UNIVERSITY OF CALIFORNIA,
IRVINE

Modeling the Dynamics of Category Learning

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Sciences

by

Jesus Manuel Villarreal Ulloa

Dissertation Committee:
Professor Michael D. Lee, Chair
Professor Aaron M. Bornstein
Professor Mark Steyvers

2024

DEDICATION

To my family and friends. As a very special person once told me, our success is not only ours, it's the result of the support of the people in our lives.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to thank my advisor Michael D. Lee, thank you for all the support in this work and during my stay here. Thank you for listening to my crazy ideas and for having the patience to help me focus. If I'm ever a good scientist, it will be thanks to what you have taught me these past five years. To the members of my committee, Aaron Bornstein and Mark Styvers, if this work has improved over the last few years, it is in part thanks to your insightful comments. Thank you for taking the time to read this work.

There are a lot of people that without their support this would not have been possible. First and foremost to my mom Odilia Ulloa Padilla, thank you for all the patience and love that you have given me throughout the years, without you I would not be who I am today. To my father Juan Manuel Villarreal Trujillo for all your patience and love, I know that things were not always easy but you have always tried to be there for me. To Miguel Alonso Raya, thank you for always being there for me and my mom, without you everything would have been more difficult. To my cousin David Ricardo Guzmán Ulloa, thank you for always being there when our family needed you. It would take too long to thank every member of my family that has been there for me over the last five years, but I would like you all to know how much I appreciate your words and encouragement. I love you all.

To Talía Vianney Román López. We have been together for a long time, and there are many things that I could thank you for. I know that these past few years have been complicated, being apart is not easy, however, you were always there when I needed you the most. Thank you for all your patience, support, and care. Our conversations have helped me get through some difficult times, times when I felt lost or scared, and for that I will always be grateful. I hope we can continue to share more adventures together until we have finally filled that book.

To my friends Angela Shen and Prawit Sriwat (yes, you're on the same line), for all the laughs, noodles and support these past two and a half years, things would have been more difficult without the two of you. To Alex Etz, for all your help during the first few years, you helped make the grad school transition easier. To Nora Harhen and Nidhi Banavar, for trying to convince me that there is more to do in life than just work. To Adriana Felisa Chávez de la Peña, Jaime Islas Farias, José Luis Baroja Manzano, and Elena Villalobos Nolasco having you all here made things easier and also made Irvine feel more like home. To Joachim Vandekerckhove, thank you for all your support, you made those first years more fun and easier to navigate. To Mahbod Mehrvarz, for the smoking times where we could talk about science and life. To Alejandro Segura and Nataly Yáñes, I know that we don't get to see each other as often as we would like to but every time we are together it feels like no time has passed, thank you for all of your support.

Finally to the the Lee lab, Michael Lee and Helen Braithwaite, thank you both for all that you have done these last five years. I know that my transition to a new country wouldn't have been this easy without the two of you. I want you to know that I consider both of you dear friends and I will always cherish my memories with you... even that painful first

summer in Amsterdam. To Holly Westfall, thank you for your friendship and being like an older sister, I will always remember all the times we spent avoiding work, laughing, and throwing balls around the lab. To Lauren ELIZABETH Montgomery, thank you friend, I'm sure that what I've been able to do in this program is in part because of all that you've done. Thank you for listening to my ramblings about life and work. I bet you that this would not have been as easy without you here!

VITA

Jesus Manuel Villarreal Ulloa

EDUCATION

Doctor of Philosophy in Cognitive Sciences **2024**
University of California, Irvine *Irvine, CA*

Master of Science in Statistics **2022**
University of California, Irvine *Irvine, CA*

Specialization in Applied Statistics **2018**
Universidad Nacional Autónoma de México *CDMX, MX*

Bachelor of Science Psychology **2017**
Universidad Nacional Autónoma de México *CDMX, MX*

RESEARCH EXPERIENCE

Graduate Research Assistant **2019–2024**
University of California, Irvine *Irvine, CA*

Undergraduate Research Assistant **2011–2018**
Universidad Nacional Autónoma de México *CDMX, MX*

TEACHING EXPERIENCE

Teaching Assistant **2019–2024**
University of California, Irvine *Irvine, CA*

Professor of Instruction: Probability and Statistics for Psychology III. **Fall–2022**
University of California, Irvine *Irvine, CA*

Teaching Assistant **2012–2018**
Universidad Nacional Autónoma de México *CDMX, MX*

REFEREED JOURNAL PUBLICATIONS

A Coupled Hidden Markov Model Framework for Measuring the Dynamics of Categorization. **in review**

Bayesian graphical modeling with the circular drift diffusion model **2023**
Computational Brain and Behavior

Evaluating the complexity and falsifiability of psychological models **2023**
Psychological Review

Adaptive design optimization for a mnemonic similarity task **2022**
Journal of Mathematical Psychology

Velocity Estimation in Reinforcement Learning **2019**
Computational Brain and Behavior

Bayesian methods applied to the generalized matching law **2019**
Journal of the Experimental Analysis of Behavior

REFEREED CONFERENCE PUBLICATIONS

Categorization in Environments that Change when People Learn **July 2022**
Proceedings of the 44th Annual Conference of the Cognitive Science Society

BOOK CHAPTERS

Models in Strategic Interaction: Game Theory **In Press**
A Practical Introduction to Quantitative Models in Behaviour and Cognition

AWARDS & HONORS

Indow Fellowship for Research Excellence **2023**
University of California, Irvine

John I. Yellott Scholar Award Honorable Mention **2020**
University of California, Irvine

UC-Mexus CONACYT: Doctoral fellowship: 739644 **2019**
CONHACYT

SERVICE

Journal of Mathematical Psychology
Reviewer

Proceedings of the 41st Annual Meeting of the Cognitive Science Society
Reviewer

Journal of Judgement and Decision Making
Reviewer

ABSTRACT OF THE DISSERTATION

Modeling the Dynamics of Category Learning

By

Jesus Manuel Villarreal Ulloa

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2024

Professor Michael D. Lee, Chair

In order to organize the information that people encounter in their environments they have to develop the ability to organize their experience into categories. Category learning has been studied using both stable environments in which the underlying categories do not change, and in dynamic environments in which the underlying category structure changes independent of people's behavior. In many natural environments, however, category structures can change as a consequence of people's behavior. In order to study how people learn categories that are dynamically coupled to their behavior, we conduct two experiments in which the underlying category structure changed as people became more accurate. Results from these experiments show that people can quickly adapt to changes in an underlying category structure, and that complicated learning dynamics occur as they transition from one structure to another.

We argue that existing models of category learning are not well suited to capturing people's behavior in these tasks, and introduce a new framework to measure the dynamics of categorization. This approach is based on Coupled Hidden Markov Models (CHMMs), and allows us to directly model the effect of stimulus similarity in the categorical associations of stimuli across trials in a task. Using data from a previously published category learning study, we show that the CHMM can adequately describe people's categorization behavior, and serves as an interpretable and useful measurement model for understanding how people learn to associate stimuli with categories over time.

We further test the CHMM as a predictive model with out-of-sample generalization tests, using data from two previous experiments. One experiment involves people learning to categorize faces in terms of different category structures including gender, hair color, and trustworthiness. We show that the CHMM performs progressively better for more abstract categories, and often outperforms the well-established Generalized Context Model. The second experiment involves simple perceptual stimuli learning in a training-transfer design. We show that the CHMM again matches or outperforms the Generalized Context Model in predictive accuracy for transfer stimuli. We argue that the key advantage of the CHMM is the flexibility that motivated its development as a measurement model. While the GCM assumes people always know the true category structure, the CHMM has the flexibility to infer that some people at some stages of learning have incorrect category structures. We conclude with a discussion of future avenues for experimentation and model development that will improve our understanding of how people learn to make categorization decisions in a changing environment.

INTRODUCTION

Imagine that you have just received an email. Someone is letting you know that a distant relative has passed and you have inherited a large sum of money. The only thing you need to do is respond to the email with some personal information and everything will be arranged for you. Regrettably this is likely to be a scam, and an old one at that. Some of its features, such as the “large sum of money” and the “distant relative” make it easy for us to recognize it and categorize it as spam. Now imagine that there is a second email in your inbox. This time it has the name of a person you know. It states that they need your help and that you should contact them as soon as possible. The problem is the same. Should we categorize this email as a scam and just continue with our day? Or is this email real?

The problem of deciding if an email is legitimate or a scam highlights an important property of categorization in natural environments. When we have to make this decision, the features that we need to pay attention to may be replaced, changed, or gain more weight depending on the dynamics of the environment. In our example, the main features that made the first scam easy to spot are no longer part of the second one, and instead we need to look for new ones. Furthermore, these features could have changed as a consequence of our behavior. In this case, scam emails have changed because old ones were very easy to recognize either by the receiver or by an algorithm. The scammers needed to develop new approaches to meet their goals. In order to be able to categorize a stimulus correctly in this type of adversarial environment, humans need to be able to detect when something has changed and adapt how they weigh the information present in the stimulus.

In contrast with the problem in our motivating example, category learning studies typically involve stimulus-category relations that are stable across time (e.g., Feldman, 2000; Shepard et al., 1961; Smith & Minda, 2000). When naturally occurring stimuli are involved, these types of studies can help us understand how participants learn about a stable concept, such as how rocks are assigned to categories like basalt or granite (Nosofsky et al., 2018). When artificial stimuli are used, they can help us understand how participants learn to weigh different features as they gain more experience with a task (Kruschke, 1993a). These approaches have furthered our understanding on how people learn to categorize objects and generalize that knowledge to novel items.

Other category learning studies have instead focused on how people learn to make decisions in environments that change (e.g. Estes, 1984; Navarro et al., 2013; Speekenbrink & Shanks, 2010). These studies have shown that participants are able to track how the environment changes (Gallistel et al., 2014), and to adapt the weight they assign to specific features of a stimulus after a shift in a category structure (Kruschke, 1996). However, these types of experiments share some common shortcomings. Firstly, they use artificial stimuli which are built from only a couple of features that are then combined in order to construct artificial categories. This kind of category structures can be difficult to learn (Smith et al., 2011). Secondly, the dynamics of the environment are typically probabilistic or fixed prior to the start of the experiment. In other words, they are independent from the ability of participants to categorize the stimuli they encounter.

The study of category learning in stable environments, or in environment where the dynamics of the environment are decoupled from people's behavior, has allowed us to develop cognitive models that try to account for the dynamics of categorization. However, as the introductory example shows, not all environments behave in this way. It is not uncommon for these changes to be correlated with the decisions that we make. For example, in the human-constructed world, a filter that solves the phishing scam problem, motivates scammers to come up with new ways to evade those filters. Therefore, if we want to be able to correctly categorize a new email as spam, the filters we use have to adapt to the changes (El Kouari et al., 2020).

In Chapter 1 we study how people make categorization decisions in environments that change as they become more accurate. We present the results of two experiments in which the ability of people to correctly make categorization decisions modifies an underlying category structure. In the first experiment, we present participants with real-world animal images that participants have to categorize as being susceptible to a given disease or not. Results from this experiment show that people are able to adapt to changes in an underlying category structure that depend on the accuracy of their responses. Furthermore, the results suggest that the base-rate of the categories might play an important role in how participants adapt to a change. In a second experiment we introduce two new stimulus domains—involving travel items that may or may not be banned from plane travel, and produce that may be in season or out of season—each with their own dynamics. Similar to experiment one, the results show that the base rate of a category plays an important role in the adaptation process and that participants can learn to categorize items after a change even in situations when they only encounter a stimulus once.

We model the results of the first experiment using the ALCOVE model (Kruschke, 1992) of category learning. The results show that a model that allows forgetting for a subset of the stimuli in the task can account for participants adaptation. However, this assumption was tailored to details of the experimental design, and seems unlikely to be successful as a general model of the learning process in dynamic environments.

Motivated by the experimental results and the limitations of existing models, in Chapter 2 we introduce a new framework for measuring the dynamics of categorization. This new framework is based on a statistical method known as Coupled Hidden Markov Models (CHMM).

The key intuition behind this approach is that we can understand categorization as the result of two processes, the categorical representation of stimuli that we cannot observe, and the decisions that a participant makes trial by trial that we can observe. We develop and apply this new model to data from an experiment reported by Lee & Navarro (2002) which presents participants with artificial stimuli. The results show that the CHMM approach can accurately describe people's classification behavior in the experiment. Additionally, we show that some of the inferences are sensitive to the base rate of the categories, which was a key experimental result from Chapter 1.

One of the main problems with the application and evaluation of the CHMM approach is that its flexibility makes it difficult to evaluate without prediction and generalization tests based on out-of-sample data. In order to evaluate the CHMM approach, in Chapter 3 we apply this testing approach to two experiments reported previously on the literature. The first previous experiment was reported by Navarro et al. (2005), in which participants were presented with real-world face stimuli under four different category structures. The CHMM model is shown to have a high posterior adequacy, meaning that it is able to correctly account for the majority of the participant's label choices. To test the predictive accuracy of the model in this experiment we remove four stimuli from the inference process and show that as the category structure becomes harder to learn, the CHMM approach has a predictive accuracy that rivals the Generalized Context Model (GCM) of categorization (Nosofsky, 1988).

The second previous experiment was reported by Bartlema et al. (2014). In this experiment participants had to categorize artificial stimuli defined by two continuous features. The experimental design used learning and transfer blocks, where a subset of the stimuli were only presented during transfer. These transfer blocks allow us to test the predictive accuracy of the CHMM approach without artificially removing a subset of the items as was the case in the previous experiment. Once again, the results show that the predictive accuracy of the

CHMM is on par with the GCM. Results from these two experiments highlight some of the strengths and weaknesses of the CHMM approach and offer some interesting avenues for its future development.

Chapter 1

Categorization in Environments that Change when People Learn

Abstract

Most studies of human category learning involve category structures that do not change, or that change in a way that is independent of people's categorization behavior. We consider two experiments in which successful category learning causes categories to change. In experiment one, participants learned from feedback whether animals are healthy or diseased. Once their categorization accuracy was near-perfect, the category structure changed so that different animals became diseased. Based on exploratory data analysis and the application of two category learning models, we argue that, once they detect a category change, people retain what they have learned about healthy animals, but reset what they have learned about diseased animals. In experiment two, participants learned from feedback whether stimuli from three different domains belonged to one of two mutually exclusive categories. Each domain varied in terms of the base rate of their categories and the dynamics that control the

within domain transitions between category structures. Exploratory data analysis suggests that both variables have an effect in the patterns of leaning observed in aggregate behavior. We discuss future modeling goals and emphasize the need for learning models to study situations in which people's behavior impacts the dynamics of the environment in which learning takes place.

1.1 Introduction

Most studies of human category learning involve fixed categories (e.g., Feldman, 2000; Shepard et al., 1961; Smith & Minda, 2000). This is appropriate for understanding how people learn about stable concepts. It is reasonable to assume that many natural kinds —fruits, insects, weapons, and so on—have stable relationships between stimuli and categories. For example, the assignment of rocks to categories like obsidian, basalt, and granite involves a stable category structure (Nosofsky et al., 2018). The assignment of colors to categories like red, blue, and yellow involves cultural differences in the available categories and assignments (Regier et al., 2007), but those structures are largely stable within a culture.

Some category learning studies use more dynamic environments, in which categories change over time. The change could be a sudden reassignment of stimuli to different categories, or a gradual drift in the probability that stimuli belong to categories (e.g. Estes, 1984; Gallistel et al., 2001; Navarro et al., 2013; Speekenbrink & Shanks, 2010; Kruschke, 1996). These tasks are appropriate for understanding how people adapt to new category structures and non-stationary environments. Most of these previous studies determine the dynamics of environmental change ahead of time, and assume that change is independent of participant behavior. Category learning studies rarely consider dynamic environments that change *in* *response to* people's decisions.[1] Assuming that category learning is independent of category structure may be appropriate in some situations, at least as an approximation. For example, starting from house telephones, the technological development that led to the sudden introduction of car phones, then mobile phones, and then smartphones has required people to change how they categorize stimuli as phones. This learning process, however, has not influenced technological development. As another example, the seasons drift cyclically largely independent of the categories people learn. This means that people adapting their categorization from Finland being an undesirable vacation destination in winter to a desirable one in summer does not influence the weather in Helsinki.

These examples hint, however, at the limits of the independence assumption. People's ability to learn to use new devices as phones creates longer-term markets for technological development. Similarly, the independence of people's behavior from temperature fluctuation only holds for those categories and time scales that do not involve human-influenced global warming. It is generally not the case that the dynamics of an environment are completely decouple from people's learning and behavior in that environment.

Accordingly, it is not hard to identify real-world category learning situations in which people's learning and environmental dynamics are tightly coupled, with changes in categorization behavior leading to changes in the category structures being learned. In the natural world, one cause of virus mutation, along with copying error and select cell pressure, is a change in the immunity of potential hosts (Sugak et al., 2015). This means that as society develops better treatments the virus environment changes. Loosely speaking, as the categorization problems involved in providing immunity—correctly classifying treatments as effective or not effective—is solved, the categorization problem itself changes as a consequence. In the

[1] Perhaps the closest example is the Wisconsin Card Sorting Test (WCST: Dehaene & Changeux, 1991; D'Alessandro et al., 2020), in which participants have to organize a set of cards base on an underlying rule. As performance improves the rule can change. The main difference is that rules in the WCST are typically based on a single stimulus dimension, such as color or shape. In general, the relationship between stimuli and categories is more complicated than a single dimension.

human-constructed world, phishing scams continually need to adapt to evade spam filters (El Kouari et al., 2020). The spam filters solve a categorization problem to separate email stimuli into legitimate and blocked categories. As the accuracy of filters improves, the nature of the categorization problem changes, with new phishing attacks developed.

In this chapter we present results from two experiments aiming at studying people's category learning behavior in tasks where the category structure being leaned changes when people become sufficiently accurate. In experiment one, people are asked over a sequence of trials to categorize animals as healthy or infected with some disease, based on feedback provided after every trial. Once they reach a high level of accuracy, the category structure changes, so that a different set of animals become diseased. The environmental change is not signaled other than through the change in feedback for specific animals on individual trials. We are interested in how people perform in this learning situation, for which environmental dynamics are linked to their category learning.

In experiment two, people are asked to categorize three different sets of stimuli in a randomized block design. Each block corresponds to a different stimulus domain. The first again asked people to categorize animals as healthy or infected with some disease. However, we modified the four underlying category structures to control for the base rate of each category. The second stimulus asked to classify every-day objects as allowed or prohibited for commercial flight. The category structures in this domain were defined by published Transport Security Association (TSA) classifications in four different years. The third stimulus domain asked people to categorize fruit and vegetable produce as being in season or out of season, the category structures were defined by the four seasons. The main objectives of experiment two were to control for the base rate of a category, contrast participants learning behavior across different stimulus domains, and to compare participants learning behavior under environments that change when people become more accurate with environments that change independently of people's behavior.

The remainder of this chapter is organized as follows. First we introduce experiment one and present the behavioral results, before presenting the results of experiment two along some comparisons to the behavioral patterns observed in experiment one. We continue by presenting two category learning models and comparing their performance on data obtained for experiment one. We end this chapter by discussing some of the model's limitations and future directions for the study of categorization behavior in changing environments.

1.2 Experiment One

1.2.1 Method

Participants

38 undergraduate student participants at the University of California, Irvine completed the experiment for course credit.

Stimuli

The stimuli were images of 21 animals divided into two mutually exclusive categories, "healthy" or "infected", depending on four category structures. These category structures corresponded to four real-world diseases: cryptococcosis, foot and mouth, lentivirus, and anthrax. Figure 1.1 shows the set of stimuli, and their assignment to the healthy and diseased categories for all four diseases. The animals are represented as points using non-metric multidimensional scaling as a visualization method (Kruskal, 1964), based on similarity data reported by Westfall & Lee (2021). The animals category membership is represented by the colored regions for each of the diseases.

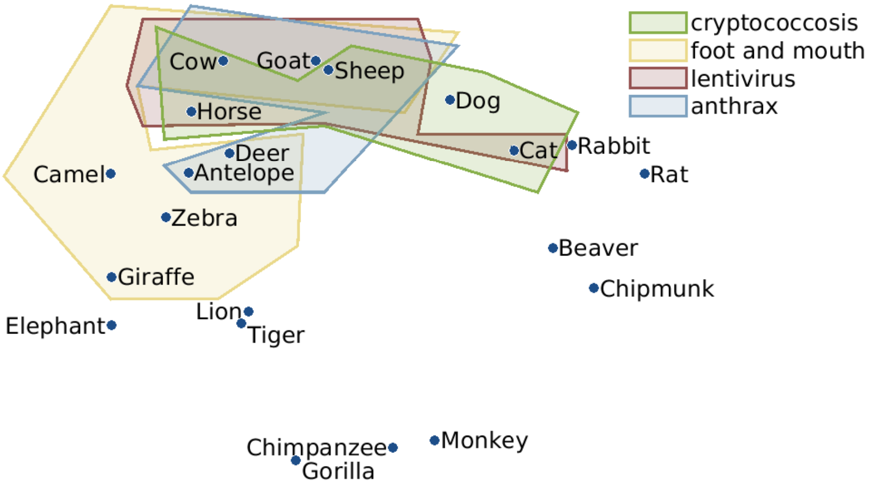

Figure 1.1: The four category structures. The 21 animal stimuli are represented as points arranged so that more semantically similar animals are located nearer each other. The four diseases are represented by colored regions that encompass those animals that have the disease.

It is clear there is considerable overlap between the four category structures and the differences between them are subtle. Sheep and cow belong to the diseased category for all of the diseases, horse and deer belong to the diseased category for exactly half the diseases, and a large number of animals always belong to the healthy category. The category structures vary between five and eight diseased animals, so disease is always the lower base-rate category.

Procedure

All participants completed 210 categorization trials. On each trial, an animal was presented as a picture with an accompanying text label. The same picture was used every time that animal was presented. Participants were required to categorize the animal as “healthy” or “diseased”. They then received feedback of the form “wrong, the horse is diseased”, informing them whether their response was correct and making explicit the correct classification. At the top of the interface a set of 210 slots was shown, corresponding to the 210 trials. Completed correct responses were shown as black circles, completed incorrect responses were shown as crosses, and trials yet to be completed were shown as gray circles. This information was updated after every trial.

If, at any point in the sequence of trials, the participant had correctly categorized 18 or more out of the last 21 animals, the category changed. The study information sheet told participants that “It is possible that whether or not a particular animal is healthy could change over the course of the experiment,” but a change in category structure was not indicated in any way during the experiment. The animals were presented in a random order, subject to the constraint that no animal be presented twice within the same disease category until all other animals had been presented. The experiment was completed after 210 trials, regardless of the participant’s accuracy.

Three different sequences of transitions from one category structure to the next were used. We refer to these sequences as conditions. Condition 1 started with anthrax, followed by lentivirus, cryptococcosis, and foot and mouth. Condition 2 started with anthrax, followed by foot and mouth, cryptococcosis, and lentivirus. Condition 3 started with foot and mouth, followed by lentivirus, cryptococcosis, and anthrax. A total of 14, 13, and 11 participants completed conditions 1, 2, and 3 respectively, in a between-participants design. These sequences were intended to allow comparisons that focus on specific research questions. For example, always having cryptococcosis as the third disease allows a controlled comparison of the impact of the previous two diseases on learning.

1.2.2 Results

Figure 1.2 shows the category learning performance of all 38 participants. Each panel corresponds to a participant and the panels are organized by condition. The colored lines show the average proportion of correct responses over the last 10 trials for the current disease category. Different diseases are indicated by different colors. For most participants, there is a clear pattern of a sudden decrease in accuracy following a change and then subsequent

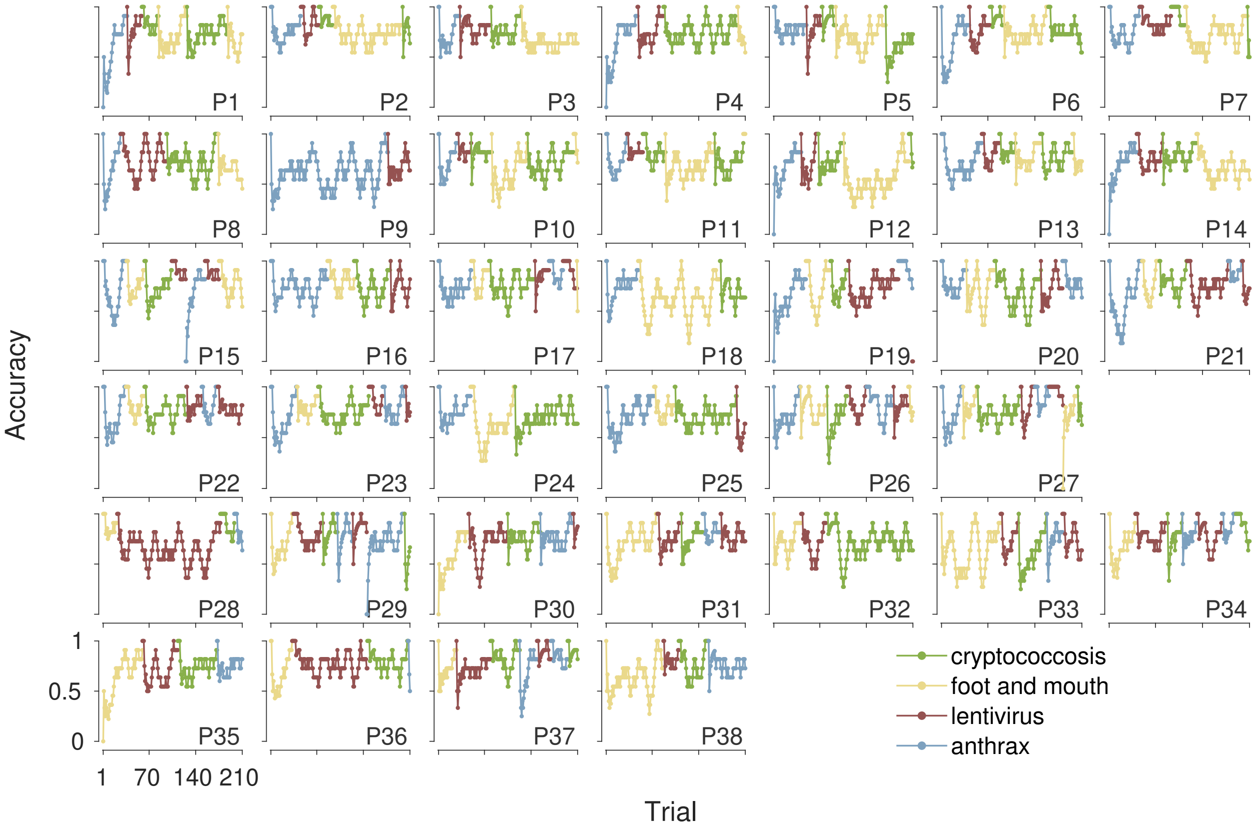

Figure 1.2: Category learning performance for all 38 participants. Each panel corresponds to a participant, with colored lines showing their average proportion of correct responses on the last 10 trials with the current category. Changes in category are shown by different colors. The participant panels are arranged so that the top two rows correspond to the first condition, the middle two rows correspond to the second condition, and the bottom two rows correspond to the third condition.

learning of the new disease category. After learning the first disease in their sequence, most participants maintain an average accuracy well above chance for the remaining diseases, which suggests some beneficial transfer of learning from one category to the next.

