Constraints of Selective Attention during Category Learning in High-Dimensional Information Space

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Yiming Wang

Graduate Program in Psychology

The Ohio State University

2025

Thesis Committee

Brandon Turner, Advisor

Peter Kvam

Vladimir Sloutsky

Abstract

In this information-dense era, humans adapt to selectively utilize a subset of information that can realize their goals while preserving cognitive resources. Understanding how people narrow down available information and seek economical solutions to optimize performance and reduce cognitive efforts is thus pivotal to understanding how humans navigate complex environments. This thesis focuses on two characteristics of selective attention during category learning – accuracy and simplicity in representation. To investigate the boundaries of these two processes, we design an experiment that varies along the environmental complexity by manipulating the number of dimensions in the learning set. The behavioral data provide strong evidence that higher dimensionality induces sparser attention, therefore increasing the risk of learners being trapped by suboptimal information and sacrificing accuracy. We present a computational framework that extends the Adaptive Attention Representation Model (AARM) to more explicitly account for the interplay of the two aforementioned goals of human learning. Our account enforces constraints on attention into the optimization process, offering a simple yet robust computational mechanism to capture the nature of selective attention during category learning in different environments.

Acknowledgments

Two days before Thanksgiving, as I work on this most personal part of my master's thesis, it feels like exactly the right moment to express my gratitude and love to the people around me.

First and foremost, I would like to offer my sincere thanks to my advisors. Dr. Brandon Turner, thank you for recognizing my potential when I felt uncertain about my direction, for always encouraging every small step of progress I made, and for stimulating me with your insights. Dr. Peter Kvam, thank you for always being generous and supportive as I explored the field, for sharing a lot of resources to broaden my horizons, and for patiently working with me hands-on. I am profoundly grateful to have you as the coach of my research. You have helped me grow beyond what I thought I could be.

I would also like to thank Dr. Vladimir Sloutsky as a member of my committee. Thank you for your valuable feedback from unique perspectives and conversations that helped me reflect more deeply on how I can contribute to this field.

I would like to thank my lab mates in the Model-based Cognitive Neuroscience Lab and Cognition and Decision Modeling Lab for making me always feel a sense of belonging. Nicole, thank you for being so responsive and responsible in taking care of me when I was at a loss, for being so warm and loving when I felt frustrated or sad, and for your positive energy. Benson and Progga, thank you for being such great friends and cohort-

mates, and for always patiently helping me with my confusions and puzzles. Konstantina, Andy, Robby, Eeshan, Hanrui, Abhay, and Murray, thank you all for always taking my immature thoughts seriously and sharing your experienced perspectives with me. Each of you has been a strong source of support for me here.

I would like to express my heartfelt thanks to my family. Mom and Dad, thank you for loving me so generously and selflessly, for respecting and supporting whatever I want to pursue, and for always being there for me and being proud of whatever I achieve. I could not be the strong, independent, and brave girl I am now without you, my beloved family.

I would also like to especially thank my friends. Kexin, Weiyi, Tengfei, and Tianzhen, thank you for taking such good care of me since I was a clueless teenager, for being my best cheerleaders, and for accompanying me through life wherever we are. Mia, Ari, Amber, and Jiaqi, thank you for being my friends in Columbus, for helping me fit in and making me feel at home. All of my friends, thank you for making me happy and helping me stay emotionally resilient when facing difficulties.

Life is all about who we share it with, and all of you have shaped who I am today. I again express my deepest gratitude to you for supporting me throughout this journey—both in completing this thesis and in my life more broadly.

Vita

August 2025 – Current	Graduate Research Assistant ~ PI: Brandon Turner
	The Ohio State University (Ohio, United States)
August 2023 – January 2024	Research Assistant ~ PI: Henrik Singmann
	University of Warwick (Remote)
September 2022 – September 2023	Master of Research in Cognitive Neuroscience
	University College London (London, United Kingdom)
August 2021 – October 2021	Consultancy Intern ~ Performance, Reward, and Talent
	Aon Hewitt (Beijing, China)
June 2021 – August 2021	Data Analyst Intern ~ Human Resource
	TikTok China (Beijing, China)
September 2020 – March 2021	Undergraduate Research Assistant
	Beijing Normal University (Beijing, China)
September 2018 – June 2022	Bachelor of Science in Psychology
	Beijing Normal University (Beijing, China)

