 95% HPD intervals obtained via the Dirichlet-Multinomial model for linear transfer set conditions and control condition.

$\hat{M}$	SD	HPD$_{2.5\%}$	HPD$_{97.5\%}$	Option	Condition
0.38	0.06	0.25	0.50	Cos	$2 \times Cos + Lin$
0.11	0.04	0.03	0.19	Lin	$2 \times Cos + Lin$
0.11	0.04	0.04	0.19	Lin+Cos	$2 \times Cos + Lin$
0.16	0.05	0.07	0.26	Lin+OU	$2 \times Cos + Lin$
0.09	0.04	0.03	0.16	OU	$2 \times Cos + Lin$
0.16	0.05	0.07	0.26	OU+Cos	$2 \times Cos + Lin$
0.20	0.05	0.09	0.30	Cos	$2 \times Cos + OU$
0.07	0.04	0.01	0.14	Lin	$2 \times Cos + OU$
0.21	0.05	0.12	0.32	Lin+Cos	$2 \times Cos + OU$
0.12	0.04	0.05	0.21	Lin+OU	$2 \times Cos + OU$
0.18	0.05	0.08	0.29	OU	$2 \times Cos + OU$
0.22	0.06	0.11	0.32	OU+Cos	$2 \times Cos + OU$
0.21	0.05	0.11	0.32	Cos	$Cos + Lin, Cos + OU$
0.07	0.03	0.01	0.13	Lin	$Cos + Lin, Cos + OU$
0.27	0.06	0.16	0.38	Lin+Cos	$Cos + Lin, Cos + OU$
0.12	0.04	0.05	0.21	Lin+OU	$Cos + Lin, Cos + OU$
0.11	0.04	0.04	0.19	OU	$Cos + Lin, Cos + OU$
0.22	0.05	0.11	0.31	OU+Cos	$Cos + Lin, Cos + OU$
0.36	0.06	0.24	0.48	Cos	Cos Control
0.16	0.05	0.08	0.25	Lin	Cos Control
0.16	0.05	0.06	0.26	Lin+Cos	Cos Control
0.11	0.04	0.03	0.18	Lin+OU	Cos Control
0.09	0.04	0.02	0.16	OU	Cos Control
0.12	0.04	0.05	0.20	OU+Cos	Cos Control

Table F.4: Proportion estimates (mean $\hat{M}$ and 95% HPD intervals obtained via the Dirichlet-Multinomial model for cosine transfer set conditions and control condition.

Bibliography

Alfonso-Reese, L. A., Ashby, F. G., and Brainard, D. H. (2002). What makes a categorization task difficult? *Perception & Psychophysics*, 64(4):570–583.

Ashby, F. G. and Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, 39(2):216–233.

Berry, D. C. and Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology Section A*, 36(2):209–231.

Blanchard, T., Lombrozo, T., and Nichols, S. (2018). Bayesian occam's razor is a razor of the people. *Cognitive science*, 42(4):1345–1359.

Bott, L. and Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1):38–50.

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review*, 77(6):546–556.

Bourne Jr, L. (1979). Stimulus-rule interaction in concept learning. *The American Journal of Psychology*, pages 3–17.

Brehmer, B. (1971). Effects of communication and feedback on cognitive conflict. *Scandinavian Journal of Psychology*, 12(1):205–216.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, 11(1):1–27.

Brehmer, B. (1976). Learning complex rules in probabilistic inference tasks. *Scandinavian Journal of Psychology*, 17(1):309–312.

Brehmer, B., Alm, H., and Warg, L.-E. (1985). Learning and hypothesis testing in probabilistic inference tasks. *Scandinavian Journal of Psychology*, 26(1):305–313.

Brown, M. and Lacroix, G. (2017). Underestimation in linear function learning: Anchoring to zero or xy similarity? *Canadian Journal of Experimental Psychology*, 71(4):274–282.

Brown, N. and Sandholm, T. (2018). Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424.

Bruner, J. S., Goodnow, J. J., and Austin, G. A. (1956). *A study of thinking.* John Wiley and Sons.

Busemeyer, J. R., Byun, E., Delosh, E. L., and McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In *Knowledge Concepts and Categories*. The MIT Press.

Busemeyer, J. R., Myung, I. J., and McDaniel, M. A. (1993). Cue competition effects: Empirical tests of adaptive network learning models. *Psychological Science*, 4(3):190–195.

Byun, E. (1995). *Interaction between prior knowledge and type of nonlinear relationship on function learning.* PhD thesis, Purdue University.

Carroll, J. D. (1963). *Functional learning: The learning of continuous functional mappings relating stimulus and response continua.* PhD thesis, Princeton.