There are also clear individual differences. For example, participant 28 learns the first foot and mouth disease category quickly, whereas participant 38 takes many trials to reach the criterion level of accuracy. Interestingly, however, participant 28 then takes many trials to learn the subsequent lentivirus disease, whereas participant 38 now learns quickly.

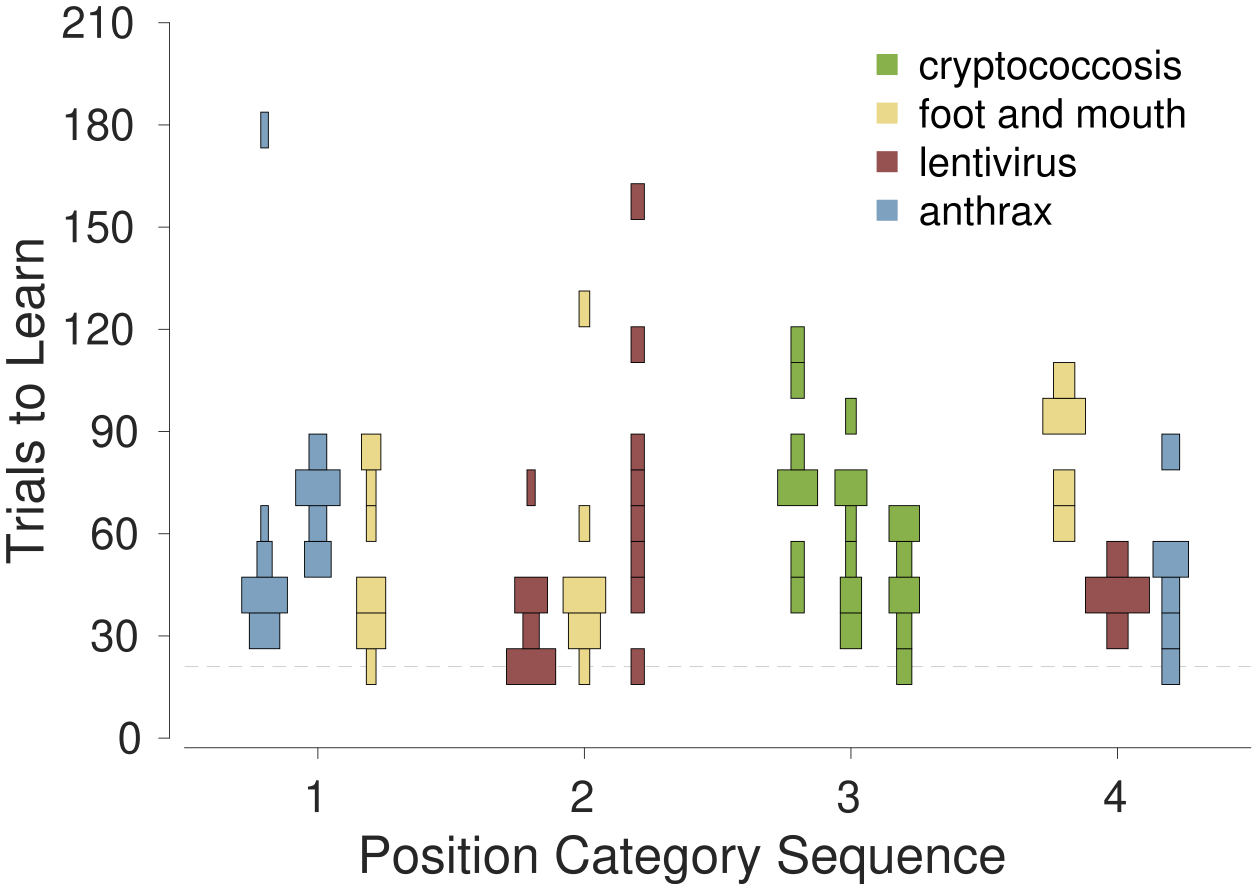

Figure 1.3: The distribution of the number of trials needed to learn the category at each position in the sequence. The distributions for conditions 1, 2, 3 are shown from left to right at each position. Distributions are colored according to the disease category. The dotted gray line shows the minimum of 21 trials needed to demonstrate learning.

Trials Needed to Learn Categories

Each of the four disease categories are about equally difficult to learn. Aggregated over all participants, and all of their attempts at learning the categories, the mean (standard deviation) number of trials to learn is 46.7 (22.0) for cryptococcosis, 52.0 (27.5) for foot and mouth, 39.0 (26.8) for lentivirus, and 41.8 (25.2) for anthrax. The Bayes factor for a one-way ANOVA is greater than 1000 in favor of these distributions having the same mean, rather than independently different means. This is evidence that the average number of trials that it takes participants to learn does not depend on the category.

Figure 1.3 shows a vertical histogram of learning times, considering both the disease category and its position in the learning sequence. The width of each square shows the frequency with which participants took that number of trials to learn that category in that position . Only the first four positions are considered, corresponding to the first time a participant encountered each disease category. For each position, there are three possibilities, corresponding to the three conditions. There is little evidence of differences in the learning distributions across the four positions. The Bayes factor for a one-way ANOVA is greater than 1000 in favor of sameness. Comparing the same disease in different positions also shows few differences. A t-test comparison of group means for anthrax in the first versus fourth positions provides an inconclusive Bayes factor of 2.0 in favor of a difference. Comparing lentivirus in the second and fourth positions provides a Bayes factor of 5.4 in favor of sameness. Comparing the three distributions of cryptococcosis which involve different prior learning experiences across the conditions, a one-way ANOVA provides a Bayes factor of 9.3 in favor of sameness.

Overall, there are neither strong nor systematic differences in the distributions of the number of trials needed to learn the different categories at different positions in the sequence. This is an interesting finding. On the one hand, the overlap between the different categories shown in Figure 1.1 means there is clearly some transfer advantage from prior learning. Many of the animals learned to be healthy, for example, will remain healthy. On the other hand, the similarity of the categories means prior learning could interfere with the fine-grained distinctions needed to master a new disease. The results in Figure 1.2 suggest these transfer and interference effects tend to balance each other out.

Accuracy Before and After Category Changes

There are four possible patterns of category association for an animal over a change in category structure. An animal can be diseased in both categories, healthy in both, change from being healthy to diseased, or change from being diseased to healthy. Figure 1.4 shows

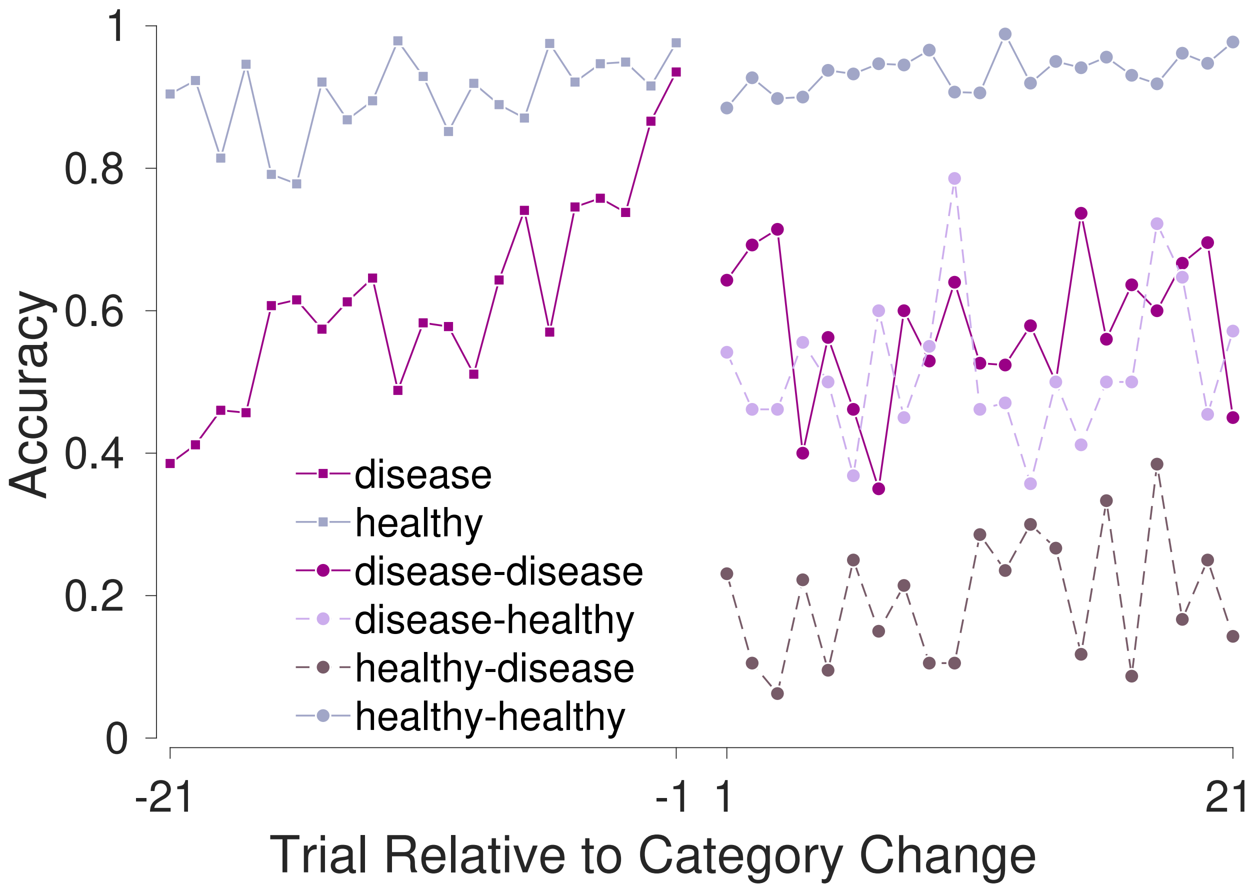

Figure 1.4: Accuracy for different patterns of change between diseased and healthy animals across category changes. The four lines correspond to the different transitions between healthy and disease categories, showing the average accuracy across all participants, stimuli, and changes for the 21 trials leading up to the change and the 21 trials after the change.

the change in accuracy for these four different possibilities, for the 21 trials following a category change. It also shows accuracy for the diseased and healthy animals in the original category structure for the 21 trials leading up to a category change. The measures of accuracy are aggregated over all participants, animal stimuli, and disease category transitions.

Leading into the category change, most learning is evident for the diseased animals, with the healthy animals generally already being accurately categorized throughout. After the category change, those animals that remain healthy continue to be accurately categorized. Animals that continue to be diseased, in contrast, are suddenly much less well categorized, with accuracy falling to around 50%. This is about the same level as animals that have changed from being healthy to diseased. There is an even more drastic drop in accuracy for animals that were healthy but have become diseased.

One interpretation of this pattern of results is that participants assume that the healthy animals continue to be healthy after a category change, but decide to re-learn the diseased animals. The assumption of stability in healthy animals is consistent with high healthy-healthy accuracy but low healthy-disease accuracy. The assumption of re-learning the diseased category is consistent with disease-disease and disease-healthy having the same moderate accuracy, independent of whether or not the now diseased animals changed category.

1.3 Experiment Two

1.3.1 Method

Participants

55 undergraduate student participants at the University of California, Irvine completed the experiment for course credit.

Stimuli

Three different stimulus domains, involving animals that could be diseased or healthy, items that could be allowed or prohibited by the TSA, and produce that could be in season or out of season, were used.

Animal Domain. The stimuli were images of 21 animals divided into two mutually exclusive categories, "healthy" or "infected", depending on four category structures. The category structures corresponded to four real-world diseases: brucellosis, echinococcus, heartwater and

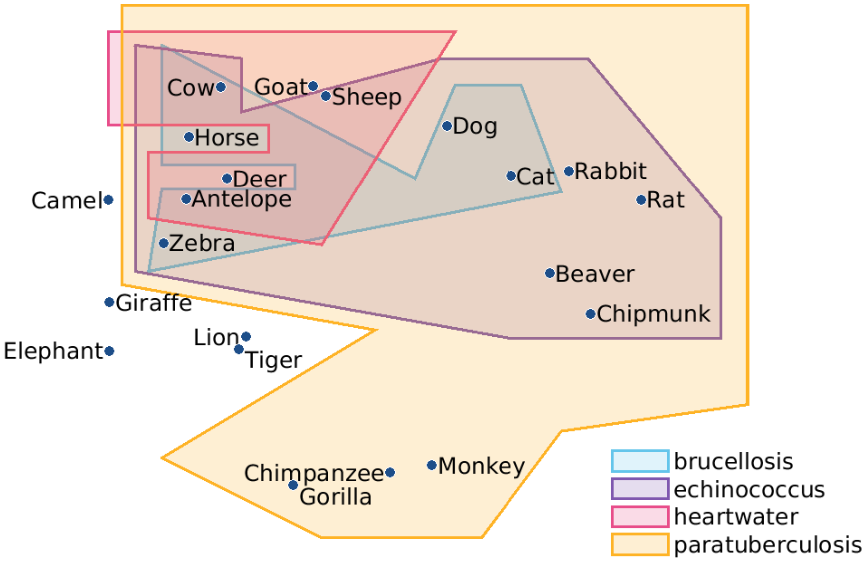

Figure 1.5: The four category structures. The 21 animal stimuli are represented as points arranged so that more semantically similar animals are located nearer each other. The four diseases are represented by colored regions that encompass those animals that have the disease.

paratuberculosis. Figure 1.5 shows the names and assignment to the “infected” category of the 21 animals. As can be seen in the figure there is significant overlap between the four category structures. For example, the Cow and Antelope are always infected while the Lion and Tiger are always healthy regardless of the category structure.

TSA Domain. The stimuli were images of 21 every-day objects divided into two mutually exclusive categories, “allowed” or “prohibited” depending on four category structures. The category structures corresponded to the list of every-day objects prohibited in flights published by the TSA in four different years. The name of the stimuli, category structures, and category membership are presented in Table 1.1. As can be seen in the table, for all four category structures, the base rate of the prohibited category is always larger, with only between 3 and 4 items at a time being allowed at any given year. This is a similar to the situation in experiment one in which the healthy category always had a larger base rate.

Produce Domain. The stimuli were images of 18 fruits and vegetables divided into two mutually exclusive categories, “in season” or “out of season” depending on four category structures. The category structures corresponded with the yearly seasons. The name of the

Stimulus	2005	2009	2014	2022
Bb guns	prohibited	prohibited	prohibited	prohibited
Brass knuckles	prohibited	prohibited	prohibited	prohibited
Cattle prods	prohibited	prohibited	prohibited	prohibited
Common lighters	prohibited	allowed	allowed	allowed
Crowbars	prohibited	prohibited	prohibited	prohibited
Gas torches	prohibited	prohibited	prohibited	prohibited
Gel heating pads	allowed	prohibited	prohibited	prohibited
Gel shoe inserts	allowed	prohibited	prohibited	prohibited
Gunpowder	prohibited	prohibited	prohibited	prohibited
Meat cleavers	prohibited	prohibited	prohibited	prohibited
Party poppers	prohibited	prohibited	prohibited	prohibited
Power tools	prohibited	prohibited	prohibited	allowed
Saws	prohibited	prohibited	prohibited	prohibited
Scissors	prohibited	allowed	allowed	allowed
Snow globes	allowed	prohibited	allowed	allowed
Sparklers	allowed	allowed	allowed	prohibited
Starter pistols	prohibited	prohibited	prohibited	prohibited
Strike-anywhere matches	prohibited	prohibited	prohibited	prohibited
Stun guns	prohibited	prohibited	prohibited	prohibited
Swords	prohibited	prohibited	prohibited	prohibited
Throwing stars	prohibited	prohibited	prohibited	prohibited

Table 1.1: Stimulus names, category structures, and category membership in the TSA domain

stimuli, category structures, and category membership are presented in Table 1.2. As can be seen in the table, the base rate of the “in season” category is larger in 3 of the four seasons, however, this changes abruptly in the case of summer, where the “out of season” category has the largest base rate of the two.

Procedure

This experiment had the same procedure as experiment one. On each trial people were presented with a single stimulus which they had to assign to one of the two mutually exclusive categories of the active domain. Participants received feedback on the accuracy of their categorization choices of the form “correct, the sage is in season” and then moved to the

Stimulus	Summer	Fall	Winter	Spring
Avocados	in season	out of season	in season	in season
Bay leaves	in season	in season	in season	out of season
Beets	in season	in season	in season	out of season
Brussels sprout	out of season	in season	in season	in season
Cabbages	out of season	in season	in season	in season
Celery	out of season	in season	in season	in season
Collard greens	out of season	in season	in season	in season
Garlic	in season	in season	out of season	in season
Kale	out of season	in season	in season	in season
Kiwifruits	out of season	in season	in season	in season
Lettuces	out of season	out of season	out of season	out of season
Mushrooms	out of season	in season	in season	in season
Onions	out of season	in season	in season	in season
Parsnips	out of season	in season	in season	in season
Pineapples	out of season	in season	in season	in season
Sage	in season	in season	in season	out of season
Swiss chard	out of season	in season	in season	in season
Turnips	out of season	in season	in season	in season

Table 1.2: Stimulus names, category structures and category membership in the produce domain.

following trial. The presentation of the three domains was randomized within participants. The experiment ended when participants had reached the criterion of the last domain. The learning criterion and the environment dynamics varied by domain.

Animal Domain. Category structures of the disease domain where presented to participant's in a randomized order and without replacement. The active category structure changed when a participant reached an accuracy level of 80% in the last 20–21 trials. This change was not signaled to the participants. Trials on the animal domain ended under two conditions, either a participant reached the learning criterion for all four category structures or after 210 stimulus presentations.

TSA Domain. Category structures in the TSA domain where presented in chronological order starting with the year 2005. If a participant reached an accuracy level of at least 80%

in the last 20–21 trials the active category changed to the following year in chronological order. Trials in this domain ended if either a participant had reached the learning criterion in all four category structures or after 210 stimulus presentations.

Produce Domain. The initial category structure in the produce domain was selected at random and changes followed a chronological order after. Changes in this domain where independent of participants behavior, and each category structure consisted of a single presentation of each of the 18 stimuli. Trials in this domain ended after the presentation of the 18 stimulus in each of the four category structures for a total of 72 trials.

1.3.2 Results

Figure 1.6 shows the number of participants who experienced one or more of the category structures in each domain. Notice that because changes in the Produce domain are independent of participant's accuracy every participants saw all four category structures. From the 55 participants in the task, only 15 experienced all four category structures in the Disease domain. This is striking contrast to the 34 participants who experienced all category structures of experiment one. The difference in the proportion of participants who completed the task in the two experiments could be attributed to the difference in the base rate of the healthy and infected categories. In the TSA domain 51 of the participants experienced all four category structures.

Results in the remaining of this section focus on the 14 participants who experienced every category stricture in the experiment. Figure 1.7 shows the category learning performance of these 14 participants for for animal diseases, TSA items, and produce seasonality from top to bottom. The panels for individual participants are shown in the same order across domains. The colored lines show the average proportion of correct responses over the last 10 trials for the current category structure and different structures are indicated by colors.

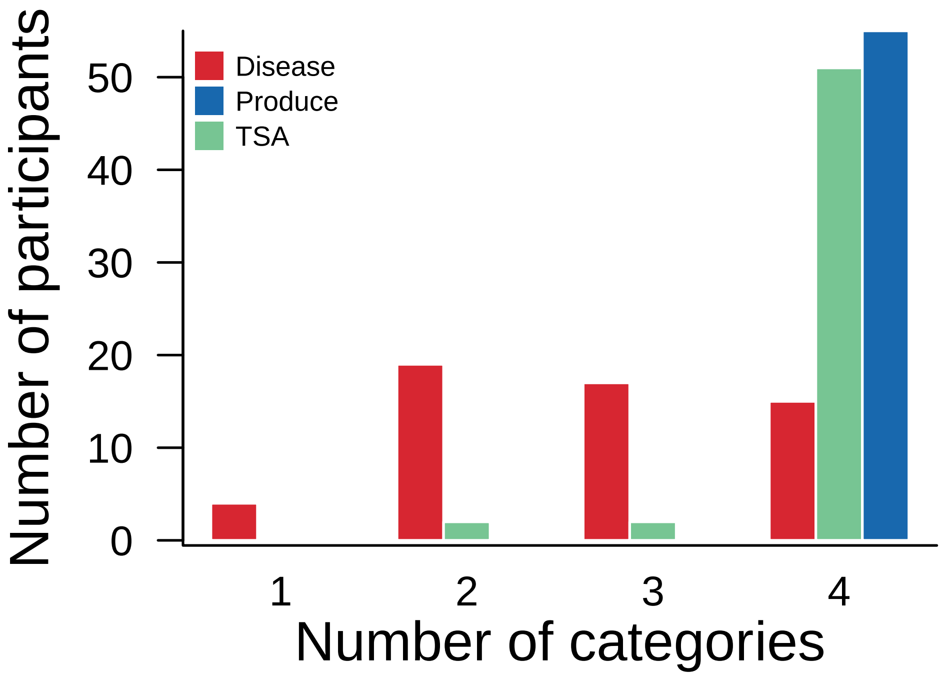

Figure 1.6: Number of categories experienced by participants in each domain. The bars represent the number of participants that experienced one, two, three, or four of the category structures in each domain of experiment two.

Participant performance in the Disease domain shows a clear patter of a sudden decrease in accuracy following a change and then a subsequent learning of the new category structure. A similar patter is observed in the TSA domain with drops in the accuracy followed by learning. This pattern is less consistently observed in the Produce domain: the learning of the category structure is observed only in some participants and category structures. For example, the accuracy of participant 6 continues to drop after a change in the category structure from summer to fall with no indication of learning the new category. However, when going from winter to spring the expected drop in accuracy followed by learning of the new category is observed.

Although the category learning performance of participants shows similar patterns in the Disease and TSA domains, there are some important differences. The average number of

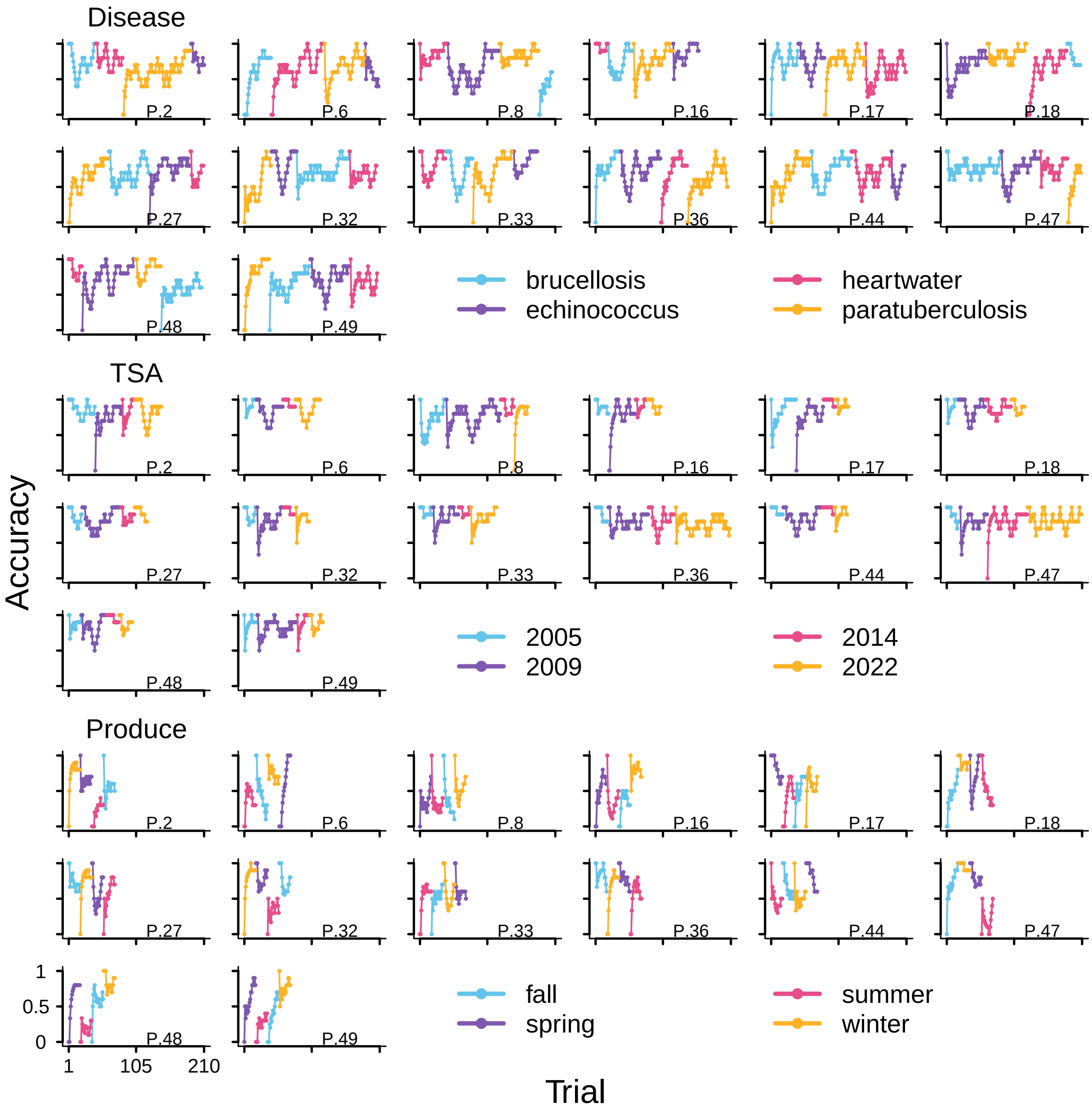

Figure 1.7: Category learning performance of 14 participants by domain. Domains are divided into sections with participant's behavior shown in the same order. Each panel presents a different participant. Colored lines represent the running average (window size 10) of the proportion of correct responses for each category structure. Changes in color represent changes in the active category

trials to reach criterion on each of the four category structures is different, with an average of 207 trials for the Disease domain and 140 for the TSA domain. In the TSA domain most participants maintain an average accuracy well above chance for most of the years, which is not the case for the Disease domain. As for experiment one, these differences can be explained by the consistency in category membership across years in the TSA domain, which promotes a beneficial transfer of learning from one category structure to the next.

Accuracy Before and After Category Changes

For each of the three domains, there are four possible patterns of category association for a stimulus over a change in the category structure for all participants in the task. Figure 1.8 shows the change in accuracy for these four possibilities of each domain for the last 18 trials leading up to a change and the 18 trials following it. The measures of accuracy are aggregated over participants, stimuli, and category structure transitions. The specific patterns in each panel do not change when looking only at the 14 participants that completed every category structure. They are, however, clearer when every participant is included in the analysis.