Fields of Study

Major Field: Psychology

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Humans are confronted with an overwhelming amount of information that far exceeds the capacity of processing. Unlike supercomputers, we are granted a low storage workspace that forces us to concentrate on a small subset. Attention, as the gatekeeper of most conscious cognitive processes, governs this early selection. Humans selectively attend to information that is relevant to their goals. Selective attention sits in close dialogue with the explore-exploit argument: there is always a question of when to keep exploring the information board for more knowledge versus when to stop and shift gear to exploiting already encoded information (Navarro et al., 2016). An inherent tendency to reduce cognitive effort and seek for simplicity is crucial to this dilemma, which may trigger an early pause on exploration even when the optimal solution has yet to be achieved. The need to simplify the environment should become more evident as the complexity of the environment increases, e.g., with either increasing amounts of information or increasing complexity in how dimensions are combined to create rules. For example, as the number of dimensions of information increases, continual exploration can create enormous search costs, making exploitation a reasonable and economic strategy for performing a task. Our study aims to understand the explore-exploit trade-off induced by selective attention during category learning in complex environments. We operationalize the complexity of an environment as the number of dimensions of the stimuli as a way to raise the demands

on information encoding and storage during category learning. Computationally, we implement the explore-exploit trade-off under the framework of the adaptive attention representation model (AARM). However, we explore a more flexible specification of the cost function by introducing an ℓ_p-norm regularization term. This cost function is more general than what was previously specified by Galdo et al. (2022), and in particular, it subsumes ridge and LASSO regularization mechanisms as we will discuss in Chapter 3.

1.1 Costs of Attention

When exposed to many dimensions of information to learn the categorization rule, humans encode the information and further process it in working memory. The huge costs of neural activities limit the amount of information that can be processed, which is referred to as the capacity of information processing (Lennie, 2003). Each spike of a single neuron carries with it a substantial cost, and as a result, there forms a metabolic limit on the number and extent of concurrently activated neurons in our cortical system. Due to these metabolic constraints, it becomes a priority of the learning system to seek out economical representations by identifying, attending to, and processing only the most relevant dimensions of information so that a minimum number of neurons will be used (Barlow, 1972; Dukas 2004; Pashler et al., 2001). Hence, the need to reduce computational costs turns into a goal of the learner.

A commonly accepted dichotomy for understanding attention is to divide it into two types of capacity-limited attention – external and internal attention (Chun et al., 2011). External attention orients the learner to the perceptual world and endows them with

limited spaces when splitting across multiple sensory loci (Jolicœur & Dell'Acqua, 1999). A typical example is the visuospatial attention – people only allocate attention to a finite number of locations and guide eye movements to these locations (Palmer, 1990; Schall & Thompson, 1999; West, 2010). This supports the capacity-limited property of external attention and also justifies gaze as a proper reflection of attention.

By contrast, internal attention refers to internal cognitive representations, such as those stored in working memory or memory in general. Decades of research on working memory suggests that internal attention has a clear capacity, such that the amount of information that can be actively maintained is 7 ± 2 chunks (Baddeley, 2003; Brady et al., 2011; Miller, 1956).

Returning to the context of category learning, when stimuli are constructed with a number of stimulus dimensions, as the number of dimensions grows, we near the capacity of internal attention. As such, the capacity of the learner becomes the driving force behind the explore-exploit tradeoff. With multiple dimensions to be considered, a rational cognitive entity needs to allocate their cognitive resources in an economical manner such that they focus on relevant dimensions of information rather than wasting resources on less relevant dimensions of information for their task.

1.2 Selective Attention

The evolutionary sparsity of neural activations and thereby the capacity-limited attention promotes an outstanding pattern within attention – selectiveness. Humans are extremely prudent when exploring the information space, economizing the amount of attention by

focusing on the most relevant pieces to jointly satisfy the goal of simplicity and the goal of accuracy (Galdo et al., 2022; Turner & Sloutsky, 2024; Weichart et al., 2024).

An aspect of selective attention is its context-specific property: humans, especially adults, can efficiently allocate attention resources to useful dimensions (Shepard et al., 1961). On one hand, people distribute their attention in accordance with the categorization rule. This task-responsive attention has been well supported by a great body of experiments using eye-tracking where people learn to fixate to a subset of diagnostic dimensions that relate to categorization (Blair et al., 2009a; Castro et al., 2020; Chen et al., 2013; Dolguikh et al., 2021; Poletti et al., 2017; Rehder & Hoffman, 2005). In addition to visual tasks, evidence of auditory attention shows that people attend to diagnostic acoustic dimensions and further enhance the cortical tracking of attended dimensions (Symons et al., 2021). Moreover, people apply distinct attention patterns based on different stimuli. According to Nosofsky and Hu's (2023) experiment, participants learned to selectively attend to specific dimensions for different subsets of stimuli. This stimulus-responsive attention allows high flexibility in attention even at the within-trial level (Blair et al., 2009b). These sophisticated mechanisms of attention is a hallmark of our fine-tuned adaptive attention system (Weichart et al., 2022).