Chomsky, N. and Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, 1(2):97–138.

Dechter, E., Malmaud, J., Adams, R. P., and Tenenbaum, J. B. (2013). Bootstrap learning via modular concept discovery. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

DeLosh, E. L., Busemeyer, J. R., and McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(4):968–986.

Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., and Zoubin, G. (2013). Structure Discovery in Nonparametric Regression through Compositional Kernel Search. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1166–1174, Atlanta, Georgia, USA.

Erickson, M. A. and Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, 127(2):107–140.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, 41(2):145–170.

French, R. M., Mareschal, D., Mermillod, M., and Quinn, P. C. (2004). The role of bottom-up processing in perceptual categorization by 3-to 4-month-old infants: Simulations and data. *Journal of Experimental Psychology: General*, 133(3):382–397.

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

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Gershman, S. (2017a). On the blessing of abstraction. *Quarterly journal of experimental psychology (2006)*, 70(3):361–365.

Gershman, S. J. (2017b). Predicting the past, remembering the future. *Current Opinion in Behavioral Sciences*, 17:7–13.

Gershman, S. J. and Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, 20(2):251–256.

Gick, M. L. and Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12:306–355.

Gilden, D. L., Thornton, T., and Mallon, M. W. (1995). $1/f$ noise in human cognition. *Science*, 267(5205):1837–1839.

Goodman, N. (1983). *Fact, fiction, and forecast.* Harvard University Press.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., and Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., and Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, 114(30):7892–7899.

Gopnik, A. and Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085.

Griffiths, T. L. (2017). Formalizing prior knowledge in causal induction. *The Oxford Handbook of Causal Reasoning*, pages 115–126.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364.

Griffiths, T. L., Lucas, C., Williams, J., and Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 553–560.

Griffiths, T. L. and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17(9):767–773.

Hahn, U., Bailey, T. M., and Elvin, L. B. (2005). Effects of category diversity on learning, memory, and generalization. *Memory & cognition*, 33(2):289–302.

Hamrick, J. B., Ballard, A. J., Pascanu, R., Vinyals, O., Heess, N., and Battaglia, P. W. (2017). Metacontrol for adaptive imagination-based optimization. *arXiv preprint arXiv:1705.02670*.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, 56(1):51–65.

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K., and Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic bulletin & review*, 26(3):1043–1050.

He, D., Dushoff, J., Day, T., Ma, J., and Earn, D. J. (2013). Inferring the causes of the three waves of the 1918 influenza pandemic in England and Wales. *Proceedings of the Royal Society B: Biological Sciences*, 280(1766):20131345.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Jäkel, F., Schölkopf, B., and Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, 15(2):256–271.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604.

Jern, A. and Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive psychology*, 66(1):85–125.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in neural information processing systems*, pages 641–648.

Kalish, M. L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition*, 41(6):886–896.

Kalish, M. L., Griffiths, T. L., and Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2):288–294.

Kalish, M. L., Lewandowsky, S., and Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, 111(4):1072–1099.

Kang, S. H., McDaniel, M. A., and Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, 18(5):998–1005.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, 119(4):685–722.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10(3):307–321.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*.

Ketchum, R. D. and Bourne Jr, L. (1980). Stimulus-rule interactions in concept verification. *The American journal of psychology*, pages 5–23.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280.

Kwantes, P. J. and Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5):1019.

Kwantes, P. J., Neal, A., and Kalish, M. (2012). Item order matters in a function learning task. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 66(2):90.

Lake, B., Salakhutdinov, R., and Tenenbaum, J. (2012). Concept learning as motor program induction: A large-scale empirical study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Lake, B. M., Lawrence, N. D., and Tenenbaum, J. B. (2018). The emergence of organizing structure in conceptual representation. *Cognitive science*.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2019). The Omniglot challenge: A 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

Little, D. R. and Shiffrin, R. (2009). Simplicity bias in the estimation of causal functions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Little, J. L. and McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & cognition*, 43(2):283–297.

Lucas, C. G., Bridgers, S., Griffiths, T. L., and Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, 131(2):284–299.

Lucas, C. G. and Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, 34(1):113–147.

Lucas, C. G., Griffiths, T. L., Williams, J. J., and Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, 22(5):1193–1215.

Lucas, C. G., Sterling, D., and Kemp, C. (2012). Superspace extrapolation reveals inductive biases in function learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, number 34 in 34.

MacKay, D. (2003). *Information theory, inference and learning algorithms.* Cambridge University Press.