The top panel in Figure 1.8 shows the accuracy measures before and after a change in the animal disease domain. Leading up to a change in the category structure both healthy and infected categories have a similar accuracy. This result is different from experiment one, in which the accuracy for one of the categories increased on the trials leading up to a change. After the change in the category structure the accuracy of categorizing stimuli that have switched from healthy to infected and vice-versa increases. On the other hand, the accuracy for stimuli that preserve their category seems to decline slightly. These results suggest that after the category change, participants could be “exploring” the new environment by modifying previous category assignments.

The middle panel in Figure 1.8 shows the accuracy measures before and after a change in the Produce domain. The accuracy measure for both in-season and out-of-season items do not seem to increase on trials leading up to a change. However, we can observe some evidence for learning after a change in the category structure. For example, both the purple and

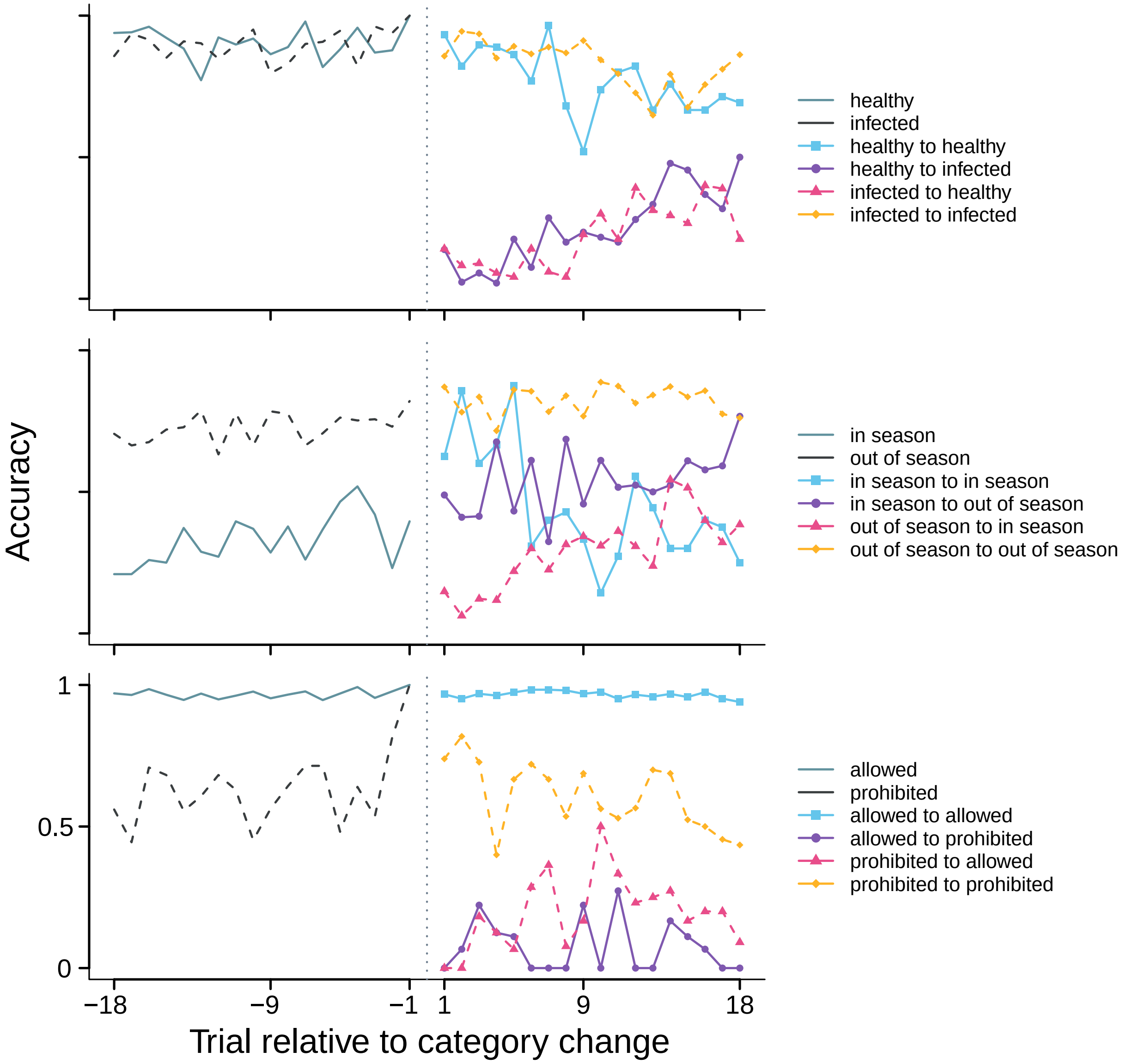

Figure 1.8: Accuracy for different patterns of change between positive and negative categories across domains and category structures. The four color lines correspond to the different transitions between positive and negative categories, showing the average accuracy across all participants, stimuli, and changes for the 18 trials leading up to the change and the 18 trials after change.

pink lines which represent the accuracy for items that have switched labels after a change in the category structure increase as trials move away from the change point. In comparison, the accuracy for items that remain “in season” decreases. This change could be explained by participants being more willing to change their category assignments for stimuli in that category. Nonetheless, it would not explain learning of items that switch from the "out of seasson" category.

The bottom panel of Figure 1.8 shows the accuracy measures before and after a change in the TSA domain. Leading up to a change in the category structure, accuracy for the allowed category is above 0.9 and remains high leading up to a change in the category. In comparison, the accuracy for stimuli in the prohibited category increases as trials approach the change point. This is the opposite of the pattern observed in experiment one, in which accuracy for the low base-rate category increased on trials leading up to a change, and accuracy for the high base-rate category was above 0.75 and stable. In the TSA domain, accuracy for items in the low base-rate category remains high and stable on trials following a change. In comparison, the accuracy for items in the high base-rate category decreases. This suggests that participants might be modifying some of their categorization choices for the prohibited category after a change. For items whose category membership changes, accuracy increases slightly which is an indication of learning.

1.4 Modeling

Many standard models of trial-by-trial category learning with feedback rely on incremental learning rules that adjust the associations between stimuli and categories (Kruschke, 2008; Shanks, 1991). Combining this approach to learning with similarity-based generalization gradients based on exemplar representation of stimuli, has been shown to avoid catastrophic forgetting (Kruschke, 1993b), and leads to the influential ALCOVE model (Kruschke, 1992) and its variants. We base our modeling on the version of ALCOVE developed by Lee & Navarro (2002) that relies on feature-based representations, since the animal stimuli seem better represented in terms of high-level cognitive features than low-level perceptual dimensions.

Our empirical results for experiment one, especially through the analysis in Figure 1.4 suggest that adaptation to category change can be understood in terms of what associations are preserved and reset when the category changes. To explore this intuition, we compare a basic model with only incremental learning against an extended model that resets the associations for previously diseased animals after each category change.

1.4.1 Two Learning Models

Formally, the similarity between animal i and j is represented by s_{ij} which are calculated as

$$s_{ij} = \exp\left(-\sigma\left[\sum_x f_{ix}(1-f_{jx}) + \sum_x (1-f_{ix})f_{jx}\right]\right), \tag{1.1}$$

where the f_{ix} are binary features, with $f_{ix} = 1$ if the animal has feature x and $f_{ix} = 0$ if it does not. The combination of features in Equation 1.1 provides a measure of the difference between animals i and j, consistent with the contrast model (Tversky, 1977). The exponentiation corresponds to a standard form of generalization gradient (Shepard, 1987), with a decay parameter $\sigma > 0$. We use the features for the animal stimuli found using similarity modeling by Westfall & Lee (2021). The association between the animal i and category k (i.e., healthy or diseased) is represented by a weight w_{ik}, all of which start at zero for the first category. When animal i is presented, the overall response strength for each category is calculated as

$$r_{ik} = \sum_j s_{ij} w_{jk}, \tag{1.2}$$

which provides a response probability

$$P\left(R=k \mid i\right)=\frac{\exp\left(\phi r_{ik}\right)}{\sum_{g}\exp\left(\phi r_{ig}\right)}, \tag{1.3}$$

where $\phi > 0$ is a response determinism parameter.

Once a decision has been made and feedback received, a standard delta learning rule is used to update the association weights (Sutton & Barto, 1998)

$$\Delta w_{jk}=\lambda(t_k-r_{ik})s_{ij}, \tag{1.4}$$

where $0 \leq \lambda \leq 1$ is a learning rate parameter and t_k is the teacher signal for category C_k. Following Kruschke (1992) the "humble" teacher feedback is

$$t_k=\begin{cases}\max\left(+1,r_{ik}\right) & \text{if } i \ \in \ C_k\\ \min\left(-1,r_{ik}\right) & \text{if } i \ \notin \ C_k.\end{cases} \tag{1.5}$$

The learning rule for the weights is designed to minimize the sum-squared difference between the response strengths and teacher values.

In the extended "reset" model, associations for the disease category are reset to zero when a category changes. This is consistent with the observation that accuracy for animals that had previously been diseased is relatively low after the category change, regardless of their new category membership.

1.4.2 Modeling Results

We applied both the basic and reset models to the category decisions made by the 34 participants who performed well enough to encounter each disease category at least once. The

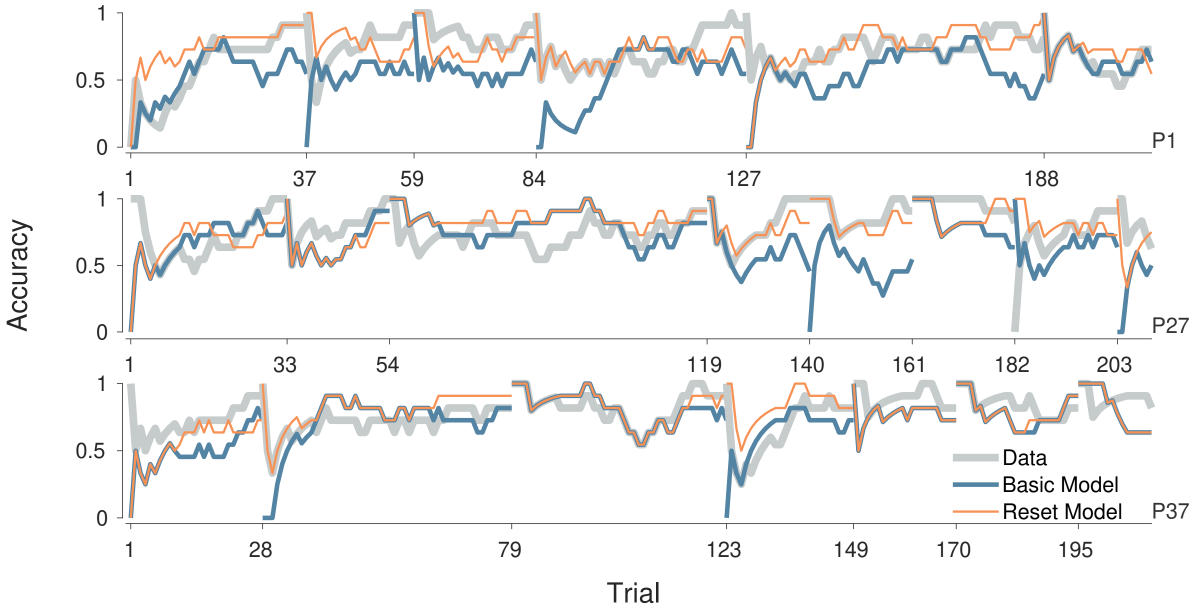

Figure 1.9: Model performance for three representative participants. The lines show the smoothed accuracy of the behavioral data, the basic model, and the reset model. The x-axis tick marks indicate the trials at which a category change occurred.

two models were fit independently for each participant using maximum likelihood, optimizing the σ, ϕ, and λ parameters. At the maximum-likelihood values, the basic model agrees with the behavioral data for 76% of trials. The reset model agrees on 81% of trials. This improvement is consistently shown at the participant level, with 30 out of 34 participants better described by the reset model.

Figure 1.9 provides insight into how the reset model improves upon the basic model. It shows the accuracy over trials of both models and the behavioral data for three representative participants, with one participant chosen from each condition. The reset model is generally able to describe performance between category changes a little better than the basic model, although both are far from perfect. The reset model is often significantly better, however, at describing the participants' behavior immediately after a category change. The basic model regularly shows very low accuracy for a few trials, whereas the reset model shows patterns of accuracy qualitatively more consistent with participant performance. This discrepancy does not happen after every category change. There are exceptions in which a participant does drop to very low accuracy after a category change. Overall, however, the additional assumptions in the reset model seem to capture an important aspect of participant behavior that often occurs.

1.5 Discussion

We considered two category learning experiments in which people's success in learning an underlying category structure caused changes in those categories. Our analysis of experiment one suggested that people adapt to the changing categories by preserving their knowledge about the high base rate category, healthy animals, but discarding information about diseased animals. A model that incorporated this insight provided a better account of people's behavior than a standard incremental associative learning model.

An interesting question is why there is an asymmetry between the disease and healthy categories. It seems participants actively re-learned the diseased animals but not the healthy animals after a category change. One possible explanation for this difference is in terms of the category structures shown in Figure 1.1. Many animals are healthy in all of the categories, whereas only two animals are always diseased. A different, potentially complementary, explanation is in terms of the semantics of the categories. It seems natural to treat the category learning task as requiring the concept of "diseased animals" to be learned. This would naturally lead to an emphasis on positive instances of the category (Tenenbaum & Griffiths, 2001), meaning animals categorized as diseased are the focus of re-learning after a category change.

Experiment two tries to address these two possibilities experimentally by presenting participants with a category learning task in three different domains. Our analysis from the second experiment suggest that, although a categories base rate plays a role in the adaptation to a change in the category structure, this role is not as straight forward as experiment one suggests. For example, participant's behavior in the TSA domain showed similar patterns to participant's behavior in experiment one, but with a change in the roles of the low and high base-rate categories. In this case, accuracy for the low base-rate category was always higher and stable across trials, similar to what was observed for the high base-rate category in experiment one. This could be explained by an interaction between a categories base rate and the ability of participants to memorize category structures built from a small number of stimuli.

Furthermore, in contrast to the results of experiment one, in the animal domain of experiment two we did not observe an asymmetry in participants' accuracy between healthy and infected categories. Our analysis of this domain suggest that participants adapt to the new category structure by relabeling items from both categories. This is indicated by a decrease in participants' accuracy for stimuli that maintain their category membership after a change in the category structure, and an increase in accuracy for stimuli that have switched. These two distinct patterns of results go against the category semantics explanation, which posits that participants treat the category learning task as requiring the "disease animals" concept to be learned.

Another interesting result from experiment two are the instances of learning observed in the produce domain. The dynamics of the environment in the produce domain were such that participants encountered each stimulus only once before a change in the category structure. Therefore, patterns in participants' accuracy that suggest learning have to be driven by the interactions between stimuli, the consistency of stimuli representations across category structures, or a combination of both.

People learn about their world in order to guide their future decisions and actions. This means that learning can impact the world, and introduces a coupling between what people learn as their environment changes and how the environment actually does change. Thus, restricting the study of category learning, or any learning or decision-making process, to static environments or environments that change independent of people's behavior, fails to consider an important aspect of human learning.

1.5.1 Implications for Modeling

One of the objectives of experiment two was to test the generality of our findings in the first experiment using other stimuli, categories and category structures. In particular we were interested in testing the effects of a categories base rate, and to compare our results with environments in which the underlying category structures had more variability.

Comparing the results of the two experiments suggests that assuming there is a mechanism that resets the associations between stimuli and a single category might not be general enough. Similar modeling approaches have been proposed before. For example, Kruschke (2003) introduced a model where attention shifts provide a mechanism that could account for how people quickly learn as categories change. Similar to our modeling results, an attentional shift could explain a subset of the results from the second experiment, in which accuracy changes for both categories. However, this assumption would not account for the results of experiment one or of the TSA domain in experiment 2, in which participants' accuracy changes only for one category when a new category structure is introduced.

There is also a broader cognitive science literature from artificial intelligence and machine learning in concept and context drift that is relevant for understanding how participants performed in our task (Iwashita & Papa, 2019; Widmer & Kubat, 1996). For example, Devaney & Ram (1996) study small changes in category structure for the same set of stimuli. This is the same basic situation as we studied, as the overlap between categories in Figure 1.1 shows. They develop a COBWEB account of this sort of category drift, using a modeling framework rooted in economic models of market fluctuations. As alternative modeling approaches, Maloof (2003) develops a system that adapts by removing irrelevant examples of old concepts, Koychev (2007) presents a statistical method based on making inferences about change points, and D'Alessandro et al. (2020) develops a Bayesian model that relies on a probability distribution over possible states and the interactions between responses and feedback.

Although there are some models that could account for aspects of participants' behavior in our experiments, one of our major challenges is to detect when the categorical representation of a stimulus has changed. For example, when a participant is categorizing animals in the first experiment, if we observe two different categorization decisions for the gorilla, an important question is when this change in the categorical representation of the gorilla occurred. This change in the representation could have a rippling effect on how other stimuli are categorized in the same context. To solve this problem we need models that can measure or infer the categorical representation of the stimuli in a task across time.

An alternative framework that could be used to solve this problem is one based on Coupled Hidden Markov Models (CHMM). In a CHMM model, the category that a stimulus is assigned to in every trial can be represented as a hidden process that needs to be inferred. This inference process is guided by participant's categorization decisions on every trial and the assumed stimulus-stimulus interactions. Therefore, this approach allows us to model how the categorical representations of stimuli in a task interact across time. This is accomplished by making the category assignment of a stimulus a function of the classification of all other stimuli in the same task.

Motivated by these experimental results and the limitations of existing models, the following chapters introduce the CHMM approach to category learning tasks and then test its ability to make inferences in experimental designs with transfer stimuli.

1.6 Publication Note

Parts of this chapter were previously published as Villarreal, M., Vaday, S., & Lee, M. D. (2022). Categorization in environments that change when people learn. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society* Toronto, CA. (pp. 680–686).

Chapter 2

A Coupled Hidden Markov Model Framework for Measuring the Dynamics of Categorization

Abstract

We introduce a new framework for measuring the dynamics of category learning using Coupled Hidden Markov Models (CHMMs). The key assumptions of the framework are that people maintain a latent assignment of every stimulus to a category, and can update the assignments for all stimuli whenever they encounter any stimulus. These assumptions contrast with many existing accounts of category learning, which either do not allow for what is learned about one stimulus to influence the category association of others, or allow only for indirect influence. The CHMM framework allows tailored models to be developed for specific category learning tasks, taking as input the stimulus sequence and category responses people make, and producing as output inferences about the underlying dynamics of category

assignments and the mechanics of the response processes. We demonstrate the framework by developing a model of a categorization task considered by Lee & Navarro (2002), showing how the model measures the change in participants' latent category assignments as they learn the category structure. We conclude by discussing potential applications of the CHMM framework to category learning situations involving prior knowledge, changing category structures, and category learning tasks that involve the consideration of multiple stimuli at one time.

2.1 Introduction

Wittgenstein (1958) famously considered the challenge of deciding what activities can be categorized as a game. Imagine that you are walking down a street and see a child bouncing a ball from a wall and catching it. Would you consider the child to be playing a game? You continue your walk and now see two children bouncing a ball from the wall. They take turns throwing and catching. Is this a game? If you thought the first activity was not a game, but that the second was, an additional question arises. Should you now change you beliefs and regard the first activity as a game? This simple example highlights a property of people's category learning. The category labels that we assign to newly encountered stimuli can potentially change our beliefs about previously encountered stimuli. Even if we are only observed to be labeling the stimulus in front of us, our latent beliefs about the category structure of all stimuli can be changing.

Some category learning models do not allow for what is learned about one stimulus to affect the category assignments of other stimuli. An example is the simple two-layer connectionist network developed by Shanks (1991). In this model, there are associations between nodes representing stimuli and nodes representing categories. The associations for a stimulus can only be changed by the learning rule when the stimulus is presented. This means that the category associations of other stimuli are not affected by what has been learned for the presented stimulus.

More often, category learning models do incorporate some form of stimulus-stimulus interaction, but only in an indirect way. One class of models with this property includes ALCOVE (Kruschke, 1992) and its extensions (Kruschke, 1993a; Kruschke & Johansen, 1999; Lee & Navarro, 2002). In these models, the learning rule again only adjusts the associations between the presented stimulus and the categories. There are, however, similarity relationships between the stimuli, often incorporating aspects such as selective attention, that are also learned. The relationship between stimuli means that changes in the associations of a presented stimulus A have indirect consequences for the category assignments of another stimulus A'. But these consequences can only be determined by calculating the response that would be made if A' was to be presented. The hypothetical response for A' is sensitive to the category associations for A, according to the similarity between A and A', and so what has just been learned about A impacts future responses to A'. This modeling approach only incorporate stimulus-stimulus interactions to the extent that hypothetical future responses can be considered surrogates for latent category associations.

A second class of models that incorporate stimulus-stimulus interactions indirectly includes those based on associations between the properties of stimuli and the categories. This class includes models such as SUSTAIN (Love et al., 2004) and DIVA (Kurtz, 2007). In these models, learning about a stimulus involves learning about its dimensions, features, shared clusters with other stimuli, or some other set of properties. This means that what is learned about the properties of A affects the category assignment of another stimulus A' that shares some of those properties. Once again, however, the consequences can only be determined by calculating the hypothetical response that would be made if A' was to be presented.

In this article, we introduce a new framework for understanding the dynamics of category learning that takes a more direct approach. The framework directly models category representations, separating latent beliefs about category assignments from the processes that produce observable responses. The framework also directly models the stimulus-stimulus interactions that control how latent category assignments are maintained and updated. This means that what is learned about one stimulus can have an immediate impact on the latent category assignments of other stimuli.

Our framework is not designed to produce models of category learning. It does not operate by being presented with a sequence of stimuli and applying learning rules to make predictions about categorization responses. Instead, it is designed to produce measurement models that are capable of inferring from a person's responses how their category representation changed during learning. One way to think of these measurement goals is by analogy to the analysis of EEG data. In EEG analysis, a person does a task and electrophysiological data relevant to their cognitive experience are collected. After the task is complete, the data are analyzed to understand the underlying cognition, such as by identifying ERPs in the time-series of the EEG signal. In the same way, our approach analyzes a person's sequence of responses in a category learning task to understand how their underlying category assignments evolved.

With its measurement goals, our framework most closely resembles the Generalized Context Model of categorization (GCM; Nosofsky, 1986). Standard applications of the GCM assume that a participant knows perfectly the category assignment of a set of training stimuli for which extensive feedback is provided. It then measures how a participant represents the overall category structure using their responses to a set of different transfer stimuli for which no feedback is provided. Our measurement framework is more ambitious by focusing on measuring the dynamics of how a participant learns category assignments for all stimuli. The key assumptions are that people explicitly represent the probability that every stimulus belongs to the available categories, and that they can change these probabilities for any stimulus at any stage, including for stimuli not currently present in the task environment.

Our approach is based on interacting Hidden Markov Models, which is an active and well-developed area of research in statistics (Brand, 1997; Brand et al., 1997; Sherlock et al., 2013; Zhong & Ghosh, 2002). There are multiple ways in which two or more Hidden Markov Models can interact. For example, in a Linked Hidden Markov Model a hidden process depends on its own previous value and on the current value of other processes. In a Hidden Markov Decision Tree, hidden processes are organized hierarchically so that the current state of one process can affect the state of lower processes in the hierarchy. Of the various possibilities, the organization of a Coupled Hidden Markov Model (CHMM) best matches our intuitions about the dynamics of stimulus-stimulus interaction in category learning. A CHMM consists of two or more interacting Hidden Markov Models chains, each of which is built from two stochastic processes—a hidden one and an observable one—with the observed process depending on the hidden one. In the context of a categorization task, it is natural to think of the hidden processes as the latent category representations being maintained over time for all stimuli, and the observable processes as the trial by trial responses to presented stimuli.

The key assumption of the CHMM organization is that the category a stimulus is assigned to is only a function of the category assignments of all other stimuli on the previous trial. There is no interdependence between current assignments, just a dependence of each stimulus on the previous assignment of the others. This is known as the Markov property, and it allows for stimulus-stimulus interactions to be modeled directly by defining a transition function.

The remainder of this article is organized as follows. In the next section we formally develop the CHMM framework. We then demonstrate the framework by developing and applying a specific CHMM model designed to account for people's behavior in a simple categorization learning experiment reported by Lee & Navarro (2002). We show that the model describes the observed category responses well, by inferring changes in people's beliefs about category assignments that are highly interpretable and psychologically plausible. We conclude with a discussion of potential applications of the framework to more complicated category learning tasks, and future avenues for developing the general CHMM approach.

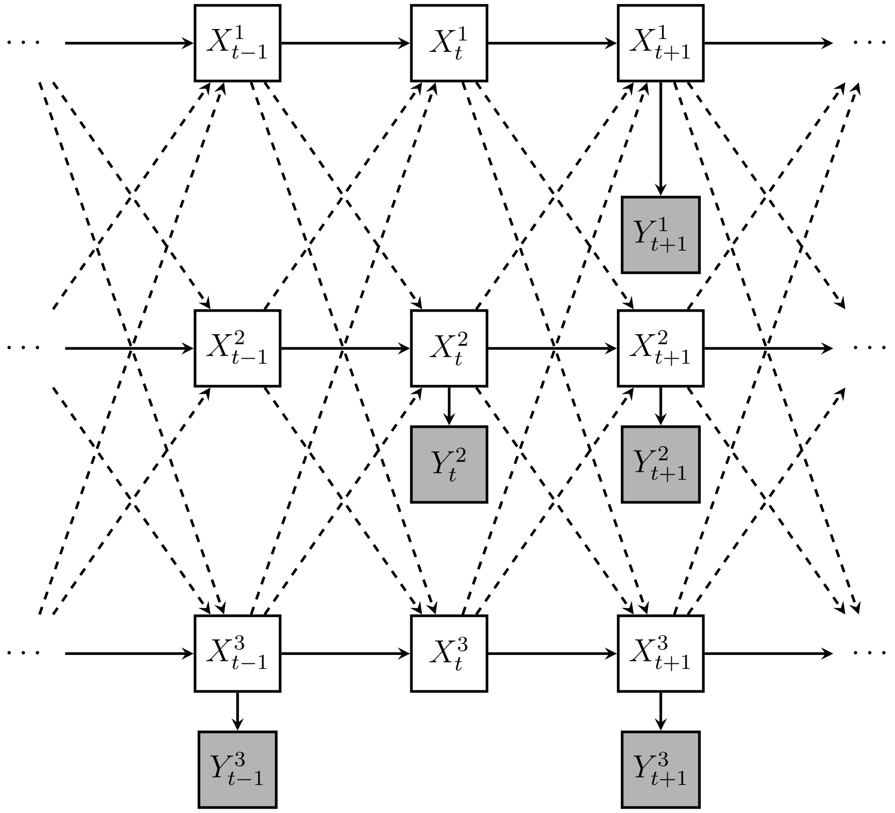

Figure 2.1: Graphical representation of a Coupled Hidden Markov Model. Shaded nodes represent the observed process, unshaded nodes represent the hidden process. Superscripts index different Markov chains that can interact with one another. Solid and dashed arrows represent the dependency structure of the model.