Although selective attention promotes an efficient, economical use of biological resources on task-relevant information, simplifying the representation of the task can produce other costs. When one's attention focus is too narrow, they may form inaccurate or incomplete representation of the learning set, which is identified as the "learning trap" (Rich & Gureckis, 2018). On one hand, one can fail to identify the most diagnostic

dimension if the avoidance of attentional efforts overweighs the pursuit of accuracy. In this case, the performance detriment is just a consequence of the over-simplified representation of the environment. On the other hand, selective attention may blind people in dynamic environments such that they fail to adapt to the new characteristics of the environment if they change (Lee et al., 2024; Plebanek & Sloutsky, 2017). For example, people learn to shift attention away from the irrelevant information yet fail to draw their attention back even it becomes relevant at a later point (learned inattention, Blanco & Sloutsky, 2019; Kruschke & Blair, 2000; Hoffman & Rehder, 2010).

We believe that human's need for simplicity in selective attention during category learning especially stands out in complex environments requiring high perceptual loads (Lavie et al., 2014). According to Lavie's Load Theory of Attention (1995), in cognitive tasks involving a low amount of information, distractors will be automatically processed with spare capacity, thereby interfering with targets. The availability of perceptual processing in a lightly demanding context may attenuate the selectivity of attention. However, selective attention significantly functions provided heavy perceptual loads, with no spare attention spillover to irrelevant information and instead top-down attention prioritized for the task-relevant subset (Lavie et al., 2014). Therefore, manipulating the dimensionality of information is effective to inducing selective attention. We expect that with higher-dimensional stimulus, people will focus on smaller subsets even though these subsets might not be optimal for task performance.

In this thesis, we investigate how the goal of accuracy interacts with the goal of simplicity by systematically increasing the number of dimensions. Our hypothesis is that

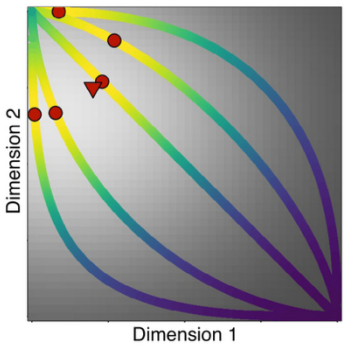
as the number of dimensions grows, the need for simplicity will become more evident in terms of the number of sampled dimensions. As a result, the probability of identifying the most diagnostic dimensions, reflected in the accuracy, should decrease as the number of information dimensions are increased.

1.3 Attention Constraints in Categorization Models

As indicated by the explore-exploit dilemma in selective attention, people balance their cognitive cost and performance. One might rather pay less attention to obtain a fairly good accuracy than exhaust all attention resources to achieve ceiling accuracy. This quest for simplicity is implemented in categorization models in different ways (Galdo et al., 2022; Kruschke, 2001; Nosofsky, 1986; Paskewitz & Jones, 2020). In this thesis, we argue that previous solutions may either over-constrain attention or over-parameterize attention constraints by allowing several independent mechanisms. We then introduce our approach of modeling attention constraints through the ℓ_p-norm. For consistency, we will denote attention vector as $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_D]$, with attention to dimension j as α_j.

A prevailing approach in prior categorization models is to constrain attention with the sum-to-constant or norm-to-constant assumption. At very first, the Generalized Context Model (GCM) assumes $\sum_j \alpha_j = 1$ that total attention sums up to one (Nosofsky, 1986). Similarly, some models assume the sum of certain transformed α_j to be constant. For example, ALCOVE (Kruschke, 1992) assumes $\left(\sum_j \alpha_j^p\right)^{\frac{1}{p}} = 1$ while EXIT (Kruschke, 2001; Paskewitz & Jones, 2020) assumes $\left(\sum_j \alpha_j^p\right)^{\frac{1}{p}} = k$ (constant). A nice illustration