Martin, J. B., Griffiths, T. L., and Sanborn, A. N. (2012). Testing the efficiency of Markov chain Monte Carlo with People using facial affect categories. *Cognitive science*, 36(1):150–162.

Mathy, F. and Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review*, 16(6):1050–1057.

McDaniel, M. A. and Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, 12(1):24–42.

Mervis, C. B., Mervis, C. A., Johnson, K. E., and Bertrand, J. (1992). Studying early lexical development: The value of the systematic diary method. *Advances in infancy research.*

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518:529–533.

National Geographic (2020). How some cities flattened the curve during the 1918 flu pandemic. `"http://www.nationalgeographic.com/history/2020/03/how-cities-flattened-curve-1918-spanish-flu-pandemic-coronavirus/"`. Last accessed: 09.06.2020.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., and Shafir, E. (1990). Category-based induction. *Psychological Review*, 97(2):185–200.

Pachur, T. and Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, 65(2):207–240.

Piantadosi, S. T. and Jacobs, R. A. (2016). Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science*, 25(1):54–59.

Piantadosi, S. T., Tenenbaum, J. B., and Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123(4):392–424.

Pothos, E. M. and Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive science*, 26(3):303–343.

Quiroga, F., Schulz, E., Speekenbrink, M., and Harvey, N. (2018). Structured priors in human forecasting. *BioRxiv*, page 285668.

Ramlee, F., Sanborn, A. N., and Tang, N. K. (2017). What sways people's judgment of sleep quality? A quantitative choice-making study with good and poor sleepers. *Sleep.*

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.

Reimers, S. and Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting*, 27(4):1196–1214.

Reznick, J. S., Ketchum, R. D., and Bourne, L. E. (1978). Rule-specific dimensional interaction effects in concept learning. *Bulletin of the Psychonomic Society*, 12(4):314–316.

Roberts, G., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.

Salatas, H. and Bourne, L. (1974). Learning conceptual rules: III. Processes contributing to rule difficulty. *Memory & Cognition*, 2(3):549–553.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55.

Samuelson, L. K. and Smith, L. B. (2005). They call it like they see it: Spontaneous naming and attention to shape. *Developmental Science*, 8(2):182–198.

Sanborn, A., Griffiths, T., and Navarro, D. (2006). A more rational model of categorization. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, number 28 in 28.

Sanborn, A. N., Griffiths, T. L., and Shiffrin, R. M. (2010). Uncovering mental representations with Markov Chain Monte Carlo. *Cognitive Psychology*, 60(2):63–106.

Schulz, E., Speekenbrink, M., and Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., and Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, 99:44–79.

Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M., and Gershman, S. (2015). Assessing the perceived predictability of functions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, number 37 in 37.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.

Shepard, R. N., Hovland, C. I., and Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, 75(13):1.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484.

Stojic, H., Eldar, E., Hassan, B., Dayan, P., and Dolan, R. J. (2018). Are you sure about that? On the origins of confidence in concept learning. In *Proceedings of the Annual Conference on Cognitive Computational Neuroscience*. Citeseer.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In *Advances in neural information processing systems*, pages 59–68.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Theocharis, Z., Smith, L. A., and Harvey, N. (2019). The influence of graphical format on judgmental forecasting accuracy: Lines versus points. *Futures & Foresight Science*, 1(1):e7.

Ullman, T., Stuhlmüller, A., Goodman, N., and Tenenbaum, J. B. (2014). Learning physics from dynamical scenes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, number 36 in 36, pages 1640–1645.

Ullman, T. D., Siegel, M., Tenenbaum, J. B., and Gershman, S. J. (2016). Coalescing the vapors of human experience into a viable and meaningful comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 38. Cognitive Science Society.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *arXiv preprint arXiv:1903.08008*.

von Helversen, B. and Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4):867–889.

Vul, E., Goodman, N., Griffiths, T. L., and Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4):599–637.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science*, 31(2):233–256.

Washington Post (2020). The White Houses self-serving approach to estimating the deadliness of the coronavirus. `"http://www.washingtonpost.com/politics/2020/05/05/white-houses-self-serving-approach-estimating-deadliness-coronavirus/"`. Last accessed: 09.06.2020.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

Wilson, A. G., Dann, C., Lucas, C. G., and Xing, E. P. (2015). The human kernel. In *Advances in neural information processing systems*, pages 2854–2862.

Wilson, R. C. and Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in Human Neuroscience*, 5:189.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., and Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2(12):915.

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review*, 126(6):841–864.

Xu, F., Dewar, K., and Perfors, A. (2009). Induction, overhypotheses, and the shape bias: Some arguments and evidence for rational constructivism. *The origins of object knowledge*, pages 263–284.