2.2 Coupled Hidden Markov Models

Formally, a CHMM is a stochastic process built from two or more interacting Hidden Markov models that run concurrently. Each Hidden Markov Model is comprised of an observable process $Y_{1:T} = (Y_1, Y_2, \ldots, Y_T)$ and a hidden process $X_{1:T} = (X_1, X_2, \ldots, X_T)$ over trials $1, \ldots, T$. The hidden process can take values on any trial from a set of finite and discrete values known as states $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$. An initial probability distribution $P(X_0 = \omega_i)$ over the set of states denotes the probability that the process will start with state ω_i. A transition probability function $P(X_t = \omega_i \mid X_{t-1}, \ldots, X_1)$ indicates the probability of moving to state ω_i conditional on the previous states of the process starting from $t-1$ and down to

1. Finally, an outcome probability function $P(Y_t \mid X_t, X_{t-1}, \ldots, X_1)$ links the observations $Y_{1:T}$ to the states of the hidden process.

There are two simplifying assumptions in a Hidden Markov Model. The first is the Markov property. It assumes that, conditional on the present state, the future of the process is independent of the past. This property is reflected in the transition function

$$P\left(X_t \mid X_{t-1}, \ldots, X_1\right) = P\left(X_t \mid X_{t-1}\right). \tag{2.1}$$

The second assumption is the conditional independence of the observable process given the current state, which is expressed as

$$P\left(Y_t \mid X_t, X_{t-1}, \ldots, X_1\right) = P\left(Y_t \mid X_t\right). \tag{2.2}$$

This equality indicates that, conditional on the present, outcomes are independent of the history of the hidden process.

Figure 2.1 shows the organization of a CHMM with three chains. The square unshaded nodes represent the hidden process, while the shaded nodes represent the observable process. The solid and dashed arrows represent the dependency between states and the coupling structure between them. The shaded nodes are connected only to their own hidden process, which means that this coupling structure preserves the conditional independence between states and outcomes as in Equation 2.2. Let $X^m_{1:T}$ and $Y^m_{1:T}$ represent the mth chain in a CHMM, then

$$P\big(Y^m_t \mid X^m_{1:t}, \boldsymbol{X}^{-m}_{1:t}\big) = P\big(Y^m_t \mid X^m_t\big), \tag{2.3}$$

where $\boldsymbol{X}^{-m}_{1:t} = \big(X^1_{1:t}, \ldots, X^{m-1}_{1:t}, X^{m+1}_{1:t}, \ldots X^M_{1:t}\big)$ represents the hidden process from trial 1 up to trial t of all chains except the mth one.

The main feature of a CHMM is integrated in the transition probability function. Unlike the basic Markov property in Equation 2.1, the transition function includes the influence of other hidden processes such that

$$P\big(X_t^m \mid X_{1:t-1}^m, \boldsymbol{X}_{1:t-1}^{-m}\big) = P\big(X_t^m \mid X_{t-1}^m, \boldsymbol{X}_{t-1}^{-m}\big). \tag{2.4}$$

This set of dependencies means that the mth hidden process depends not only on its own previous state, but also on the previous state of all other chains.

2.3 A CHMM Framework for Categorization

In a typical categorization task, participants are presented with a sequence of stimuli to label. When a participant assigns a label to a stimulus they are often presented with feedback about the accuracy of their response. From the perspective of a CHMM we can think of a single stimulus and a participant's responses to it over the experiment as a Hidden Markov Model. The hidden process $X_{1:T}^{m,p}$ denotes the categorical representation of stimulus m for participant p across trials $t = 1, 2, \ldots, T$ as one of the available categories from the set Ω. The labels that a participant assigns to that stimulus when it is presented are represented by the observable process $Y_{1:T}^{m,p}$. These labels form a different set of outcomes $O = \{o_1, \ldots, o_N\}$, but we assume that there is a one-to-one correspondence between the elements in Ω and elements in O. That is, we assume that the state ω_i corresponds to the label o_i. The CHMM thus allows the categorical representation of all stimuli $\boldsymbol{X}_t^{1:M,p}$ at any trial t to be inferred based on the observed labeling responses $\boldsymbol{Y}_t^{1:M,p}$. We also assume that participants share no information between them and thus the model is independent, which allows the p notation to be suppressed.

2.3.1 Initial probability

The initial probability function defines the probability that the mth stimulus will be assigned to one of the N available categories $\Omega = \omega_1, \ldots, \omega_N$ at the start of the task, and it can be interpreted as a prior distribution of the category assignments of all stimuli in the experiment. For a task with two or more available categories, a prior is

$$X_1^m \sim \text{categorical}\left(\gamma_1, \ldots, \gamma_N\right), \tag{2.5}$$

where $\gamma_i = P\left(X_1^m = \omega_i\right)$ is constrained such that $\gamma_i \geq 0$ and $\sum_{i=1}^{N} \gamma_i = 1$. Given these constraints, a general prior for the parameter $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)$ is

$$\boldsymbol{\gamma} \sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right), \tag{2.6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$. This allows an intuitive interpretation of α_i as the expected number of stimuli assigned to category ω_i at the start of the experiment, which is likely to depend on prior knowledge about the stimuli and possibly task instructions. For example, people often have reasonable prior expectations about the likely assignments of stimuli for familiar natural categories (e.g., Hemmer & Steyvers, 2009).

2.3.2 Transition function

The transition function defines the trial-by-trial probability of a change in the category assignment of a stimulus conditional on the representation of all of the other stimuli on the previous trial. This function models the interactions between the category assignments of stimuli in the experiment. We propose transition functions with the form

$$P\big(X_t^m = x_t^m \mid \boldsymbol{X}_{t-1}^{1:M} = \boldsymbol{x}_{t-1}^{1:M},\ \boldsymbol{\beta}\big) = \begin{cases} \text{logit}^{-1}\left(\beta_i + \bar{\eta}_i^m(t-1)\right) & \text{if } \ x_{t-1}^m = \omega_i \text{ and } \ x_t^m = \omega_i \\ \\ \text{logit}^{-1}\big(\beta_i + \bar{\eta}_j^m(t-1)\big) & \text{if } \ x_{t-1}^m = \omega_i \text{ and } \ x_t^m = \omega_j. \end{cases} \tag{2.7}$$

This function ensures that the transition probabilities from state ω_i to ω_j sum to one for all $i, j \ \in \ \Omega$. The "stickiness" parameter $\beta_i > 0$ can be interpreted as the tendency of stimuli to remain in category ω_i between consecutive trials, and is given the general prior

$$\beta_i \sim \text{gamma}\,(a, b)\,, \tag{2.8}$$

the choice of a determines the scale while b controls the rate of the distribution. As the value of β_i increases it is less likely that the latent category assignment of a stimulus will change from that category to another one. It would be reasonable to expect the stickiness of a parameter to be smaller if, for example, the task environment allowed for the stimuli in the category to change (e.g., Estes, 1984; Kruschke, 1996; Gallistel et al., 2001; Navarro et al., 2013; Speekenbrink & Shanks, 2010; Villarreal et al., 2022), or if the feedback received about whether a stimulus is in the category was probabilistic (e.g., Ashby & Maddox, 2005; Kruschke & Johansen, 1999; Knowlton et al., 1994; Gluck et al., 2002).

The function $\bar{\eta}_i^m(t-1)$ defines how the representation of all other stimuli in the task m' interact with stimulus m. We propose that this function takes the form

$$\bar{\eta}_i^m(t-1) = \frac{1}{\sum_{m' \neq m} \mathbf{1}_\Omega(x_{t-1}^{m'} = \omega_i)} \sum_{m' \neq m} \mathbf{1}_\Omega(x_{t-1}^{m'} = \omega_i) s_{mm'}, \tag{2.9}$$

where $s_{mm'}$ represents the similarity between stimulus m and m' for all stimuli $m' \ \neq m$ and $\mathbf{1}_\Omega(x_t^m = \omega_i)$ is the indicator function defined as

$$\mathbf{1}_\Omega(x_t^m = \omega_i) = \begin{cases} 1 & \text{if } x_t^m = \omega_i \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.10}$$

We refer to $\bar{\eta}_i^m(t-1)$ as the average similarity of stimulus m to items in the ω_i category at trial $t-1$. Many ways of modeling stimulus similarity have been developed as theories of mental representation (Shepard, 1980; Tversky, 1977). A number of these have been used in the categorization literature, including generalization curves applied to psychological distances (e.g., Kruschke, 1992; Nosofsky, 1986; Shepard, 1987), comparisons of common and distinctive features (e.g., Lee & Navarro, 2002), and rule-based measures (e.g. Nosofsky et al., 1994; Schlegelmilch et al., 2022). The appropriate measure of similarity will usually depend on the nature of the stimulus domain in a categorization task.

Together Equations 2.7 and 2.9 mean that the probability that a stimulus will continue to be represented as category ω_i on the next trial is a function of the stickiness β_i and the average similarity of stimulus m to all other stimuli represented as category ω_i in the previous trial. In this way, these two functions model how the category assignments of stimuli in the task interact with one another. The interaction is based on the similarity between items in the same category, consistent with exemplar models of categorization. The key feature of the CHMM approach, however, is that the inference about the category representation of every stimulus is made simultaneously on every trial.

An important special case of the transition function occurs if there are only two categories, which is the most common situation in experimental studies of category learning. In this case, it is often appropriate to use the difference between the average similarity of stimulus m to each category in the transition probability function in Equation 2.7. For two categories p and q, the average similarity of stimulus m to stimuli in category p relative to category q

is defined as

$$\bar{\eta}^m_{p,q}(t) = \bar{\eta}^m_p(t) - \bar{\eta}^m_q(t).$$

This leads to the transition probability function

$$P\big(X^m_t = x^m_t \mid \boldsymbol{X}^{1:M}_{t-1} = \boldsymbol{x}^{1:M}_{t-1},\ \boldsymbol{\beta}\big) = \begin{cases} \text{logit}^{-1}\left(\beta_i + \bar{\eta}^m_{p,q}(t-1)\right) & \text{if } x^m_{t-1} = p \text{ and } x^m_t = p \\ 1 - \text{logit}^{-1}\left(\beta_i + \bar{\eta}^m_{p,q}(t-1)\right) & \text{if } x^m_{t-1} = p \text{ and } x^m_t = q \\ 1 - \text{logit}^{-1}\left(\beta_i - \bar{\eta}^m_{p,q}(t-1)\right) & \text{if } x^m_{t-1} = q \text{ and } x^m_t = p \\ \text{logit}^{-1}\left(\beta_i - \bar{\eta}^m_{p,q}(t-1)\right) & \text{if } x^m_{t-1} = q \text{ and } x^m_t = q. \end{cases} \tag{2.11}$$

Intuitively, the probability that a stimulus will continue to be represented as category p is a function of its average similarity to other stimuli in category p in comparison to its similarity to stimuli in category q.

2.3.3 Response function

The response function links the state of the hidden process of stimulus m at a time t with the observed label choice. We denote the set of available label options as the set $O = (o_1, o_2, \ldots, o_N)$ and assume that there is a unique label option for each category in Ω. The conditional independence assumption in Equation 2.2 implies that responses to the presentation of stimulus m depend only on the category representation of the given stimulus. We assume, however, that people do not always respond according to the current category assignment of a stimulus. Instead, we assume that participants respond following a trembling-hand choice rule (Zilker, 2022).

We can define a vector $\boldsymbol{I}_{\omega_i}(x^m_t)$ that has a value of 0 on all but its ith entry, using this vector, the response rule can be expressed as

$$P\big(Y_t^m = o_i \mid X_t^m = x_t^m\big) \sim \text{categorical}\left(\ (\boldsymbol{I}_{\omega_i}(x_t^m), 1 - \boldsymbol{I}_{\omega_i}(x_t^m))\,\vec{\epsilon}'\ \right), \tag{2.12}$$

where $\vec{\epsilon}' = (\epsilon, 1-\epsilon)$ represents the probability of responding with a label that does not correspond to the current category assignment. We assume a general prior distribution

$$\epsilon \sim \text{beta}\,(d, g)\,. \tag{2.13}$$

for the probability of trembling-hand errors. Specific values of d and g correspond to assumptions about the likelihoods of errors in a given experiment. For example, if participants completed the task under time pressure or cognitive load, it would be reasonable to assume a greater probability of errors.

2.3.4 Inference

We used Bayesian computational sampling methods to make inferences about CHMM parameters applied to behavioral data. This requires sampling the joint posterior distribution of the initial probability $\boldsymbol{\gamma}$, the stickiness $\boldsymbol{\beta}$ of the categories, the trembling-hand error ϵ and category assignments of the hidden process $\boldsymbol{X}_{1:T}^{1:M}$ for all stimuli:

$$P\left(\boldsymbol{X}_{1:T}^{1:M}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \epsilon \mid \boldsymbol{Y}_{1:T}^{1:M}\right).$$

The parameters $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and ϵ can be sampled using conventional Markov-chain Monte Carlo methods such as Gibbs sampling (Chen et al., 2000) or Hamiltonian Monte Carlo (Neal, 2011).

Inference for the hidden process requires more specialized sampling approaches. One of the main algorithms developed for this purpose is the "Full Forward Filtering Backward Sampling" algorithm (Carter & Kohn, 1994; Chib, 1996), which is based on the forward-backward algorithm used for maximum likelihood inference in Hidden Markov Models (Baum, 1972). This sampling method involves the transformation of the original CHMM into an equivalent Hidden Markov Model where each state represents the Cartesian product of the individual chains X_t^m for each $m \in 1, \ldots, M$ and $t \in 1, \ldots, T$. The total number of states of this newly formed Hidden Markov Model at trial t is $|\Omega|^M$ where M is the total number of chains, which in our framework corresponds to the number of stimuli in a categorization task. As the number of possible states N or chains M increases this method becomes computationally expensive.

Instead, we adapted a computational inference algorithm based on the "Individual FFBS" approach developed by Touloupou et al. (2020). There are two steps involved in this method. First, a forward sweep is used to compute the modified conditional filtered probabilities

$$P\big(X_t^m = x_t^m \mid \boldsymbol{X}_{1:t+1}^{-m}, \boldsymbol{Y}_{1:t}^m, \boldsymbol{\gamma}, \boldsymbol{\beta}, \epsilon\big),$$

for $t = 1, \ldots, T$. These values can then be used to sample the hidden process starting from X_T^m and moving backwards to sample X_t^m, conditioning on the value of X_{t+1}^m for $t = T-1, T-2, \ldots, 1$ from the conditional distribution:

$$P\big(X_t^m \mid X_{t+1}^m = x_{t+1}^m, \boldsymbol{X}_{1:t+1}^{-m}, \boldsymbol{Y}_{1:t}^m, \boldsymbol{\gamma}, \boldsymbol{\beta}, \epsilon\big).$$

These two steps result in the full conditional distribution of the hidden process $X_{1:T}^m$ in closed form. This means that we can consider the samples from the backward sweep as the draw step from a Gibbs sampling algorithm (see Touloupou et al., 2020, for details), producing a

Algorithm 1: MCMC algorithm for the Coupled Hidden Markov model with individual FFBS (from Touloupou et al., 2020, see Algorithm 1).

```
Begin
    Sample: $\boldsymbol{\gamma}, \boldsymbol{\beta}, \epsilon \sim \pi(\boldsymbol{\gamma}, \boldsymbol{\beta}, \epsilon)$
    Generate: $\boldsymbol{X}_{1:T}^{1:M} \sim P\big(\boldsymbol{X}_{1:T}^{1:M} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \epsilon\big)$
    for $i = 1, 2, \ldots K$ do
        for $m = 1, 2, \ldots, M$ do
            Update: $\boldsymbol{X}_{1:T}^{m} \sim P\big(\boldsymbol{X}_{1:T}^{m} \mid \boldsymbol{X}_{1:T}^{-m}, \boldsymbol{Y}_{1:T}^{m}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \epsilon\big)$ using iFFBS.
        end for
        Update: $\boldsymbol{\beta} \sim P\big(\boldsymbol{\beta} \mid \boldsymbol{X}_{1:T}^{1:M}, \boldsymbol{Y}_{1:T}^{1:M}\big)$ with HMC.
        Update: $\boldsymbol{\gamma} \sim P\big(\boldsymbol{\gamma} \mid \boldsymbol{X}_{1:T}^{1:M}, \boldsymbol{Y}_{1:T}^{1:M}\big)$ with Gibbs.
        Update: $\epsilon \sim P\big(\epsilon \mid \boldsymbol{X}_{1:T}^{1:M}, \boldsymbol{Y}_{1:T}^{1:M}\big)$ with Gibbs.
    end for
End
```

single sample of the full conditional posterior distribution of the m-th hidden process:

$$P\big(\boldsymbol{X}_{1:T}^{m} \mid \boldsymbol{X}_{1:T}^{-m}, \boldsymbol{Y}_{1:T}^{m}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \epsilon\big).$$

Sampling is applied one hidden process at a time, conditioning on the sampled values on previous iterations of the algorithm. For details on these probability functions see Touloupou et al. (2020, Equations 4–8).

As in other Bayesian computational sampling methods, the model parameters can be updated in blocks using suitable sampling routines by conditioning on the sample of the hidden states. Because the parameters $\boldsymbol{\gamma}$ and ϵ have a closed form we use a Gibbs sampling algorithm. The sampling of the stickiness parameters is performed jointly using a Hamiltonian Monte Carlo algorithm. Pseudocode describing the entire Markov-chain Monte Carlo algorithm, based on (Touloupou et al., 2020, Algorithm 1) is presented in Algorithm 1.

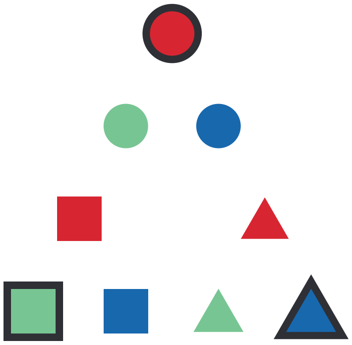

Figure 2.2: An example of a type IV category structure in Lee & Navarro (2002). The black borders indicate category membership, dividing the stimuli into a smaller and larger category with three and six stimuli respectively.

2.4 An Application of the CHMM

2.4.1 Lee & Navarro (2002) Experiment

Lee & Navarro (2002) report the results of a category learning task designed to test and extend the ALCOVE model to stimuli represented by discrete features. The experiment was based on the seminal task introduced by Shepard et al. (1961), who studied six different category structures for eight stimuli constructed in terms of three binary features. Lee & Navarro (2002) used nine stimuli constructed in terms of two ternary features: three shapes (square, circle and triangle) with three colors (red, green and blue). They tested how people learned four different category structures dividing the stimuli into a smaller category of three stimuli and a larger category of six. We focus on the data from what they termed the Type IV category structure. An example of a categorization task using these structure is shown in Figure 2.2, dividing the red circle, green square, and blue triangle from the remaining six stimuli.[1] Note that this category structure is analogous to the Shepard et al. (1961) Type VI structure. Neither structure can be more efficiently learned by selectively attending to specific features of the stimuli. Instead, the category structures need to be rote learned.

[1] The black borders on the stimuli in Figure 2.2 are used for explanation of the category structure, and were not part of the visual features presented to participants.

Twenty two participants completed the category learning task. The color and shape features were determined randomly for each participant. For example, the small category learned by one participant could be formed by the green square, blue triangle and red circle, as in Figure 2.2, but for another participant by the green circle, blue square, and red triangle. The two categories were randomly assigned to one of the response option labels “X” or “Y”. Participants were not aware of the number of stimuli assigned to each category.

Stimuli were presented one at a time in a randomized order in blocks, with the constraint that two successive blocks contained exactly two presentations of each stimulus. When a stimulus was presented participants were required to give a categorization response, using the mouse to select one of the two response options within 5 seconds. Feedback was then provided for 3 seconds by showing the correct category label before continuing to the next trial. This process continued until participants reached a criterion of 36 consecutive correct responses, or until a total of 50 presentations of each stimulus.

We consider the behavior of only 12 participants. Five were removed due to errors in the coding of the response variables, and five were removed because they took many more trials than the others to reach criterion, suggesting they were not engaged in the learning task. The lines in Figure 2.3 present the smoothed running average of the proportion of correct responses of the remaining 12 participants. The markers represent the number of trials to criterion. The histogram shows the distribution of the total number of trials for the participants. It is clear that all 12 participants reached the learning criterion in 100 or fewer trials, suggesting that they are good candidates for measuring the dynamics of learning.

2.4.2 CHMM for the Lee & Navarro (2002) Experiment

Applying the CHMM framework to a specific task requires specifying the values of the hyperparameters $\boldsymbol{\alpha}$ in Equation 2.6, a and b in Equation 2.8, and d and g in Equation 2.13, as well

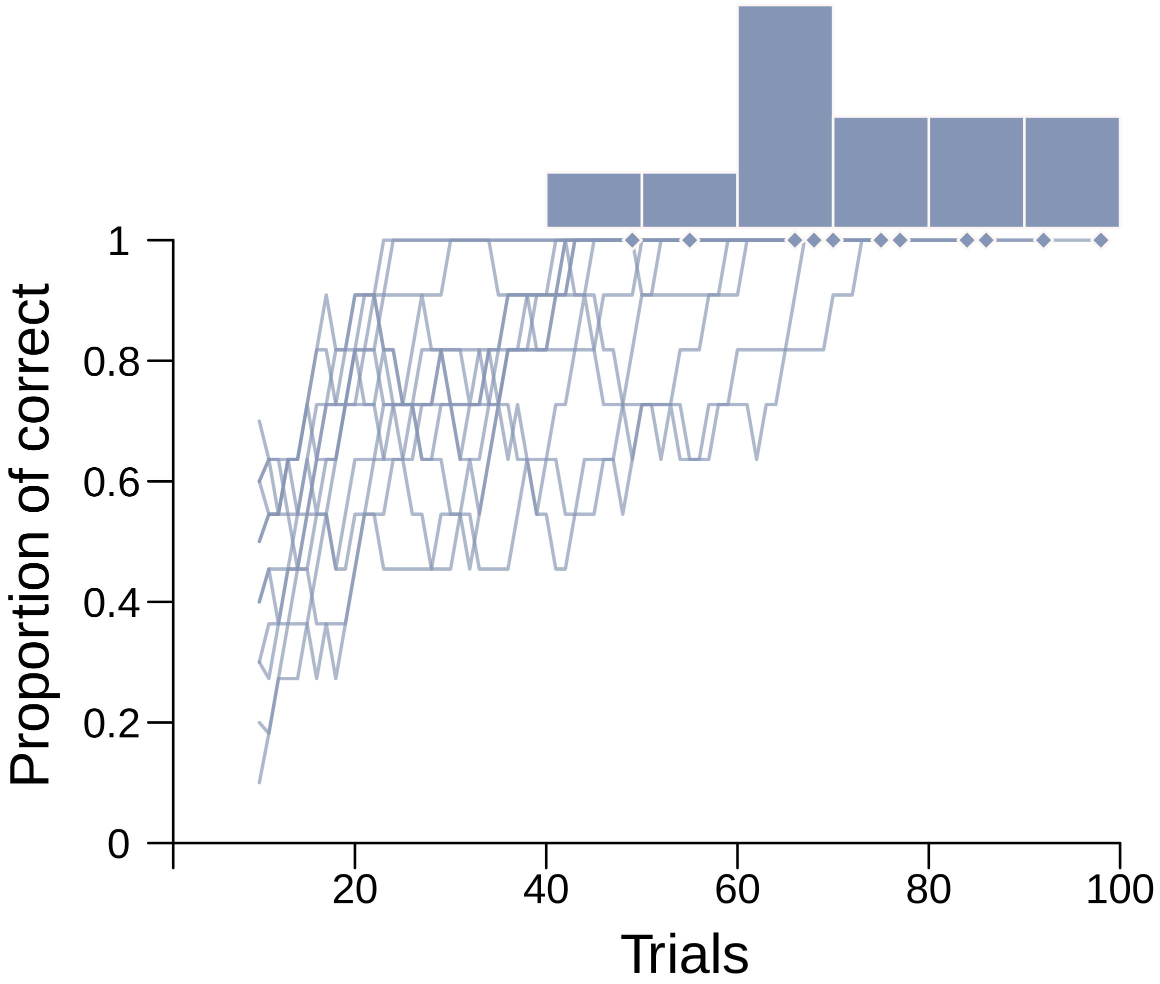

Figure 2.3: Learning performance of the 12 participants from Lee & Navarro (2002). The lines show the running average of the proportion of correct responses as a function of trials, smoothed by a window 10 trials wide. The markers show the number of trials to criterion for each participant, and the histogram shows the distribution of these total trials.

as defining the similarity function $s_{mm'}$ in Equation 2.9. The choice of these values defines the prior distribution of the initial probabilities in Equations 2.5, the stickiness parameter β_i in the transition function in Equation 2.7, and the trembling-hand error probability in Equation 2.12 respectively.

For the Lee & Navarro (2002) experiment, we assume that the hyperparameters for the initial probabilities are $\alpha_1 = \alpha_2 = 1$, so that any pattern of prior assignment of the stimuli to the categories is equally probable. This seems reasonable, since there is no prior knowledge about how simple shape stimuli belong to two arbitrary categories, and the participants do not have any prior information about one category having more stimuli than the other.

We assume that the hyperparameters for the transition function are $a = 2$ and $b = 1$. This is a vague prior because we do not have strong expectations about how willing participants are to change the latent category assignment of a stimulus based on the assignments of the other stimuli. This distribution does, however, allow for large values of β_i because it is possible in a simple artificial category learning task with fixed categories that participants may be resistant to changing an established category assignment.

We assume that the hyperparameters in Equation 2.13 are $d = 10$ and $g = 888$. These values capture the assumption that trembling-hand errors are unlikely in the setting of a simple cognitive task. The exact values were calculated so that that a participant's response is expected to misrepresent the underlying category assignment on very close to 1% of trials, and on no more than about 5% of trials in an extreme case.

Because Lee & Navarro (2002) used colored shape stimuli made up of simple nominal features, we assume that the similarity between two stimuli depends on the feature distance between them. Specifically, we use the similarity function introduced by Lee & Navarro (2002, see Equation 12, pp. 54), based on Tversky's (1977) contrast model of categorization. The similarity between stimulus m and m' is

$$s_{mm'} = \exp\left\{-\left(\sum_k f_k^m(1 - f_k^{m'}) + \sum_k (1 - f_k^m) f_k^{m'}\right)\right\}, \tag{2.14}$$

where f_k^m is a feature indicator defined as $f_k^m = 1$ if stimulus m has feature k and $f_k^m = 0$ if it does not. This means that the similarity between two stimuli decays as a function of the number of non-shared features between them. No selective attention or feature salience mechanisms are incorporated in the similarity model, consistent with the Type IV category structure not being determined by a subset of stimulus features. The transition probability function for the two-category case in Equation 2.11 was used.

The model was applied to all 12 participants independently, using a single chain that collected 5000 samples after 10,000 discarded burn-in samples used for the adaptation of the Hamiltonian Monte Carlo algorithm. We checked convergence using the method developed by Geweke (1992) and visual inspection of the chains for each parameter.