from Galdo et al. (2022) visualizes how these sum-to-constant assumptions impose a structural constraint on the attention space. Figure 1 shows that, in a two-dimensional attention space of $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]$, attention to Dimension 1 is fully determined through $\left(\sum_{j=1}^{2} \alpha_j^p\right)^{\frac{1}{p}} = 1$ when attention to Dimension 2 is specified. This figure suggests that these sum-to-constant assumptions constrain one's attention state to only move along one of the colored traces specified by p and k. For example, the diagonal can be defined with $p = 1$, i.e., $\sum_{j=1}^{2} \alpha_j = 1$, which is the same as the assumption of GCM. The convex traces in the upper triangle are defined with $p > 1$ while the concave traces in the lower triangle are defined with $p < 1$, as assumed by ALCOVE and EXIT.



This figure shows how attention is constrained with sum-to-constant assumptions. In this figure, x and y axes represent attention to Dimension 1 (α_1) and Dimension 2 (α_2). Each trace composes a series of $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]$ that satisfy $\left(\sum_{j=1}^{2} \alpha_j^p\right)^{\frac{1}{p}} = 1$ with different p. The color on the trace from yellow to purple reflects the loss (i.e., cross entropy) related to $\boldsymbol{\alpha}$ at that point, with yellow indicating $\boldsymbol{\alpha}$ related to lower loss. The red triangle represents the global maximum of $\boldsymbol{\alpha}$ over the full attention space, while red dots represent locally optimal $\boldsymbol{\alpha}$ on each trace with certain constraint.

Figure 1 Structural attention constraints from Galdo et al. (2022)

The solution from Galdo et al. (2022) argues that attention allocation should not be a zero-sum game. In addition to competition, they identify two computational efficiency goals and systematically determine which one has more contribution by implementing them through independent regularization terms. First, humans limit the total amount of attention in cognitive tasks, and this is realized through the Ridge regularization. What's more, people limit the number of dimensions they attend to, which is referred to as the dimension-reduction tendency or quest for rule simplicity (Blair et al., 2009b; Matsuka & Corter, 2008; McColeman et al., 2014; Meier & Blair, 2013). They apply the Lasso regularization to constrain the number of attended dimensions. To introduce a separate mechanism that dimensions compete for attention resources, they use the competition parameter in the gradient inhibition apart from the regularization terms (Turner, 2019). However, both solutions can be improved in some way. The sum-to-constant assumptions place excessive structural restrictions over the attention space. Referring to Figure 1, each model specification only allows the attention state to vary on a discrete slice out of the whole 2D space, sometimes resulting in a huge discrepancy between the local maximum constrained by the model and the global maximum. Also, the zero-sum policy naturally presents competition together with dimension reduction and implicitly assumes that the full capacity of attention is exhausted on every trial. In contrast, Galdo et al. (2022) tune separate parameters for different efficiency goals and competition mechanisms, while the regularization and gradient inhibition may not be necessarily required at the same time. In this thesis, we maintain that there is some common thread among different constraints of attention and there could be an intermediate solution that allows appropriate flexibility.

We add the ℓ_p-norm regularization term on attention with $p < 1$ into the loss function, thus the optimization process as well. Our approach realizes sparse attention in a dynamic way that preserves the interplay among various constraints through one parameter, p, in the ℓ_p regularization, while avoiding over-simplification of attention constraints. We will specify and validate the modeling approach in detail in Chapter 3.

To summarize, this thesis investigates attention dynamics in the category learning environment that vary in terms of complexity. In Chapter 2, we first examine the response and sampling data from our experiment. In Chapter 3, we introduce our model based on AARM that optimizes attention on the loss function that accommodate error minimization and attention sparsity. In Chapter 4, we show model fits to the data from our experiment. We close with Chapter 5, a discussion of the implications of this work.

Chapter 2. Behavioral Experiment

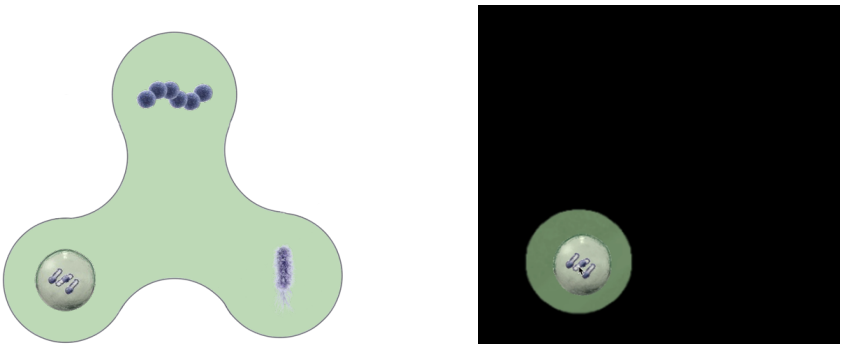
To examine how selective attention manifests during category learning, we designed a task that systematically increases the number of dimensions in the learning set. With more dimensions, it becomes more necessary to selectively attend to subsets of the dimensions, which accentuates the tradeoffs between the learning goal of accuracy and the cost of information sampling.