2.4.3 Descriptive Adequacy

We first examined the descriptive adequacy of the model. This is a logical first step, because if a model cannot pass the basic test of being able to re-describe data which it has seen, then it is unclear that the model's inferences are meaningful. Descriptive adequacy was examined by posterior predictive checking, which compares the agreement between the posterior predictive distribution and participants' responses (Gelman et al., 2004). The posterior predictive distribution indicates the probability of an outcome $\hat{\boldsymbol{y}}$, and is found by integrating the information contained in the posterior distribution of the model's parameters $\boldsymbol{\theta}$. Intuitively, the posterior predictive distribution measures the ability of the model to re-describe the data it used to make posterior inferences about parameters. Formally, it can be calculated as

$$\overbrace{p\left(\hat{\boldsymbol{y}} \mid M\right)}^{\text{posterior predictive}} = \int_{\Theta} \overbrace{p\left(\hat{\boldsymbol{y}} \mid \boldsymbol{\theta}, M\right)}^{\text{likelihood}} \overbrace{p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, M\right)}^{\text{posterior}} \, \mathrm{d}\boldsymbol{\theta}. \tag{2.15}$$

For the CHMM model, sampling from the posterior predictive distribution can be achieved by taking the posterior samples of the hidden process $\boldsymbol{X}_{1:T}^{1:M}$ and the trembling-hand probability ϵ and substituting them into Equation 2.12. This is a consequence of the conditional independence assumption described in Equation 2.3 which states that, conditional on the current sample of $\boldsymbol{X}_t^m$ and the trembling-hand parameter ϵ, the observed label response Y_t^m

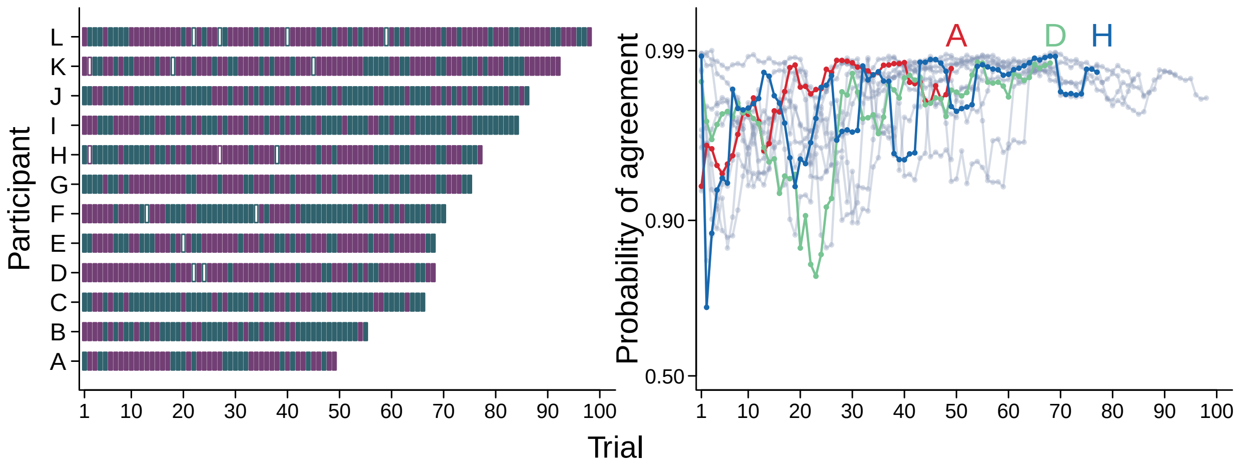

Figure 2.4: Tests of descriptive adequacy. The left panel shows the agreement between the posterior predictive distribution and participant responses. Each row represents the label choices and model predictions for a single participant trial by trial. Filled rectangles represent trials where the model's posterior predictive distribution agrees with the participant's choice with colors representing a different label response. Empty rectangles represent a lack of agreement. The right panel shows the running average of the probability of agreement between the posterior predictive distribution and participant responses for all 12 participants. Three participants labelled Participants A, D, and H with different levels of performance are highlighted.

is independent of other hidden processes and parameters.

Figure 2.4 shows the model's posterior adequacy for each participant and trial. Participant responses are organized in rows with each rectangle representing a different trial and each color a different label response. Filled rectangles represent the agreement between the mode of the posterior predictive distribution, and a participant response, while empty ones represent a lack of agreement. Overall, the modal posterior predicted response of the model describes 98% of the 1066 decisions participants made in the experiment. The greatest number of mismatched responses for any participant is four, and many participants are perfectly described.

The rare descriptive failures of the model tend to occur systematically in two places. The majority are located in the first half of a participant's trials. Because the model makes use of the whole sequence of responses to inform its inferences, this pattern of results could indicate higher levels of response variability at the beginning of the experiment or a learning process that is not accounted for. Descriptive failures are also more likely to occur for low frequency responses. For example, the mode of the posterior predictive distribution of participant D in Figure 2.4 does not agree with two responses for the yellow category. This labeling response is uncommon in the first half of the session, so the model considers the response to be the result of a trembling-hand error rather than a change in the representation.

2.4.4 Category Assignments for Representative Individual Participants

The most useful inferences of the model occur at the level of individual participants, focusing on the dynamics of their assignment of stimuli to categories. We present detailed results from three participants labelled as Participants A, H, and D in the right panel of Figure 2.4, with different levels of learning performance.

Participant A

Figure 2.5 shows the category assignment inferences made by the model for the best-performed Participant A, who completed the task in 49 trials. The x-axis corresponds to the experimental trials and the y-axis corresponds to the stimuli. The squares indicate the trials at which the stimulus on the row below was presented to the participant, and are shaded according to their response. Black squares indicate that the participant chose the smaller highlighted category with three stimuli while white squares indicate they chose the larger category with the remaining six stimuli. For example, Participant A was presented with the red circle on trials 12, 22, 28, 32, 39, and 47. They categorized the stimulus as belonging to the larger category the first time it was presented, but subsequently always categorized it as belonging to the smaller category. Notice that the feedback participants received is

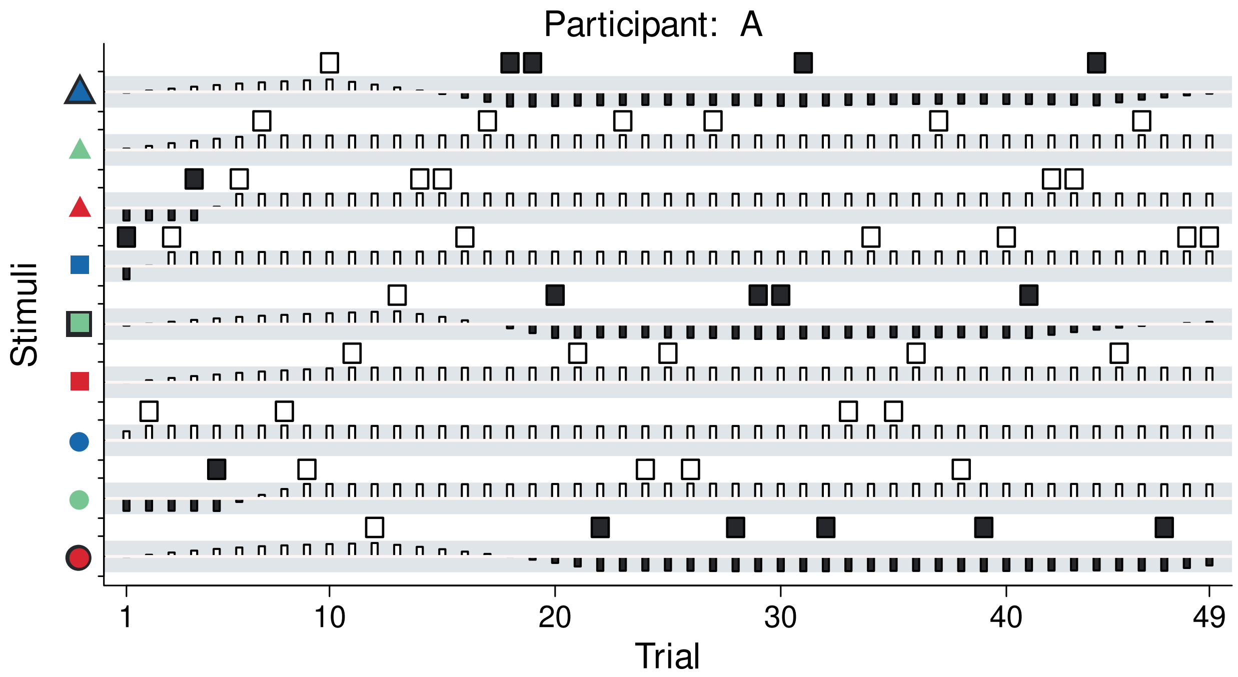

Figure 2.5: Category assignment inferences for Participant A from the Lee & Navarro (2002) experiment. The black and white squares indicate the trials at which the stimulus on the row below was presented. The shading indicates the participant's response: black squares indicate that the participant chose the smaller category, and white squares indicate they chose the larger category. The shaded bars on each row show the posterior probability of the category assignment for each stimulus on each trial.

not explicitly shown, but can be determined from the true category of each stimulus. For example, the red circle has a black border, and so belongs to the smaller category. This means that the participant received feedback that they were incorrect on trial 12 but that they were correct thereafter.

The black and white histograms in Figure 2.5 show the posterior mean of X_t^m for stimulus m on trial t. Because there are only two categories, the mean corresponds to the probability of the stimulus being assigned to the larger category. Upward white bars show posterior means greater than 0.5, while downward black bars show posterior means below 0.5. There is no bar shown if the posterior mean is exactly 0.5. Visually, this display format means that larger white bars indicate greater confidence that the stimulus is assigned to the larger category, while larger black bars indicate greater confidence that the stimulus is assigned to the smaller category. As a concrete example, consider again the red circle stimulus. For the first few trials, its latent category assignment is uncertain, but the probability that it belongs to the larger category increases before it is first presented on trial 12. After this first presentation, the latent assignment probability changes towards it more likely belonging to the smaller category. This assignment is near certain from the second presentation at trial 22 until its final presentation at trial 49. The certainty decreases slightly for the final two trials of the experiment.

Figure 2.5 demonstrates several key features of the CHMM analysis of this experiment. One feature is that the model infers that feedback can lead participants to change their category assignment suddenly. A good example is provided by the blue square presented on the first trial, for which the participant's response is incorrect. The inference is that the participant initially believed the stimulus likely belonged to the smaller category but quickly changed the assignment to the larger category. They then respond correctly when the blue square was presented again on trial 3.

Another feature of the analysis relates to the role of other stimuli in influencing the assignment of a stimulus. A good example is provided by the red circle discussed earlier. The model infers a gradual increase in certainty that the participant assigns it to the larger category leading up to its first presentation at trial 12. This inference arises because the model is also inferring that the participant believes many other stimuli that have already been presented with feedback—the green triangle, red triangle, blue square, blue circle, and green circle—belong to the larger category. The overall similarity of the red circle is accordingly greater to the larger category. For this category structure, which does not divide stimuli systematically into categories based on shape or color, the role of similarity can naturally be interpreted as a base-rate effect. The sensitivity to the number of stimuli in each category is also responsible for the decrease in certainty for smaller category stimuli without repeated responses that confirm this is what the participant believes. This is why the certainty of

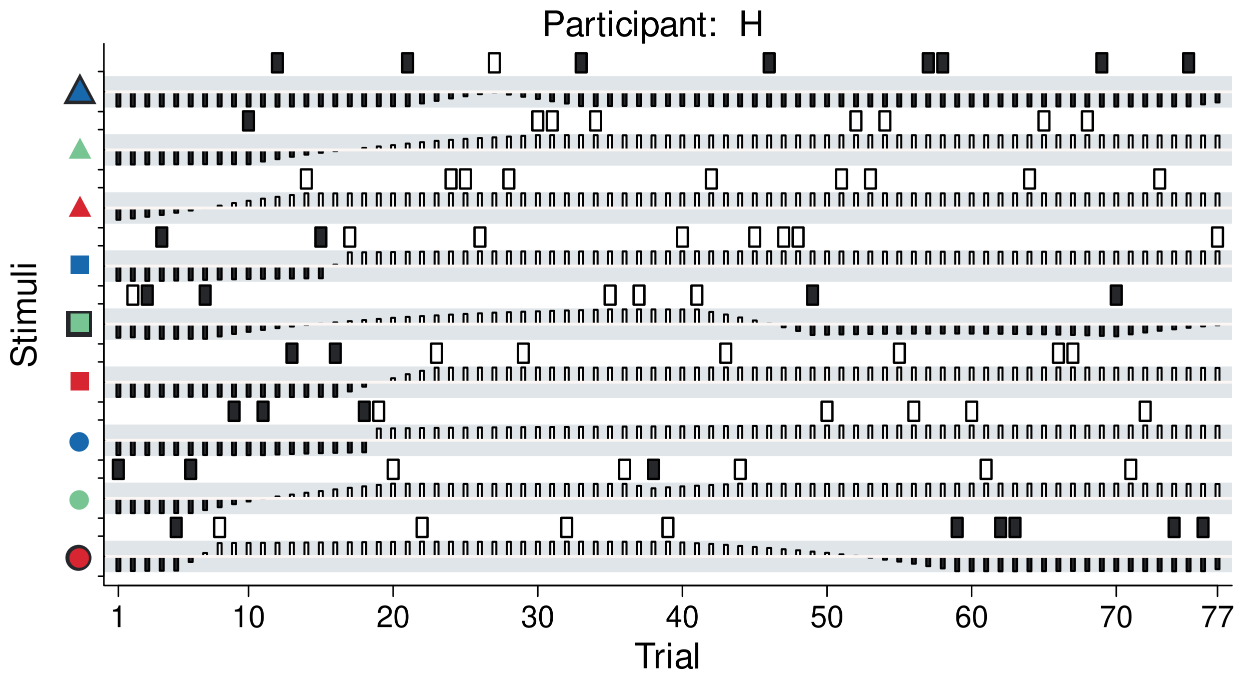

Figure 2.6: Category assignment inferences for Participant H from the Lee & Navarro (2002) experiment, using the same visual presentation as for Figure 2.5.

assignment for the blue triangle, green square, and red circle all decrease slowly after their final presentation in the experiment.

Participant H

Figure 2.6 shows the trial-by-trial category representation analysis for Participant H, who learned less quickly and required 79 trials to reach criterion. This analysis demonstrates several additional features of the CHMM model. It shows that the sensitivity to base-rate arising from stimulus similarity can interact with the sensitivity to learning from feedback. In particular, the model infers more rapid changes in category assignments from the smaller to the larger category. As an example, compare the transition of the assignment of the green square after its presentation on trial 7 to the transition of the red circle after trial 39. The first transition to the larger category occurs more quickly than the second transition to the smaller category. This example also underscores that the CHMM model is best understood as a measurement model. The green square transition is incorrect from the perspective of learning the category structure of the stimuli, but the modeling goal is not to learn categories. The goal is to understand the responses of a participant who is trying to learn the categories, and what their pattern of responses reveal about the dynamics of their latent representation of the categories. In order to explain Participant H's responses up to trial 42, the model infers that they switched the assignment of the green square to the larger category sometime after trial 7.

A final feature of the analysis shown in Figure 2.6 is that the model can account for some responses as trembling-hand errors. A good example is provided by the green square stimulus. The participant is presented with this stimulus consecutively on trials 2 and 3, making an incorrect categorization response the first time but correcting that response the second time. The model infers that this change is not due to learning, which would involve a shift in the underlying category assignment. Instead, it infers that the assignment was always to the smaller category, and that the first response simply did not reflect that underlying belief. The possibility of response error also interacts with the base rate. This is shown, for example, by the error the participant makes for the blue triangle on trial 27. The latent assignment remains primarily to the smaller category surrounding this trial, but the assignment is far less certain.

Collectively, the results for the three participants in Figure 2.5 and 2.6 show the ability of the model to provide insights into how the categories were learned, and the basis for the responses made. The analysis is at a fine level of resolution, showing trial-by-trial changes in category assignments, and the details of every response. It is also highly interpretable, focused on what category the participant believes the stimulus belong to, and whether that belief is accurately reflected in observed responses.

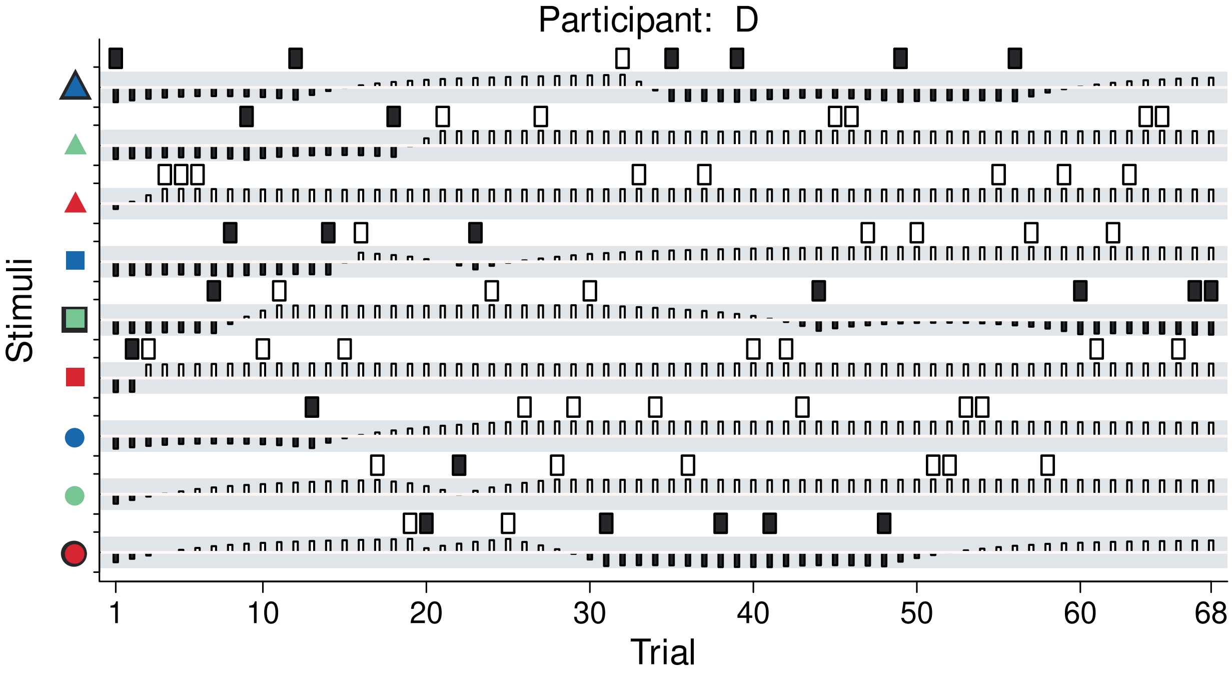

Figure 2.7: Category assignment inferences for Participant D from the Lee & Navarro (2002) experiment, using the same visual presentation as for Figure 2.5.

2.4.5 Inferences About Category Stickiness

Participant D

Figure 2.7 shows the trial-by-trial category representation analysis for Participant D, who took 68 trials to learn. An insight from their analysis involves the last presentation on trial 56 of the blue triangle, which belongs to the smaller category. The model infers that they become less confident of the assignment to the small category after this presentation, despite having responded correctly to this stimulus since trial 35. This inference arises because the participant's small category responses are less consistent than their large category responses. For example, large category responses are made for the red square consistently from trial 3 onwards, while responses to the red circle switch three times in the span of 10 trials. The model explains this pattern by inferring that the small category is not sticky in absolute terms, with posterior mean $\beta = 2.90$, and less sticky than the large category, with posterior

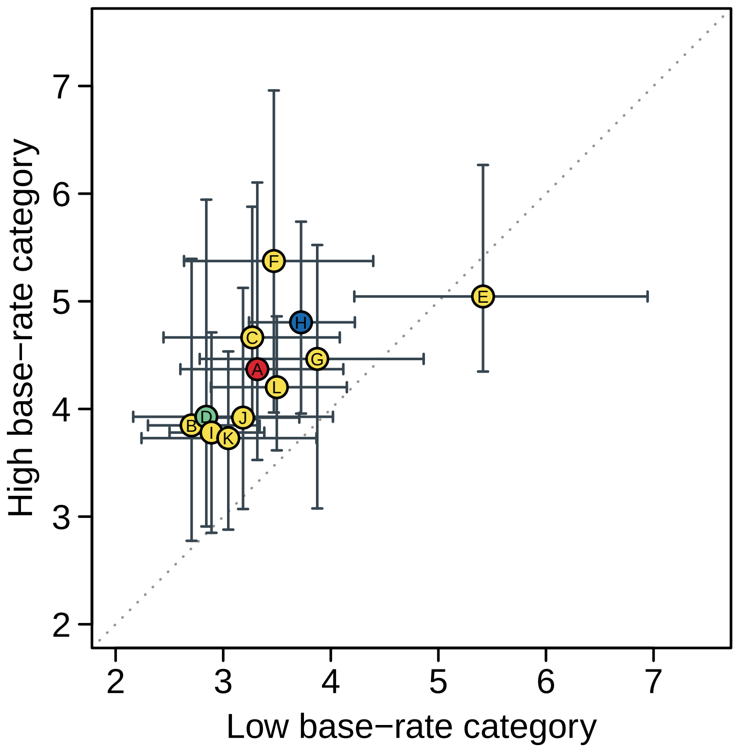

Figure 2.8: Means and 95% credible intervals of the joint posterior distribution of the stickiness parameters β_i for all participants, with Participants A, D, and H highlighted. Values on the x-axis correspond to the parameter associated to the small base-rate category, while values on the y-axis correspond to the large base-rate category.

mean $\beta = 3.95$. These stickiness values make it probable that category assignments can change from the small to the large category.

All Participants

Figure 2.8 shows the inferences about category stickiness for all participants. The labeled circles correspond to posterior means and the error bars show 95% credible intervals. For all but Participant E, the means are above the identity represented by the dashed line. This means that the probability that a stimulus will continue to be represented as part of the large category is higher than the probability of it staying in the small category. A possible explanation for the difference in the probability of switching category assignments is that the large category is easier to learn, most likely because participants are sensitive to the difference in base-rate.

2.5 Discussion

We developed a new framework based on coupled hidden Markov modeling for understanding the dynamics of categorization. The framework is not a model of how people learn categories. It is not designed to encounter a stimulus, produce a categorization response, receive feedback, and execute some sort of learning process, ready for the next trial. Instead, the framework is designed to allow measurement models to be defined for specific category learning tasks. The measurement approach means the model is supplied with information about what happens in an experiment—the sequences of stimuli a participant saw and the responses they made—and makes inferences about the underlying cognitive dynamics of the participant's learning.

To provide a formal structure for measurement, our framework does make cognitive assumptions about how people represent categories and how they use those representations to produce categorization responses. The core assumption is the one highlighted by the example of children playing games: we assume that people explicitly maintain an assignment of each stimulus to a category, and update those assignments for every stimulus after every trial. This means that the category assignment of a stimulus can change even if it is not presented. This assumption is different from many existing models of category learning that only indirectly account for the influence that learning about the current stimulus has on the assignments of others.

Our cognitive assumptions also differ from those made by the GCM measurement model (Nosofsky, 1986), which assumes participants know perfectly the category assignment of training stimuli, and measures the underlying category representation using transfer stimuli.

The CHMM framework allows for much richer dynamics in the learning process, making inferences about whether and how the category assignments of all stimuli change at any point during learning. The other assumptions our framework makes about category learning are more standard. We assume similarity-based generalization between stimuli, and allow for the possibility of response errors. It would be possible to develop the framework further to include other common assumptions, such as selective attention to relevant dimensions or features for learning specific category structures, or bias for different responses based on base-rate or payoff properties of a category learning environment.

We demonstrated the usefulness of the CHMM framework by developing a specific model for a simple category learning presented by Lee & Navarro (2002). The inferences of the model show the time course of latent category assignments for every stimulus on every trial. These inferences provide insights into how people learn, detailing how feedback about a stimulus can lead to sudden changes in its category assignment but also a more graded influence on the assignment of other stimuli. We think the analogy to EEG analysis is a good one to understand the usefulness of the CHMM framework. We think of the measurement process as taking all of the observable information in the whole category learning task, and processing it to provide a quantification of the underlying categorization dynamics.

The demonstration did not require a number of features of the framework that could be useful for other category learning tasks. The way stimulus similarity is defined allows the flexibility to tailor models to different stimulus domains. The Lee & Navarro (2002) stimuli seem best modeled in terms of features, but many category learning experiments involve stimuli best models in terms of underlying continuous features (e.g. Bartlema et al., 2014; Kruschke, 1993a; Nosofsky, 1986), and some require more sophisticated relational mechanisms for measuring similarity (e.g. Reheder & Ross, 2001; Schlegelmilch et al., 2022). Another capability of the model relates to stimulus domains like natural kinds for which people have relevant prior knowledge. This knowledge can be accommodated by informative priors in setting the initial probabilities for category assignments. The CHMM framework is also especially well suited to modeling category learning in changing environments. Some experiments involve pre-determined changes in the category membership of stimuli (e.g., Kruschke, 1993a). Other experiments involve changes that depend on the responses of participants (e.g., Villarreal et al., 2022). In both cases, the theoretical question of interest is how people learn to adapt to the new category structure. The CHMM framework offers the ability to make inferences about changes in latent category assignments that occur immediately following the first time the change is signalled through feedback.

Because category learning assignments are explicitly maintained for all stimuli at all times, the CHMM framework opens up possibilities for the design of category learning experiments that are not routinely considered in the modeling literature. A task structure used in studying the GCM (e.g., Bartlema et al., 2014; Nosofsky, 1986) involves training trials with feedback followed by transfer trials that use different stimuli and do not have feedback. This structure can be accommodated directly in the CHMM framework, which will allow inferences to be made about the category assignments of the transfer stimuli from the beginning of training. For the same reason, it would be straightforward to model experiments that present more than one stimulus on a learning trial (e.g. Ross & Murphy, 1999; Patterson & Kurtz, 2020). The CHMM framework automatically makes predictions about what the response would be for every stimulus on every trial, but in its application to a standard one-stimulus-at-a-time experiment is forced to treat most of these predicted responses as missing data. Experiments that presented more stimuli would provide more information about the underlying cognitive representations, and likely improve the acuity of measurement.

The CHMM framework is very flexible. This is both a strength and a weakness. The weakness is that, without significant additional cognitive assumptions, it is not a good cognitive model for predicting categorization behavior. It is capable of describing a wide range of observed behavior, making it difficult to falsify (Roberts & Pashler, 2000; Veksler et al., 2015;

Vanpaemel, 2020; Villarreal et al., 2023). The strength is that flexibility allows the CHMM framework to serve as an effective measurement model. The possible extensions to category learning situations involving prior knowledge, changing environments, or simultaneous encounters with multiple stimuli are specific benefits of this flexibility. The CHMM framework provides a way of measuring and understanding the dynamics of category learning based on the responses people make as they encounter stimuli. These measurements are based on interpretable and intuitive cognitive assumptions about the nature of category learning, involving stimulus-stimulus interactions that allow for the constant updating of category associations, and the separation of latent category knowledge from observable response processes. For these reasons, we believe the CHMM framework provides a new capability to construct task-specific measurement models that will help us understand the core cognitive capability of category learning.

2.6 Publication Note

The current version of this chapter has been submitted and is currently under review at the *Journal of Mathematical Psychology*.

Chapter 3

Generalization and Transfer Tests of the CHMM

Abstract

We present an extension to the Coupled Hidden Markov Model (CHMM) framework introduced in Chapter 2. This extension aims to model the weight of between stimulus similarity and the impact of generalization in the dynamics of categorization. We implement and test the predictions of this extension using data from two previously published studies. In the first study reported by Navarro et al. (2005) participants had to classify real world face stimuli into two categories according to of four different category structures. Results show that the CHMM approach is able to adequately describe participant's labeling responses in the task. To test the predictions of the model we removed four stimuli from the data of all participants. We compare the predictions of the CHMM approach to these stimuli with the predictions of the Generalized Context Model (GCM) of categorization (Nosofsky, 1988). This comparison showed that the predictive accuracy of the CHMM approach can

match and sometimes outperform the GCM. In a second study reported by Bartlema et al. (2014) participants had to classify artificial stimuli constructed from the combination of two continuous features. The design of the experiment presented participants with learning and transfer blocks where only a subset of stimuli where presented during learning. We leverage this design in order to test the predictions of the CHMM framework without the need to artificially remove stimuli from the analysis. Results showed that the predictive accuracy of the CHMM approach again matched the GCM at the participant-stimulus level, and that it outperforms the former when averaging predictive accuracy across stimuli. We conclude by presenting some future avenues for the development of the CHMM framework.

3.1 Introduction

Imagine yourself in the following situation. A friend has invited you to a dinner party, you do not know exactly what will be served but you have been told that it will be Mexican food. In order to choose if you will attend the party you want to know if you would like the food or not. How can you make your decision? You can solve this problem by thinking about your past experience with Mexican food and then extrapolating from them to decide if you will enjoy it or not. In other words, if you have enjoyed Mexican food in the past you will probably enjoy having it again and should accept the invitation. This solution to the problem involves a process known as generalization (Shepard, 1987), where what has been learned about a stimulus in the past is used to evaluate future observations.

Generalization has been proposed as the core cognitive capability (Shepard, 1987) and categorization models have made use of this principle. In particular, similarity-based categorization and category learning models assume that what is learned about a stimulus generalizes to other stimuli in a magnitude that is proportional to the inverse of their distance in psychological space (e.g. Nosofsky, 1986; Kruschke, 1992; Smith & Minda, 1998, 2000). Another way to think about this is that what is leaned about a stimulus will affect other stimuli that are similar to it. These type of models are part of a family known as similarity-based models.

Similarity-based generalization gives categorization models the ability to make predictions about the category assignments of previously unseen stimuli in an experiment. Thus, a common approach in the literature to compare categorization models has been to use a transfer designs (e.g. Kruschke, 1993a; Nosofsky, 1986). Participants in a transfer design categorization task first are exposed to a subset of stimuli that are assumed to share some psychological space. Then, after learning the underlying category structure of the task they are faced with a second subset of stimuli that they have to classify. The key intuition is that different models will make different predictions about the category assignments of these previously unseen stimuli.

In Chapter 2 we introduced the Coupled Hidden Markov Model (CHMM) framework as a way to measure the dynamics of categorization. The results of the application of the CHMM showed that the model was adequately able to describe the observed data in a categorization task in which stimulus similarity played only a small role in the underlying category structure. In that application, the unequal sizes of the categories played a more important role, because the category assignments did not depend on stimulus similarity. The CHMM framework, however, assumes that the dynamics of the categorical representation of stimuli in a task depend on both the similarity between stimuli and the base-rate of the categories. While the application showed some important results from the model and allowed us to highlight some of its properties, the small role that stimulus similarity plays in the underlying category structure of the task makes it less useful when one wants to evaluate the predictions of the model that depend on similarity.

As with other similarity-based models of categorization, the CHMM can be adapted to test transfer stimulus predictions in a categorization task. To this end, we first introduce a parameter that measures the effect of generalization. As previously mentioned, the principle of generalization allows categorization models to make predictions about the category representations of items that have not yet been encountered. Secondly, we introduce a parameter that measures the weight that similarity has on the dynamics of categorization for a given task. Both of these changes can be implemented directly in the transition function of the model, which controls the dynamics of the categorical representation of items according to the CHMM.

Furthermore, the addition of similarity-based generalization to the CHMM framework enables testing the model's predictions using out of sample data, consistent with the sort of testing a transfer experimental task tries to achieve by design. More specifically, we test the predictions of a model by first restricting its access to a subset of stimuli and then comparing its predictions with participants' behavior. This approach has the advantage of allowing us to test a model's predictive accuracy in experimental designs that do not include a transfer phase.

In this chapter we present the results of the application of the CHMM framework to two experiments in the literature. The first one is a category learning task with real-world face stimuli (Navarro et al., 2005). The second is a learning-transfer task, in which participants learn a category structure using only a subset of items before categorizing all of the stimuli in a transfer phase. In the first section we summarize the relevant main intuitions of the CHMM framework including the extensions to support similarity-based generalization. In the second section we describe the Navarro et al. (2005) experimental design, present some behavioral results, and compare the predictions of the CHMM approach to the Generalized Context Model (GCM: Nosofsky, 1986) of categorization. In the third section, we introduce the learning-transfer experiment reported by Bartlema et al. (2014), present some behavioral results, and compare the predictions of the CHMM and GCM in the transfer phase of the experiment. We finish with a discussion of the limitations of the CHMM framework and possible avenues for development.

3.2 Coupled Hidden Markov Models

In Chapter 2 we introduced the Coupled Hidden Markov Model (CHMM) approach to measuring the dynamics of categorization. In this framework the categorical representation of a stimulus and a participant's labeling choices are represented as part of a chain with two components. The first one is a hidden process that denotes the state or category that the participant believes the stimulus belongs to. The second one is an observed process of the labels assigned by the participant to the presented stimulus at any given trial. These two processes are linked together by a response function, which formalizes the probability that a given label will be assigned to a stimulus conditional on its categorical representation. Furthermore, in the CHMM framework, the hidden components of the chains are allowed to interact with one another. In the context of categorization, this means that the category that a stimulus is assigned to will affect the representation of all other stimuli in a task. These stimulus-stimulus interactions are instantiated in the transition probability function.

Figure 3.1 shows a graphical representation of a CHMM for a categorization task with three stimuli. The unshaded nodes represent the hidden process of the chain, in this case the category that the stimulus is assigned to at trial t. The shaded nodes represent the observed process or the participant's labeling choice in a given trial. Note that not all hidden nodes are associated with an observed node. This means that in some trials we only observe a single response, for example, in trial t the participant is presented with stimulus two and gives a response. This follows the structure of most categorization tasks, in which participants are presented with a single stimulus to label on each trial. Finally, the arrows represent the dependency between the variables in the model. Note that all hidden process are connected by an arrow with all previous ones, but the responses are only connected to their respective hidden process. This is known as a conditional independence, and it means that once we know which category the participant believes a stimulus to be in, the response to that stimulus is independent of the representation of all other stimuli. In other words it is only

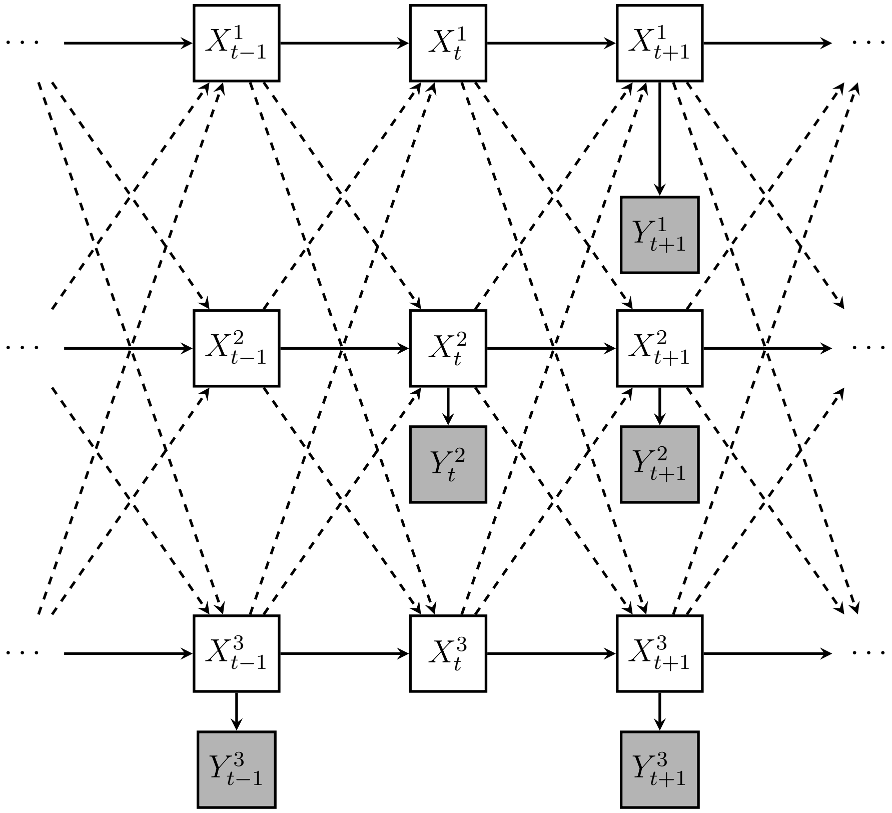

Figure 3.1: Graphical representation of a Coupled Hidden Markov Model. Shaded nodes represent the observed process, unshaded nodes represent the hidden process. Solid and dashed arrows represent the dependency structure of the model.

the representation of the stimulus that determines the response to it.

The experiments introduced in the following sections all have two category options labeled "A" and "B" and two response options "a" and "b". Therefore, to implement the CHMM framework we need to specify an initial probability that defines how likely it is that a stimulus would be assigned to each of the two categories at the start of the experiment. We can express this initial probability as:

$$X_1^m \sim \text{Bernoulli}\left(\gamma\right), \tag{3.1}$$

which specifies that at the start of the experiment, any stimulus will be assigned to category "B" with probability γ. We also need a response function that links the categorical representation of the stimulus to a participant's choice. We use the trembling hand response function introduced in Chapter 2:

$$P\big(Y_t^m = b \mid X_t^m = x_t^m\big) \sim \text{Bernoulli}\left(\epsilon \; \mathbf{1}(x_t^m = A) + (1-\epsilon) \; \mathbf{1}(x_t^m = B)\right), \tag{3.2}$$

where ϵ is the probability of responding with label "b" to a stimulus currently assigned to category "A".

Finally we need to specify a transition probability function. As noted above, it is in this function that we can introduce the parameter that measures the effect of generalization, and a parameter that measures the weight that similarity has on the dynamics of categorization. To this end, we propose the following transition probability function:

$$P\big(X_t^m = x_t^m \mid \boldsymbol{X}_{t-1}^{1:M} = \boldsymbol{x}_{t-1}^{1:M},\ \boldsymbol{\beta}\big) = \begin{cases} \text{logit}^{-1}\left(\alpha_i - \beta\ \eta_{B,A}^m(t-1)\right) & \text{if } x_{t-1}^m = A \text{ and } x_t^m = A \\ 1 - \text{logit}^{-1}\left(\alpha_i - \beta\ \eta_{B,A}^m(t-1)\right) & \text{if } x_{t-1}^m = A \text{ and } x_t^m = B \\ 1 - \text{logit}^{-1}\left(\alpha_i + \beta\ \eta_{B,A}^m(t-1)\right) & \text{if } x_{t-1}^m = B \text{ and } x_t^m = A \\ \text{logit}^{-1}\left(\alpha_i + \beta\ \eta_{B,A}^m(t-1)\right) & \text{if } x_{t-1}^m = B \text{ and } x_t^m = B, \end{cases} \tag{3.3}$$

where the parameter α_i represents the "stickiness" or the consistency with which a stimulus is assigned to category i across trials. The parameter β represents the weight of similarity in the dynamics of categorization. As the value of this parameter increases, stimuli are more likely to switch to or stay in the category that has items with the higher total similarity.

The function $\eta_{B,A}^m$ represents the total similarity of stimulus m to stimuli assigned to category "B" at trial $t-1$ relative to those assigned to category "A". This function models the interactions between stimuli in the task. We propose a total similarity function of the following form:

$$\eta_{B,A}^{m}(t-1) = \sum_{m' \neq m} \mathbf{1}(x_{t-1}^{m'} = B)\; e^{-\sigma d_{m,m'}} - \mathbf{1}(x_{t-1}^{m'} = A)\; e^{-\sigma d_{m,m'}}, \tag{3.4}$$

where $\mathbf{1}(x_{t-1}^{m'})$ is the indicator function, and $d_{m,m'}$ represents the distance between stimulus m and m' in psychological space. Considering that we have a continuous representation of the psychological space of the stimuli in both experiments, we decided to use the Euclidean distance as a metric. Finally, the parameter σ in Equation 3.4 measures the effect of generalization on a categorization task. As the value of the parameter increases, similarity decays faster as a function of distance. This means that what is learned about a stimulus would only be generalized to those items that are near in psychological space. As the value of the parameter decreases the slower similarity decays with distance, and therefore, what is learned about the same stimuli generalizes to more dissimilar stimuli.

3.3 Out-of-Sample data Navarro et al. (2005)

Navarro et al. (2005) present data from a between-participants category learning task using real world face stimuli with four conditions. Figure 3.2 shows the stimuli that participants had to categorize during the experiment. The stimuli are arranged using non-metric multidimensional scaling as a visualization method (Kruskal, 1964) obtained from previous similarity ratings (O'Doherty & Lee, 2002).

The main objective of the experiment was to test how learning a particular category structure could influence a participant's stimulus similarity judgements, through the phenomenon of learned categorical perception (Goldstone & Hendrickson, 2010). Conditions in the task where defined by different category structures, each with an increasing level of abstraction. Forty people participated in the experiment, ten in each condition. We focus on three of the

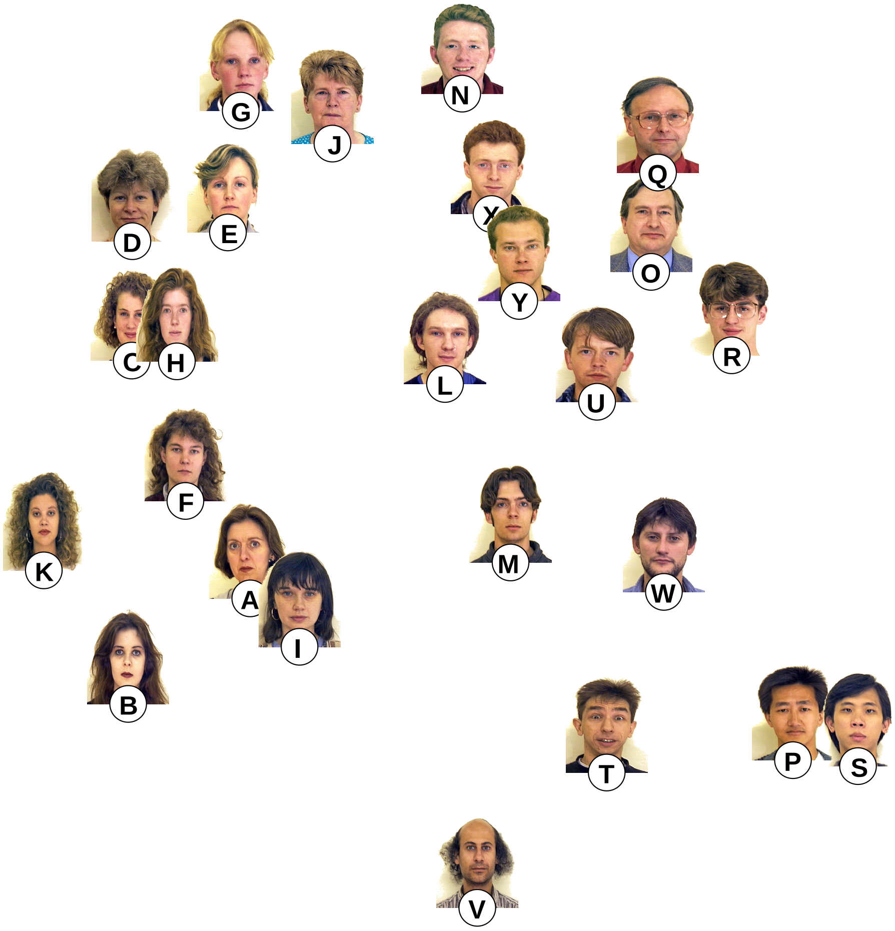

Figure 3.2: Graphical representation of the 25 stimuli in Navarro et al. (2005). Stimuli are arranged such that more similar faces are located near each other.

four conditions in the original experiment, involving learning category structures based on the gender, hair and trustworthiness of the face stimuli.

The three conditions presented were divided into 8 learning blocks. Each block consisted of a single presentation of each of the 25 stimuli. The presentation order was randomized for each participant. On each trial, a participant was presented with a face to label. Once a label response was produced, they received feedback on the accuracy of their label according to the category structure corresponding to their assigned condition. The next trial began after the presentation of feedback until all 8 blocks had been presented. This resulted in a total of 200 trials for each participant in their condition.

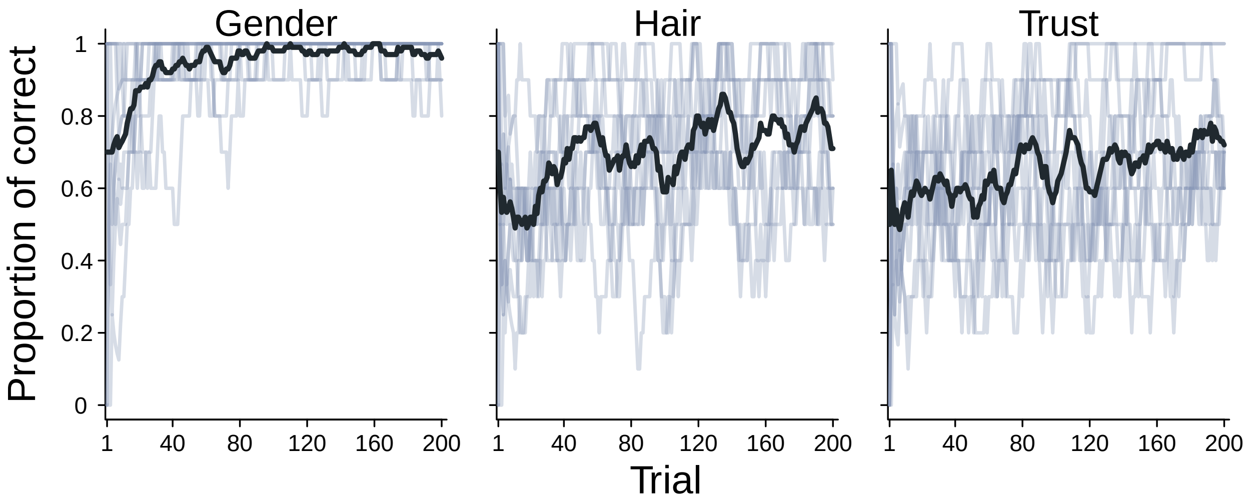

Figure 3.3: Learning performance of the 10 participants in the gender, hair color, and trustworthiness conditions from Navarro et al. (2005). From left to right, the panels show participants' performance in the respective conditions. Gray lines represent an individual participant's running average of the proportion of correct responses, smoothed by a window 10 trials wide. The black lines represent the average of the smoothed proportions across participants.

3.3.1 Results

Figure 3.3 shows the running average of the proportion of correct responses as a function of the trial for the three chosen conditions from Navarro et al. (2005). Each panels shows a different condition. Gray lines represent the running average of individual participants, smoothed by a window 10 trials wide. The black line in shows the average of the smoothed proportions across participants. In the gender condition, most participants show a steep learning curve at the beginning of the task, with almost all of them reaching an accuracy around 80% or higher before the end of the second block at trial 50. In the hair condition, participant accuracy shows more variability. There is still, however, some evidence of learning of this category structure, at least in the first two blocks. In the trustworthiness condition, participants' accuracy at the individual level does not show any clear indication that learning of the category structure is taking place.

The black lines on the panels of Figure 3.3 show the average of the smoothed proportion of correct responses across participants. For the gender condition, the group accuracy seems to reach levels around 90% after the first two blocks and remains consistently at that level. In contrast, in the hair condition there is some increase in group accuracy after the first two blocks and a smaller subsequent increase. In the trustworthiness condition learning at the group level is slower. In comparison with the other two groups, there is no abrupt change in the average in the initial portion of the task. However, there seems to be a small increasing trend, which suggest that at least some participants might be learning the category structure although in a longer time scale in comparison with the other two conditions.

It is worth noting that, from the three category structures, the trustworthiness is the most abstract one. While both gender and hair categories are easy to define in terms of perceptual features, the trustworthiness category is much more loosely defined. Face stimuli that had features such as large eyes, large forehead, small chin, softer face shape, high eyebrows, and a smiling expression being assigned to the trustworthy category. The combinations of these features might even be hard to detect for some participants, and thus it is not surprising that learning is slower.

3.3.2 Descriptive Adequacy

Figure 3.4 show the agreement between the posterior predictive distribution of the CHMM and participant's responses in the gender condition. In the top panel participants are organized by rows, with each rectangle representing a trial, and colors indicating a different label response. Filled rectangles represent trials in which the mean of the model's posterior predictive distribution agrees with the participant's label response. Unfilled rectangles represent a lack of agreement. Overall, the model can adequately describe 99% of participants responses. Furthermore, the model's mean prediction disagrees with at most seven responses of one participant, and perfectly describes the label responses of three participants.

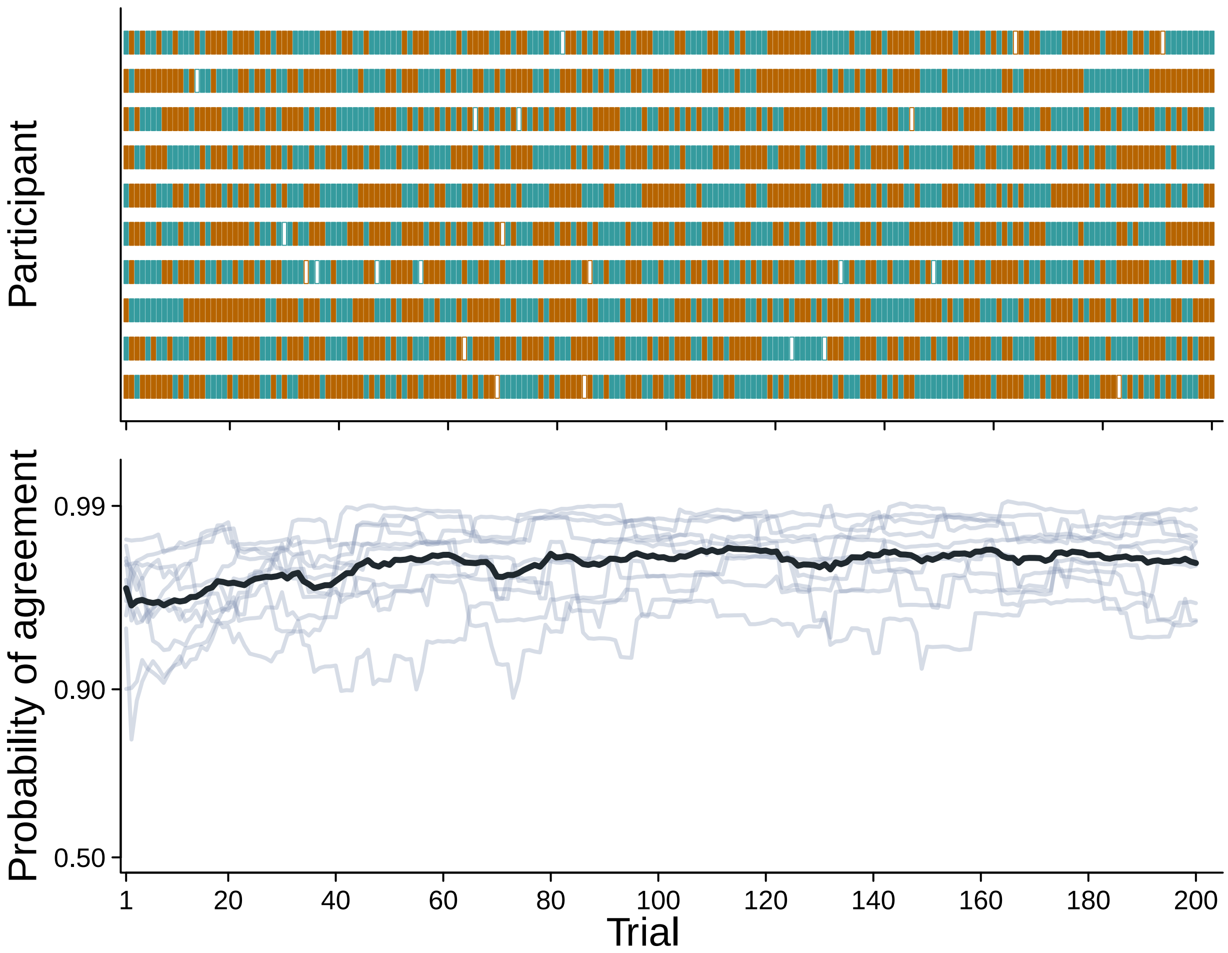

Figure 3.4: Tests of descriptive adequacy for Navarro et al. (2005) data. The top panel shows the agreement between the mean of the posterior predictive distribution and participant's responses. Participants are organized in rows. Each rectangle represents a trial, and colors represent a different label response. Filled rectangles indicate an agreement between the model's mean posterior prediction and the participant's response. Empty rectangles indicate a lack of agreement. The bottom panel shows the running average of the probability of agreement between the mean of the model's posterior predictive distribution and participant's responses for all 10 participants, smoothed by a window 10 trials wide. The black line represents the average of the smoothed agreement probabilities across participants.

The bottom panel of Figure 3.4 shows the running average of the agreement probability between the model's mean posterior prediction participant's responses, smoothed with a window of size 10 trials wide. Each gray line represents a different participant while the black line represents the average of the smoothed probabilities across participants. Note that even though we observe a rapid change in participants proportion of correct responses in the previous section, the model is able to account for those changes. Additionally, the running average was consistently above 90% for all participants. This indicates that the model's predictions assign a probability of at least 0.9 to participant's trial by trial label choices.

3.3.3 Posterior Prediction: Missing Stimuli

One of the main problems with the posterior predictive adequacy approach to model evaluation is that models have access to the data before making a "prediction". In this sense, these methods are better understood as a measure of the ability of the model to describe what it has already seen. A failure to do so could indicate that there is something wrong with the model. However, a high descriptive adequacy is a sign only that the model is capable of reproducing the information used to make the current inference.

This is particularly true in the case of the CHMM approach. As discussed in Chapter 2, the CHMM uses all of the information available in order to make a prediction. The model has access to a participant's future responses when making its predictions about how a stimulus will be represented in the following trials. A more useful way to evaluate the predictions of a model that does not suffer from these problems is the prediction of out of sample observations. As its name suggests, this means making predictions about samples that have not been used to inform the inferences in a model. In the case of Navarro et al. (2005) experiment, one way to do this is to prevent the CHMM from accessing participant responses to a given stimulus in the experiment.