2.1 Methods

2.1.1 Materials

We designed a category learning task to address the theoretical questions of interest. The task is built on amoeba stimuli (Blair et al., 2009b; McColeman et al., 2014). In biological contexts, the amoeba is the unicellular organism with arm-like structures representing dimensions of info. Each arm contains one organelle, referred to as a feature (Figure 2a). Each feature has two levels, which results in some unique combinations of amoebas divided into two categories.

(a) (b)



(a) An exemplar amoeba with three arms, each containing one feature; (b) A timepoint during a trial of the mouse-contingent paradigm, when the cursor reveals an area around it containing one feature.

Figure 2 Exemplar stimulus and trial

The amoeba stimuli, with its evenly distributed arms, arrange features that are spatially separable, which enables a close link between what is visually sampled from and what is attended to. The cursor-contingent design, combined with a black screen cover, further enables a clear interpretation of what features can possibly be attended at every moment in time (Figure 2b). In particular, when people explore with the cursor, only a small circular area around the cursor is revealed synchronously and serves as the window for people to see the amoeba. The size of the visible field is tuned according to the size of arms so that at most one feature can be revealed at a time. Theoretically, this constrains visual attention to the only visible feature. This cursor-contingent paradigm alleviates some of the limitations of the standard eye-tracking experiments in two aspects: first, it enables online data collection by recording cursor traces in replace of in-person eye-

tracking; second, it produces cleaner data with precise locations of the visible field, while the eye-tracking experiments can have calibration issues.

Features differ in their diagnosticity for the category. One feature is deterministic of the category labels – one level always indicates the category A while the other indicates the category B. Another feature probably indicates the correct category – one level is related to the category A (correct category) in 75% cases while the category B in the rest 25% cases. The last feature is irrelevant to the category labels such that each level appears in both categories for 50% of the time. For reference, we call the first type of completely diagnostic features the deterministic or D feature, the second that is mostly diagnostic the probabilistic or P feature, and the last one the irrelevant or I feature. The discrepancy in feature diagnosticity can successfully drive selective attention to features that produce a higher accuracy, such as D and P features (Blanco et al., 2023; Chen et al., 2013). Another product of this D-P-I feature structure is the conflict trials or conflict exemplars where D and P features provide conflict or inconsistent cues about the category. When there are multiple P features, conflict trials are specific for each P against D feature. During the learning phase, trials involving the non-conflict exemplars will drive selective attention away from I to D and/or P features. However, conflict exemplars particularly serve to incentivize selective attention toward D feature – the only feature that indicates correct responses on these trials. In turn, accuracies on conflict trials reflects what information people are prioritizing when making categorization decisions.

To investigate people's attention when learning category structures in a high-dimensional space, we extended the amoeba stimulus by increasing the number of arms, enabling each

amoeba to accommodate more features. We designed three types of amoebas, each with 3, 5, or 7 arms and features. Each feature still has two levels. With these stimulus sets, we created three conditions with different dimensionalities – 3, 5, or 7 dimensions. In each condition, there are one D feature and one I feature while the rest are P features. The feature-category association structure is illustrated in the Table 1.

Table 1 D-P-I feature structure, category, and trial type

Three-dimension condition

D	P	I	Category	Trial Type
0	0	0	A	Non-conflict
0	0	1	A	Non-conflict
0	0	0	A	Non-conflict
0	1	1	A	Conflict
1	1	0	B	Non-conflict
1	1	1	B	Non-conflict
1	1	0	B	Non-conflict
1	0	1	B	Conflict

Five-dimension condition

D	Pa	Pb	Pc	I	Category	Trial Type (D-Pa)
0	0	1	0	0	A	Non-conflict
0	0	0	0	1	A	Non-conflict
0	0	0	1	0	A	Non-conflict
0	1	0	0	1	A	Conflict
1	1	0	1	0	B	Non-conflict
1	1	1	0	1	B	Non-conflict
1	1	1	1	0	B	Non-conflict
1	0	1	1	1	B	Conflict