Figure 3.5 shows the posterior mean of the category assignments for Participant 8 in the gender condition for all 25 face stimuli. The x-axis corresponds to the experimental trials, while the y-axis corresponds to the stimuli. The squares indicate trials in which the stimulus in the row below was presented and are shaded according to the participant's response. Black

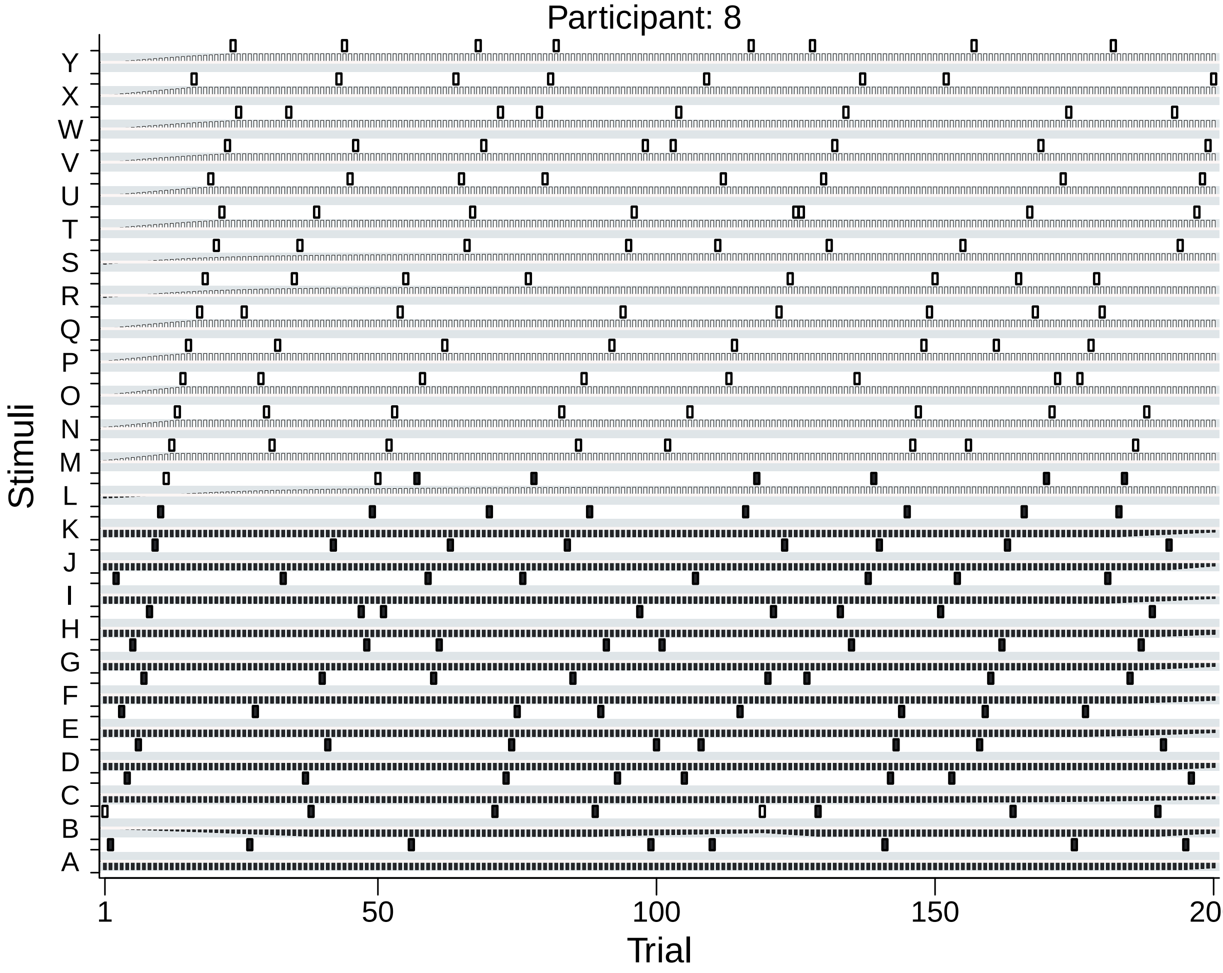

Figure 3.5: Category assignment inferences for Participant 8 in the gender condition of Navarro et al. (2005) experiment. The black and white squares indicate the trials at which the stimulus on the row below was presented. The shading indicates the participant's response: black squares indicate that the participant chose the Female category, and white squares indicate they chose the Male category. The shaded bars on each row show the posterior probability of the category assignment for each stimulus on each trial. Stimuli C, L, R, and S have been removed from the data used in the inference process.

squares indicate that the participant assigned the stimulus to the "female" category. The histograms located in the shaded regions show the posterior mean of X_t^m for stimulus m on trial t. Upward white bars indicate a higher confidence that the stimulus is assigned to the "male" category, while downward "black" bars indicate a higher confidence that the stimulus is assigned to the "female" category. A clear pattern in the graph is that the CHMM has high confidence that stimuli A to K are assigned to the "female" category, while stimuli from

M to Y are assigned to the “male” category.

One of the advantages of the CHMM framework is that the category assignment of missing stimuli can be tracked throughout the experiment. In order to test the predictions of the model we removed all of the participants responses to four faces: C, L, R, and S. These are treated as out-of-sample stimuli for which predictions can be made. Starting with face C, the model predicts it will be assigned to the “female” category, and is confident about this assignment from the beginning of the task. In contrast, for faces R and S the probability of assignment to the “male” category increases during the first 10 to 15 trials, and the model remains confident on this assignment thereafter.

Face L is a special case. Figure 3.2 makes clear that this face is represented as being closer to the “male” category. This similarity leads the model to assign this face with confidence to the “male” category. However, after two errors at the start of the task, the participant learns that this stimulus belongs to the “female” category and labels it as such for the remainder of the experiment. The same pattern of prediction and response were observed in the majority of participants in the gender condition. There are several possible reasons for the model’s failure of prediction for this face. One possibility is that the failure is a consequence of the lack of a feedback system in the model. Another possibility is the assumption that only between-stimulus similarity can affect the dynamics of categorization.

3.3.4 Out-of-Sample Predictive Accuracy

Figure 3.6 shows the out-of-sample predictive accuracy of the CHMM and the Generalized Context Model (GCM) of categorization (Nosofsky, 1988). Different types of data aggregation are organized in rows. The first row shows the predictive accuracy for each stimulus and participant. The second row shows the predictive accuracy aggregated across participants. The last row shows data aggregated across the four face stimuli removed from the inference

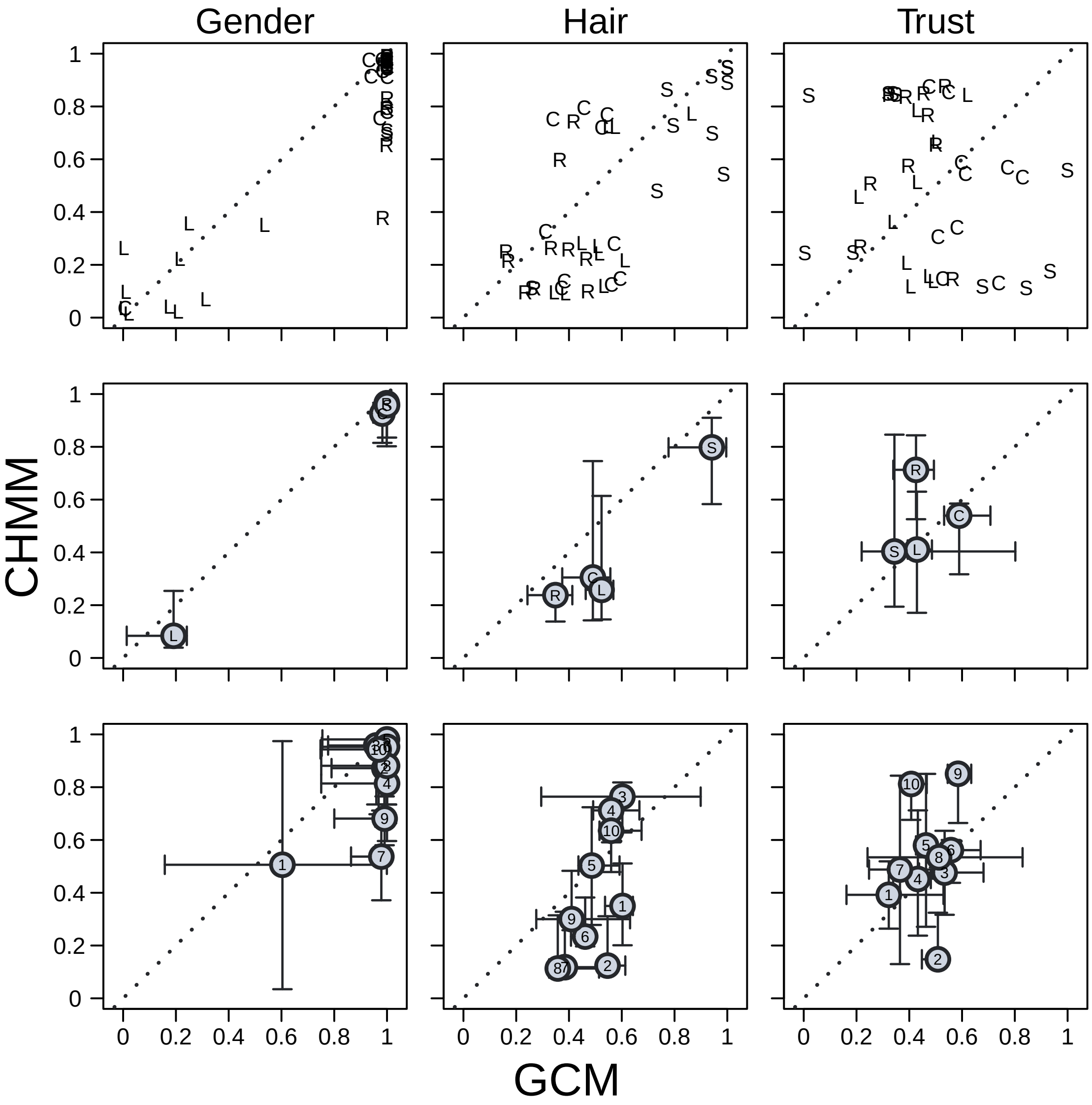

Figure 3.6: Out-of-sample predictive accuracy of the CHMM and GCM of categorization on the three conditions from Navarro et al. (2005). Different levels of data aggregation are organized in rows, while the conditions are organized in columns. The first row of panels show the predictive accuracy for each face stimulus and each participant. The second row aggregates the face stimulus predictive accuracy over participants. The last row aggregates the individual accuracy over the four withheld face stimuli. The x-axis indicates the predicted probability assigned by the GCM to the participant responses for the last trial in which the stimulus was presented. The y-axis indicates the predicted probability assigned by the CHMM. Each point in a graph represents the median while the error bars represent the 25 and 75% quantiles.

process. The columns organize each of the three conditions that we considered from the experiment of Navarro et al. (2005).

The first panel in the first row of Figure 3.6 presents the predictive accuracy of the models in the gender condition. Both models seem to be able to predict participant responses accurately for faces C and S, but both do not predict responses for face L. This face presents a "borderline" case: its representation in psychological space is closer to stimuli that belong to the opposite category. Given that both models struggle with predicting responses to this item, this suggests that a variable beyond just stimulus similarity is influencing how participants label this particular stimulus. This additional variable is likely the feedback provided to participants in the experiment. Because the stimulus is removed from the analysis neither of the models has access to this information.

The second panel in the first row presents the predictive accuracy in the hair color condition. There is more variability in the accuracy of the GCM model. Given that the GCM assumes that participants know the underlying category structure, the increase in variability is to be expected, since participants don't reach the same level of accuracy after the 200 trials in the experiment compared to the gender condition. The third panel in the top row presents the predictive accuracy in the trustworthiness condition. The predictions of the GCM show even more variability. In contrast, the predictions of the CHMM are more constrained.

The second row of panels in Figure 3.6 presents the data aggregated over participants. Each point in a graph represents the median of the models posterior predictive accuracy, and the whiskers represent the 25% and 75% quantiles. These panels show that the predictive accuracy of both models in the gender condition is close to the identity line. This means that on average both models assign similar probabilities to participant's last response to the stimulus. In comparison, GCM has a higher accuracy in the hair color condition, since most of the medians can be found under the identity function. In contrast, the CHMM generally makes better predictions in the trustworthiness condition.

The last row of panels in Figure 3.6 presents the models' predictive accuracy aggregated over stimuli, to focus on the differences between participants. Each point represents the median predictive accuracy of the models for each participant. One again, the GCM predictions generally outperform the CHMM in the gender and the hair color conditions, although some participants responses are better predicted by the CHMM in the later. In comparison, the out-of-sample predictions of the CHMM are more accurate for seven out of the ten participants in the trustworthiness condition.

All of the results in Figure 3.6 can be understood in terms of the different assumptions the GCM and CHMM make about how people represent the category structures they are learning. As the category structure becomes more abstract and difficult to learn, the more frequently the predictions of the CHMM approach outperform those of the GCM. The natural explanation for this finding is that the GCM places a strong inductive bias on the knowledge participants have about the category. It assumes people know with certainty the correct category assignment for all the faces for which feedback has been provided. This is an effective inductive bias when people are able to learn the category structure well, and so the GCM makes good prediction in these situations. The CHMM, in contrast, allows for each participant to have any possible pattern of category assignments for the face stimuli. This flexibility corresponds to a much weaker inductive bias. A consequence is that the CHMM is effective when the strong assumptions of the GCM are poor ones: that is, when a participant's representation of the category structure is significantly different from the ground truth provided by feedback. In these situations, the CHMM is able to infer the mistaken beliefs of the participant, and make predictions about out-of-sample stimuli that are more likely to correspond to the participant's actual observed behavior.

3.4 Transfer Trials in Bartlema et al. (2014)

In this section we analyze data from a category learning experiment presented by Bartlema et al. (2014), in which participants had to classify 16 simple perceptual stimuli known as

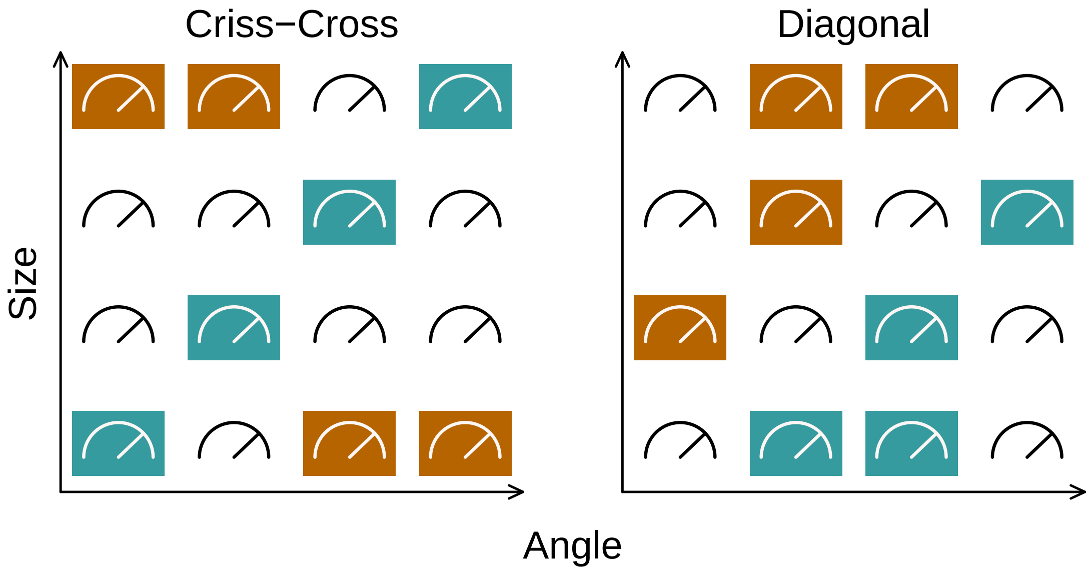

Figure 3.7: Crisscross and diagonal category structures in Bartlema et al. (2014), green and orange rectangles represent combinations of the angle and size of the radial line for stimuli assigned to categories "A" and "B" respectively.

"Shepard circles". The main objective of the study was to introduce the advantages of Bayesian hierarchical mixture models, especially in the context comparing exemplar and prototype based models of categorization. The experimental design consisted of two conditions defined by the "criss-cross" and "diagonal" category structures (Nosofsky, 1989) shown in Figure 3.7. Only analyses of data from the diagonal condition has been previously published (Bartlema et al., 2014).

A trial in the experiment consisted of the presentation of a single stimulus with a radial line with a fixed length (four levels: 0.904, 1.016, 1.128, 1.24 cm) and angle (four levels: 46, 54, 62, 70°), the Cartesian product of these levels results in 16 different stimuli. In each of the two conditions, participants were presented with 40 training blocks in which each of the colored stimuli in Figure 3.7 was presented once. After the presentation of each stimulus, participants assigned it to a category using one of two response buttons. Once their response was recorded they received corrective feedback alongside their total proportion of correct responses. Seven transfer blocks were introduced throughout the task, with each

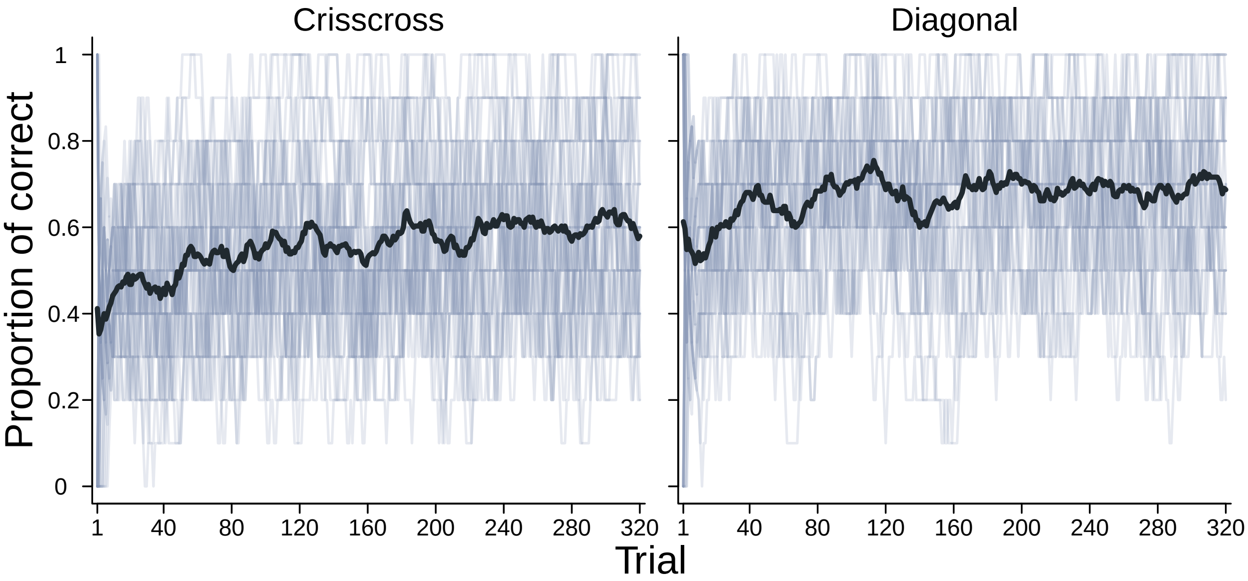

Figure 3.8: Learning performance of the 34 and 32 participants from Bartlema et al. (2014) crisscross and diagonal conditions. Gray lines show the running average of the proportion of correct responses as a function of trials, smoothed by a window 10 trials wide. The black line shows the average across participants.

consisting of a single presentation of all 16 stimuli without feedback. Transfer blocks were presented after 4, 8, 12, 16, 24, 32, and 40 learning blocks.

3.4.1 Results

Figure 3.8 shows the running average of the proportion of correct responses, smoothed with a window of size 10 trials wide, for participants in the criss-cross and diagonal conditions. The gray lines show individual participants, the black lines show the average of the smoothed proportions. We consider only responses to stimuli in the learning blocks. This is because half of the stimuli in the transfer blocks do not have a true category assignment and therefore we can't define a “correct” response. As can be observed in the figure, there is a lot of variability across participants in both conditions. On average, however, participants in the diagonal condition have a higher accuracy throughout the task. However, the group average of the smoothed proportion of correct responses in the criss-cross condition seems to increase from

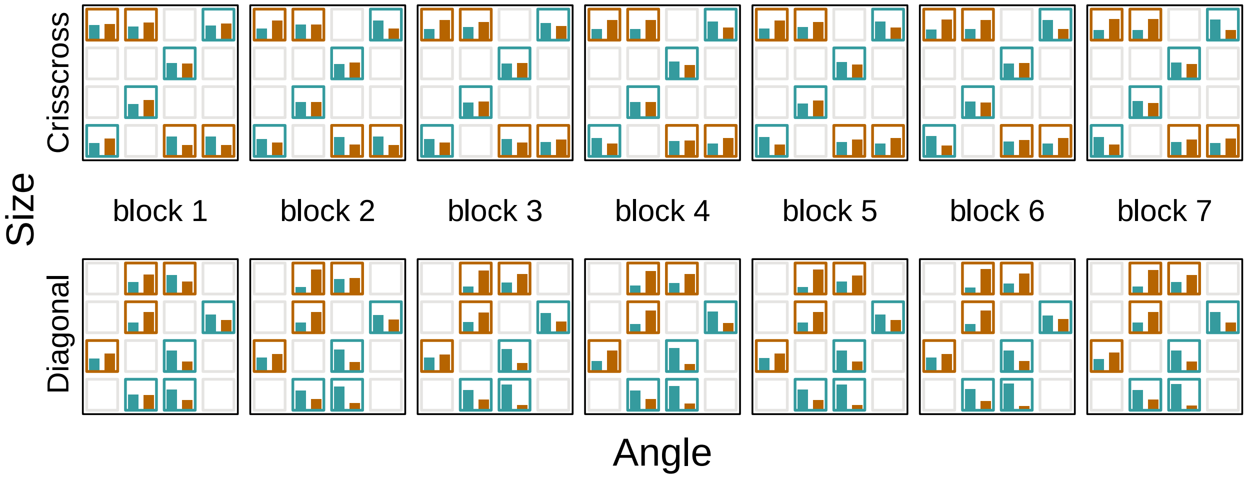

Figure 3.9: Proportion of category "A" (green) and "B" (orange) responses across participants in the Crisscross and Diagonal conditions during training blocks. The colored outlines represent the true category assignment of each stimuli.

around 0.4 to 0.6 suggesting learning, at least for a subset of participants. In comparison, the group average in the diagonal condition was more stable across trials.

Figure 3.9 presents the proportion of "A" (green) and "B" (orange) category label responses averaged across learning blocks. Conditions are organized by row and each panel is organized relative to the angle and size of each stimulus. Colored outlines represent the true category structure of each condition. In the criss-cross condition the correct category is learned for only a subset of stimuli. For example, stimuli in the top left corner, belonging to the "orange" category are correctly classified by the majority of participants in the last block. However, this proportion is lower for stimuli in the lower right corner, which belong to the same category. In contrast, in the diagonal conditions it is clear that after three learning blocks the majority of participants are correctly classifying all eight stimuli. This result contrasts with the amount of learning indicated by the curves in Figure 3.8.

Figure 3.10 shows the average proportion of label responses to stimuli that are only presented during the transfer blocks, once again averaged across participants. Similar to the learning blocks there are no clear trends in the average proportions for the criss-cross condition. This

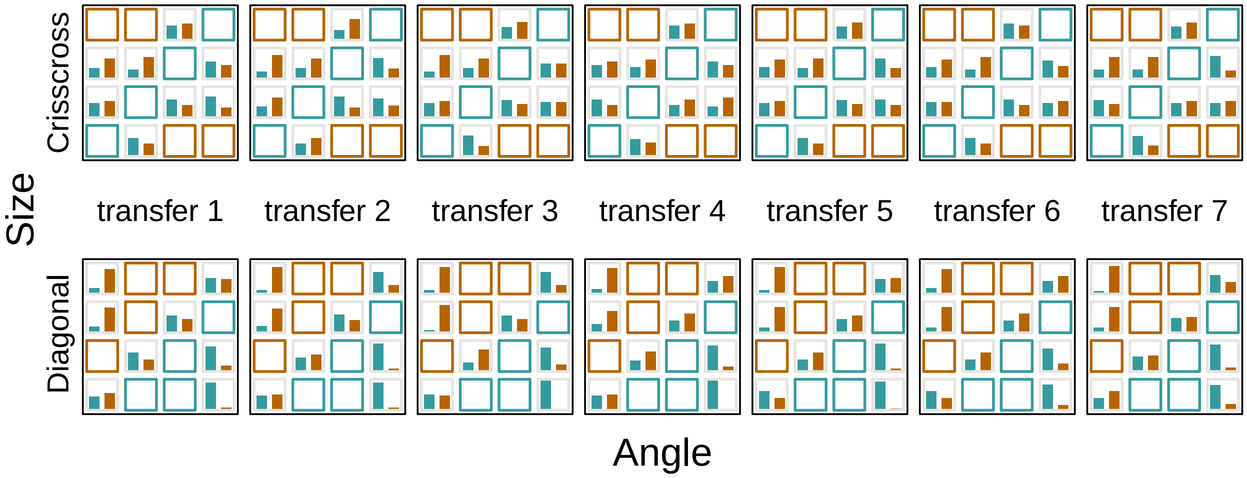

Figure 3.10: Proportion of category "A" (green) and "B" (orange) responses averaged across participants in the crisscross and diagonal conditions during transfer blocks. The colored outlines represent the true category assignment of each stimuli.

suggests that participants might be using different labeling strategies in this condition. In contrast, there is a clear pattern of labeling responses in the diagonal condition. Stimuli in the top right and lower left corner are more often assigned to the category of their nearest stimulus. Stimuli in the diagonal of the representation are split almost half and half at the end of the transfer trials.

Categorical Representation: Transfer trials

Given the difficulty of these category learning tasks, we focus subsequent analysis on a subset of participants for which a Bayes Factor showed strong evidence (Kass & Raftery, 1995) in favor of the hypothesis of a probability of a correct response at an above-chance rate. From all of the participants in the experiment only 10 and 20 met this requirement from the criss-cross and diagonal conditions respectively.

Figure 3.11 shows the label responses of these 10 participants in the criss-cross condition. Transfer blocks are organized in rows. Participants are organized in columns as a function

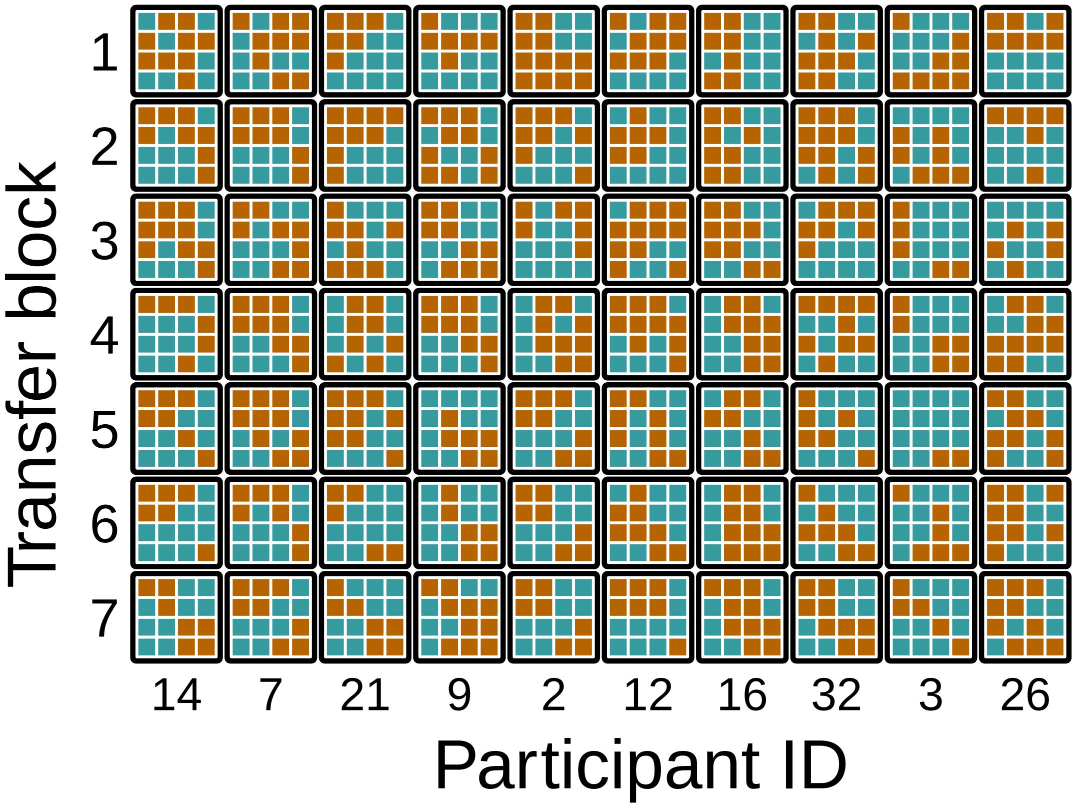

Figure 3.11: Category assignments of 10 participants in the criss-cross condition across transfer blocks. Colored squares show the category assignment of each stimulus to categories "A" (green) and "B" (orange). Transfer blocks are organized as rows while individual participants are in columns.

of their proportion of correct responses on the learning stimuli. Colored squares represent the label choices of each participant. It is clear that there is significant individual variability in classification decisions across a single transfer trial. However, it is interesting to note that by the last transfer, the majority of participants seem to be converging towards a similar categorical representation of the stimulus space. For example, participants seem consistent to classify stimuli in the upper right quadrant of the space to the green category, but they classify stimuli in the top left quadrant to the orange category.

Figure 3.12 shows the classification responses of the 20 participants in the diagonal condition. Transfer blocks are organized in rows, and participants are organized in columns. Colored squares represent classification decisions in each transfer block. Similar to the criss-cross

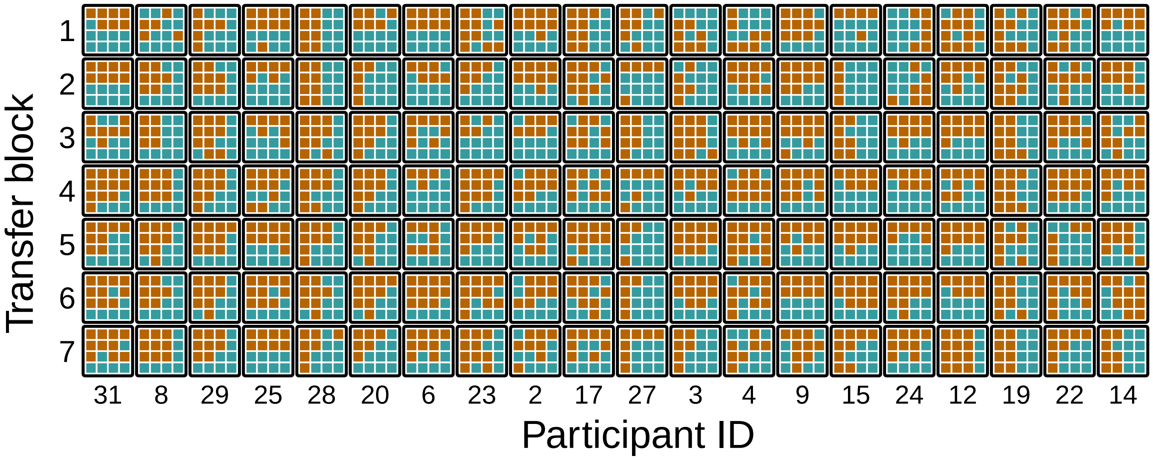

Figure 3.12: Category assignments of 20 participants in the diagonal condition across transfer blocks. Colored squares show the category assignment of each stimulus to categories “A” (green) and “B” (orange). Transfer blocks are organized as rows while individual participants are in columns.

condition, there is significant individual variability. In this case, however, participants do not seem to converge to a similar categorical representation. This shows a clear dissociation between the accuracy measures and the true categorical structure that participants build during learning. The average consistency of participants labeling choices from transfer 6 to 7 in the criss-cross condition is 73%, ranging from 44% to 100% across participants. In the diagonal condition the average is 69%, ranging from 25% to 94% across participants.

3.4.2 Comparison with the Generalized Context Model

Figure 3.13 shows the mean of the predictive distributions of the CHMM and GCM for each participant’s response and stimuli in the transfer blocks for the criss-cross and diagonal conditions. Transfer blocks are organized in columns, and conditions are represented in rows. Each point in the graph represents the mean prediction of each model for a participant’s classification behavior for a single stimulus.

The first row of panels show the predictive accuracy in the criss-cross condition. The GCM

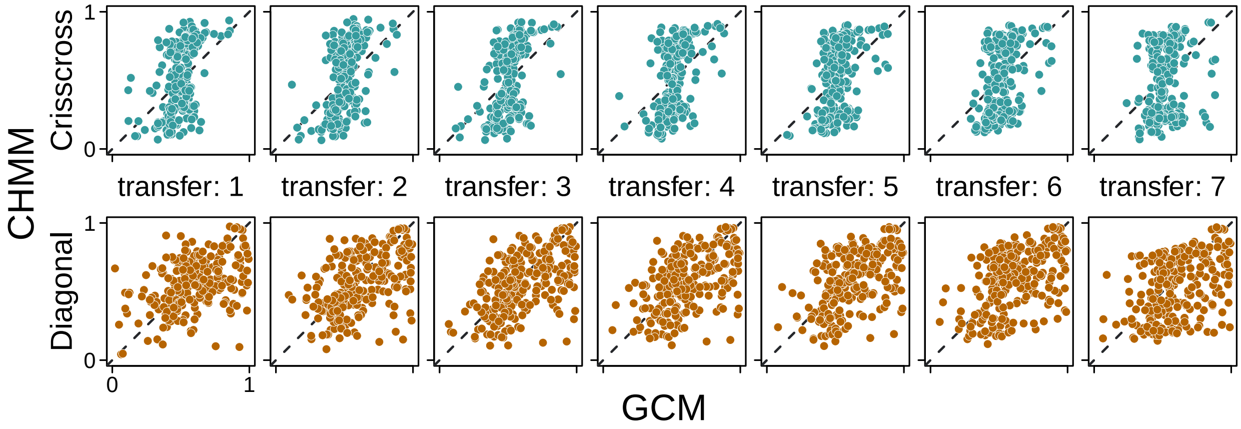

Figure 3.13: Predictive accuracy of the CHMM and GCM model for stimuli in the transfer blocks in the two conditions of Bartlema et al. (2014) experiment. Columns indicate the transfer block, with conditions organized by rows. Circles represent the mean predictive probability assigned to the participants' responses by each model.

has a predictive accuracy that is near 0.5 for almost all stimuli, participants, and transfer blocks. This indicates that the model has a significant level of uncertainty regarding which category a participant will choose. In contrast, the predictions of the CHMM approach are much more spread out, indicating that the model has more "confidence" regarding participant behavior.

The second row of Figure 3.13 shows the predictive accuracy in the diagonal condition. In this case the majority of predictions fall along the identity line, this suggests that both models are equally good at predicting the responses of participants during the transfer blocks. Later transfer blocks, however, seem to favor the GCM's predictions. In both conditions, the proportion of participants for which a model has a higher predictive accuracy is close to 50%, with proportions favoring the CHMM in the criss-cross condition and the GCM in the diagonal condition. However, when we compare the predictive performance of the models on participant's responses for the learning stimuli during transfer blocks, the CHMM outperforms the GCM in both conditions.

3.5 Discussion

In this chapter, we introduced a new version of the CHMM framework to measure the dynamics of categorization. Our main objective was to be able to integrate two parameters to the model that would allow us to measure the weight of similarity and the impact of generalization in the dynamics of the categorical representation of stimuli in a task. We did this by introducing two parameters into the transition probability function of the CHMM approach presented in Chapter 2. In order to test the predictions of the model we applied it to two experiments in the literature. The first one was a category learning experiment reported by Navarro et al. (2005), which presented participants with face stimuli. The second, a learning-transfer designed reported by Bartlema et al. (2014) who used simple visual stimuli. We compared the predictions of the CHMM approach to the predictions of a the well-established GCM (Nosofsky, 1988).

The behavioral data from Navarro et al. (2005) experiment showed that participants had more difficulty learning more abstract category structures. We applied the CHMM framework to the simplest condition in the experiment and showed that the model was able to adequately describe the data used to make the inferences, with an accuracy of approximately 95%. Even though the use of descriptive adequacy as a measure of a models success is widely relied upon, the flexibility of the CHMM makes it difficult to judge the model using only this method. Therefore, we decided to remove participant's responses for four stimuli in the task. Removing those stimuli allowed us to show how similarity drives the predictions of the model. This is shown in the predictive accuracy of the CHMM for stimuli that are assigned to the category of their closest neighbors.

We used the four withheld stimuli to compare the out-of-sample predictions of the CHMM and GCM models of categorization. This comparison showed that assuming that participants know the true underlying category structure can give the GCM model a predictive advantage in some conditions. For example, it allowed the GCM to outperform the predictions of the CHMM in the gender condition of the experiment. However, this advantage vanishes as categories become harder to learn, and people form latent representations of the category structure that do not match the ground truth. Concretely, we showed that the predictive performance of the CHMM in the trustworthiness condition was higher for the majority of participants.

An important result from the analysis of the model's predictions at the stimulus level is that both models have difficulty predicting people' behavior for a borderline stimuli. For example, both failed to predict people's classification behavior for face L in the gender condition. This particular stimulus is closer in psychological space to items in the opposite category, and thus both models make an error in their prediction. This result highlights a problem with similarity-based models of categorization that do not take the effects of feedback into account. All of the participants seemed to learn the correct category assignment of this borderline stimulus after a single instance of corrective feedback.

The behavioral results from the experiment of Bartlema et al. (2014) showed that participants in the task had difficulty learning the underlying category structure for the eight stimuli presented during the learning blocks. Of the two, the criss-cross condition seemed to be the more difficult to learn, as participants smoothed accuracy showed a large degree of variability and a slow increase. Despite this difference in accuracy, looking at the individual level label responses during the transfer trials showed that a subset of participants in the criss-cross condition converged to a single representation of the stimulus space, in which stimuli in a given quadrant are assigned to the same category. This result suggests that some participants might have used a combination of simple rules when making their decisions. In the diagonal condition we observed much more individual variability in the categorical representation of the stimuli. This suggest that, in this case, a single categorical representation might be harder to abstract from the training blocks of the experiment.

The training-transfer design implemented in Bartlema et al. (2014) experiment allowed us to test the predictions of the CHMM and GCM models of categorization without the need to remove stimuli artificially from the analysis. In this case, the inferences of each model are based on observed behavior from the learning blocks, which can then be used to generate predictions about participants' responses during transfer blocks. In the criss-cross condition, the predictive accuracy of the CHMM showed more variability than the GCM. In the diagonal condition, both models had similar predictive accuracy. However, the predictive accuracy of the CHMM approach for learned stimuli in the transfer trials outperformed that of the GCM in both conditions.

These results highlight some limitations of the current implementation of the CHMM approach. The first is that not being able to account for the effects of feedback hinders the ability to make out-of-sample predictions, at least in situations where the underlying category structure is easy to learn. A second limitation is the lack of constraints in the model. This is shown by the variability in the model's predictions for some of the experimental conditions. While the GCM constrains its predictions by assuming that participants know the underlying category structure, this is not the case for the CHMM. One of the properties of the CHMM, which may be useful under other circumstances, is that it allows the hidden process to move to its stationary distribution over time. This hinders its ability to predict responses to stimuli whose classification has been learned perfectly by the end of the experiment, as was observed in the gender and hair conditions of the Navarro et al. (2005) experiment. Despite these limitations, the CHMM approach showed a predictive accuracy that was close and sometimes higher, for specific stimuli and specific participant, than the GCM.

CONCLUSION

The main objective of this work has been to study the dynamics of people's categorization decisions, especially in situations in which the dynamics of the environment are coupled with people's behavior. To this end, we designed two tasks in which participants had to categorize real-world stimuli according to an underlying category structure that changed as participants became more accurate. Results from these experiment showed that people are able to adapt their categorization choices after a change in the category structure. However, accounting for this adaptation process has proved to be challenging.

In order to be able to measure the dynamics of categorization, we developed a new framework based on Coupled Hidden Markov Models (CHMMs). Using data from previously published studies we showed that this new framework was able to describe adequately the behavior of participants in tasks that present both real-world and artificial stimuli. Furthermore, we showed that the CHMM's accuracy for out-of-sample data increased with the difficulty of learning category structures, and that in many cases it matches or outperforms the predictive accuracy of the Generalized Context Model of categorization.

A clear avenue for future work is to apply the CHMM approach to participant's behavior in the two experiments introduced in Chapter 1. This would allow us to measure how the categorical representation of stimuli in the task changes as participant's interact with the dynamic environment that is sensitive to their behavior. Inferences drawn from this framework would allow us to test the effects of a change in the underlying category structure on how stimuli are categorized. It would also allow us to see if the effect of the change in the assignment of one stimulus ripples to affect the classification of similar stimuli.

A key aspect of the applications of the CHMM framework that we need to consider is that the experimental designs we have used to test the model, represent the worst possible scenario for it. As we mention in Chapter 2, when participants label a single stimulus at a time, the model interprets the responses to all other stimuli in the task as missing data. Therefore, the inferences made by the model have to take into account that there is missing information between two consecutive responses to the same stimulus. An experimental design that would prove more informative from the CHMM perspective would be a category learning task in which participants have to label multiple stimuli in a single trial. This type of experiment would give more information for the model to make inferences about how stimuli are categorized on any given trial. We think it is also likely to show an advantage of the CHMM framework in comparison to other categorization models, since making inferences about the category assignments of multiple stimuli in the same trial is a natural extension of the CHMM.

A final important line for future work is to improve models using the CHMM approach to categorization. There are two clear possible directions for extensions. One is to add a method that allows the CHMM to assign a different weight to dimensions of a psychological space. This is known in the literature as selective attention and established models have successfully incorporated attention mechanisms to help understand category learning. The second, and perhaps more challenging extension, is to incorporate the effects of feedback into the CHMM approach. The main reason this presents a problem is that the model requires the assumption of independence between the next categorical representation of an item and the previous response. It also works counter to the main inspiration of the CHMM as a measurement model. But, a complete account of how people learn and adapt categories over time needs to account for the entire task environment, and feedback is an important observable part of many category learning situations.

Bibliography

Ashby, F. G. & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178, https://doi.org/10.1146/annurev.psych.56.091103.070217.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150, https://doi.org/10.1016/j.jmp.2013.12.002.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In *Proceedings; Inequalities*, volume 3 (pp. 1–8). Retrived from: https://files.library.northwestern.edu/public/Files/Baum.pdf.

Brand, M. (1997). *Coupled hidden Markov models for modeling interacting processes.* Technical report, The Media Lab, Massachusetts Institute of Technology, 20 Ames Street Cambridge, MA 02139 USA. Retrived from: https://api.semanticscholar.org/CorpusID:14310282.

Brand, M., Oliver, N., & Pentland, A. (1997). Coupled hidden Markov models for complex action recognition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 994–999). https://doi.org/10.1109/CVPR.1997.609450.

Carter, C. K. & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, *81*(3), 541–553, https://doi.org/10.2307/2337125.

Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. (2000). Markov chain Monte Carlo sampling. In *Monte Carlo methods in Bayesian computation* (pp. 19–66). Springer Series in Statistics, 1st edition. https://doi.org/10.1007/978-1-4612-1276-8_2.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, *75*(1), 79–97, https://doi.org/10.1016/0304-4076(95)01770-4.

D'Alessandro, M., Radev, S. T., Voss, A., & Lombardi, L. (2020). A Bayesian brain model of adaptive behavior: An application to the Wisconsin Card Sorting task. *PeerJ*, *8*, 1–32, https://doi.org/10.7717/peerj.10316.

Dehaene, S. & Changeux, J.-P. (1991). The Wisconsin Card Sorting Test: Theoretical analysis and modeling in a neuronal network. *Cerebral Cortex*, *1*(1), 62–79, `https://doi.org/10.1093/cercor/1.1.62`.

Devaney, M. & Ram, A. (1996). Dynamically adjusting concepts to accommodate changing contexts. In M. Kubat & G. Widmer (Eds.), *Proceedings of ICML-96 Pre-Conference Workshop on Learning in Context-Sensitive Domains* Bari, Italy.

El Kouari, O., Benaboud, H., & Lazaar, S. (2020). Using machine learning to deal with Phishing and Spam Detection: An overview. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, volume 72 (pp. 1–7). New York, NY, USA: Association for Computing Machinery.

Estes, W. K. (1984). Global and local control of choice behavior by cyclically varying outcome probabilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 258–270, `https://doi.org/10.1037/0278-7393.10.2.258`.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633, `https://doi.org/10.1038/35036586`.

Gallistel, C., Mark, T. A., King, A. P., & Latham, P. (2001). The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *Journal of Experimental Psychology: Animal Behavior Processes*, *27*(4), 354–372, `https://doi.org/10.1037//0097-7403.27.4.354`.

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The Perception of Probability. *Psychological Review*, *121*(1), 96–123, `https://doi.org/10.1037//a0035232`.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. New York, NY: Chapman and Hall/CRC, third edition.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. In *Bayesian Statistics*, volume 4 (pp. 641–649). Clarendon Press. Retrived from: `https://www.jstor.org/stable/2245993`.

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic learning. *Learning and Memory*, *9*(6), 408–418, `https://doi.org/10.1101/lm.45202`.

Goldstone, R. L. & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78, `https://doi.org/doi.org/10.1002/wcs.26`.

Hemmer, P. & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202, `https://doi.org/10.1111/j.1756-8765.2008.01010.x`.

Iwashita, A. S. & Papa, J. P. (2019). An overview on concept drift learning. *IEEE Access*, *7*, 1532–1547, https://doi.org/10.1109/ACCESS.2018.2886026.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 377–395, https://doi.org/10.1080/01621459.1995.10476572.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning and Memory*, *1*, 106–120, https://doi.org/10.1101/lm.1.2.106.

Koychev, I. (2007). Experiments with two approaches for tracking drifting concepts. *Serdica Journal of Computing*, *1*(1), 27–44. Retrived from: http://eudml.org/doc/11410.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44, https://doi.org/10.1037/0033-295X.99.1.22.

Kruschke, J. K. (1993a). Human category learning: Implications for backpropagation models. *Connection Science*, *5*(1), 3–36, https://doi.org/10.1080/09540099308915683.

Kruschke, J. K. (1993b). Human category learning: Implications for backpropagation models. *Connection Science*, *5*(1), 3–36, https://doi.org/10.1080/09540099308915683.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–248, https://doi.org/10.1080/095400996116893.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science*, *12*(5), 171–175, https://doi.org/10.1111/1467-8721.01254.

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 267–301). Cambridge University Press. https://doi.org/10.1017/CBO9780511816772.013.

Kruschke, J. K. & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*(5), 1083–1119, https://doi.org/10.1037/0278-7393.25.5.1083.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1–27, https://doi.org/10.1007/BF02289565.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin and Review*, *14*(4), 560–576, https://doi.org/10.3758/BF03196806.

Lee, M. D. & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, *9*, 43–58, https://doi.org/10.3758/bf03196256.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332, https://doi.org/10.1037/0033-295X.111.2.309.

Maloof, M. A. (2003). Incremental rule learning with partial instance memory for changing concepts. In *Proceedings of the International Joint Conference on Neural Networks*, volume 4 (pp. 2764–2769). https://doi.org/10.1109/IJCNN.2003.1224005.

Navarro, D. J., Lee, M. D., & Nikkerud, H. (2005). Learned Categorical Perception for Natural Faces. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, volume 27 (pp. 1–6). Retrieved from https://escholarship.org/uc/item/4sw079dp.

Navarro, D. J., Perfors, A., & Vong, W. K. (2013). Learning time-varying categories. *Memory & Cognition*, *41*(6), 917–927, https://doi.org/10.3758/s13421-013-0309-6.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In B. Steve, G. Andrew, J. Galin, & M. Xiao-Li (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 113–162). Chapman and Hall/CRC, 1st edition. https://doi.org/10.1201/b10905.

Nosofsky, R. M. (1986). Attention, similarity and the idenitification–categorization relationship. *Journal of Experimental psychology: General*, *115*(1), 39–61, https://doi.org/10.1037//0096-3445.115.1.39.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Leaning, Memory, and Cognition*, *14*(1), 54–65, https://doi.org/10.1037/0278-7393.14.1.54.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perceprion & Psychophisics*, *45*(4), 279–290, https://doi.org/10.3758/BF03204942.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53–79, https://doi.org/10.1037/0033-295x.101.1.53.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). A formal psychological model of classification applied to natural-science category learning. *Current Directions in Psychological Science*, *27*(2), 129–135, https://doi.org/10.1177/0963721417740.

O'Doherty, K. C. & Lee, M. D. (2002). The featural representation of animals based on similarity. *Australian Journal of Psychology*, *54*(1), 60, https://doi.org/10.1080/00049530210001706513.

Patterson, J. D. & Kurtz, K. J. (2020). Comparison-based learning of relational categories (you'll never guess). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(5), 851–871, https://doi.org/10.1037/xlm0000758.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441, https://doi.org/10.1073/pnas.0610341104.

Reheder, B. & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1261–1275, https://doi.org/10.1037/0278-7393.27.5.1261.

Roberts, S. & Pashler, H. (2000). How Persuasive is a Good Fit? A Comment on Theory Testing. *Psychological Review*, *107*(2), 358–367, https://doi.org/10.1037//0033-295x.107.2.358.

Ross, B. H. & Murphy, G. L. (1999). Food for Thought: Cross-Classification and Category Organization in a Complex Real-World Domain. *Cognitive Psychology*, *38*(4), 495–553, https://doi.org/10.1006/cogp.1998.0712.

Schlegelmilch, R., Wills, A. J., & von Helversen, B. (2022). A cognitive category-learning model of rule abstraction, attention learning, and contextual modulation. *Psychological Review*, *129*(6), 1211–1248, https://doi.org/10.1037/rev0000321.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *17*(3), 433–443, https://doi.org/10.1037/0278-7393.17.3.433.

Shepard, R. N. (1980). Multidimensional scaling, tree–fitting, and clustering. *Science*, *214*(4468), 390–398, https://doi.org/10.1126/science.210.4468.390.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323, https://doi.org/10.1126/science.3629243.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42, https://doi.org/10.1037/h0093825.

Sherlock, C., Xifara, T., Telfer, S., & Begon, M. (2013). A coupled hidden Markov model for disease interactions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(4), 609–627, https://doi.org/https://doi.org/10.1111/rssc.12015.

Smith, J. D., Coutinho, M. V. C., & Couchman, J. J. (2011). The learning of exclusive–or categories by monkeys (macaca mulatta) and humans (homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(1), 20–29, https://doi.org/10.1037/a0019497.

Smith, J. D. & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436, https://doi.org/10.1037//0278-7393.24.6.1411.

Smith, J. D. & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3–27, https://doi.org/10.1037//0278-7393.26.1.3.

Speekenbrink, M. & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, *139*(2), 266–298, https://doi.org/10.1037/a0018620.

Sugak, A., Martynyuk, O., & Drozd, O. (2015). Models of the mutation and immunity in test behavioral evolution. In *2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 2 (pp. 790–795). https://doi.org/10.1109/IDAACS.2015.7341411.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning: An introduction*. Cambridge, Ma: The MIT Press, 2nd edition.

Tenenbaum, J. B. & Griffiths, T. L. (2001). Generalization, Similarity, and Bayesian Inference. *Behavioral and Brain Sciences*, *24*(4), 629–640, https://doi.org/10.1017/s0140525x01000061.

Touloupou, P., Finkenstädt, B., & Spencer, S. E. (2020). Scalable Bayesian inference for coupled hidden Markov and semi–Markov models. *Journal of Computational and Graphical Statistics*, *29*(2), 238–249, https://doi.org/10.1080/10618600.2019.1654880.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352, https://doi.org/10.1037/0033-295X.84.4.327.

Vanpaemel, W. (2020). Strong theory testing using the prior predictive and the data prior. *Psychological Review*, *127*(1), 136–145, https://doi.org/10.1037/rev0000167.

Veksler, V., Myers, C., & Gluck, K. (2015). Model flexibility analysis. *Psychological Review*, *122*(4), 755–769, https://doi.org/10.1037/a0039657.

Villarreal, M., Etz, A., & Lee, M. D. (2023). Evaluating the complexity and falsifiability of psychological models. *Psychological Review*, *130*(4), 853–872, https://doi.org/10.1037/rev0000421.

Villarreal, M., Vaday, S., & Lee, M. D. (2022). Categorization in environments that change when people learn. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, volume 44 (pp. 680–686). Retrieved from https://escholarship.org/uc/item/1j33x6qg.

Westfall, H. A. & Lee, M. D. (2021). A model-based analysis of the impairment of semantic memory. *Psychonomic Bulletin & Review*, *28*, 1484–1494, https://doi.org/10.3758/s13423-020-01875-9.

Widmer, G. & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *23*, 69–101, https://doi.org/10.1007/BF00116900.

Wittgenstein, L. (1958). *Philosophical Investigations*. Oxford, U.K: Basil Blackwell, 3rd edition. Translated by G. E. M. Anscombe. Retrived from https://books.google.com/books?id=U-sb0AEACAAJ.

Zhong, S. & Ghosh, J. (2002). HMMs and coupled HMMs for multi–channel EEG classification. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, volume 2 (pp. 1154–1159). Honolulu, HI, USA. `https://doi.org/10.1109/IJCNN.2002.1007657`.

Zilker, V. (2022). Choice rules can affect the informativeness of model comparisons. *Computational Brain & Behavior*, *5*(3), 397–421, `https://doi.org/10.1007/s42113-022-00142-5`.