Continued

Table 1 Continued

Seven-dimension condition

D	Pa	Pb	Pc	Pd	Pe	I	Category	Trial Type (D-Pa)
0	0	1	0	0	1	0	A	Non-conflict
0	0	0	0	1	0	1	A	Non-conflict
0	0	0	1	0	0	0	A	Non-conflict
0	1	0	0	0	0	1	A	Conflict
1	1	0	1	1	1	0	B	Non-conflict
1	1	1	0	1	1	1	B	Non-conflict
1	1	1	1	0	1	0	B	Non-conflict
1	0	1	1	1	0	1	B	Conflict

2.1.2 Procedure

For this task, we supply a cover story that researchers find new species of amoebas with several organelles and group them into two categories, CEQ and GEQ. Participants need to learn categorizing them according to the common characteristics within the same group while tolerating the individual differences. The task consists of a training session, a test session, and a reflection session, with the cursor-contingent paradigm. The locations of features relative to the arms and their diagnosticity are randomly determined before the task starts and stay consistent throughout the task.

The training session involves 48 trials of six blocks, each of which contains eight amoebas of two categories while maintaining the diagnosticity of all features. The combinations of different features are counterbalanced between blocks such that each level of features is displayed the same number of times during training, and no unique combination appears extraordinarily frequent. This helps reduce the possible confounder of the exposure effect. Each training trial is followed by feedback on whether the

response is correct or not. Incorrect responses will trigger a mandatory extra waiting time of 5 seconds before the next trial can be started by pressing the spacebar. This is designed to urge people to learn and strive for correct responses. A trial will automatically end in 30 seconds with no response. The exemplar blocks of trials under each condition are shown in Table 1. The display order of trials is randomized within each block such that there are two conflict trials every eight trials.

Participants enter a break after the training session and can continue to the testing session by pressing the spacebar whenever they are ready. Before testing, participants are told that the following is to test their knowledge and contains no feedback. There are 16 trials in the testing session of the 3-dimension and 5-dimension conditions and 24 testing trials of the 7-dimension condition, all without feedback. We carefully construct the testing session so that for each P feature, half of the testing trials are non-conflict trials while the other half are conflict trials. This provides full insights into which specific P feature(s) are weighed more over the D feature during categorization. Especially under 5- and 7-dimension conditions with multiple P features, the strategy is well reflected in the performance discrepancy between these non-conflict and conflict trials based on different P features. Trials are displayed in a random order.

The testing session is followed by another break, and participants are prompted into the reflection session of some singleton trials to report how much they think each organelle (feature) is associated with the category labels. Each organelle is tested separately on the scale from CEQ to GEQ with a slider to report both the category label and how strong they believe an organelle is indicating a category. The number of singleton trials is

dependent on the number of dimensions – there are 6, 10, and 14 singleton trials respectively for 3-, 5-, and 7-dimension conditions. The singleton responses reflect the learned diagnosticity of each feature and the rule participants use to categorize amoebas, not necessarily memory of organelles. Noe that this thesis does not contain in-depth analysis of singleton trials.

The cursor traces during training and testing sessions were recorded to indicate sampling behavior during and after category learning. The task takes at most 30 minutes.

2.1.3 Data Collection

Data were collected through an online experimentation platform, Pavlovia. Participants were undergraduate students taking the Psych 1100 course in The Ohio State University in 2024 Fall. They registered for this study through the OSU Research Hour Program and could take the online task on their own device including a laptop, tablet, or others with a trackable cursor. Students received 0.5 research hour credit after completion. This study was approved by the Ohio State Behavioral and Social Sciences IRB under #2023B0365.

2.2 Data Analysis

2.2.1 Participants

We recruited 30 participants for each condition of dimension numbers, all with normal or corrected-to-normal vision. Because we especially focus on attention dynamics of young adults, we only include people between 18 and 25 years old with complete data recorded. We also excluded those whose average accuracy on the non-conflict testing trials is lower than 55%. By implementing this criterion, we ideally capture whoever learned to

categorize using either D or P features in the formal analysis and exclude those who completely relied on I feature or random guess. Finally, 23 participants under the 3-dimension condition, 19 under 5-dimension, and 19 under the 7-dimension conditions are included in the analysis and model fitting.

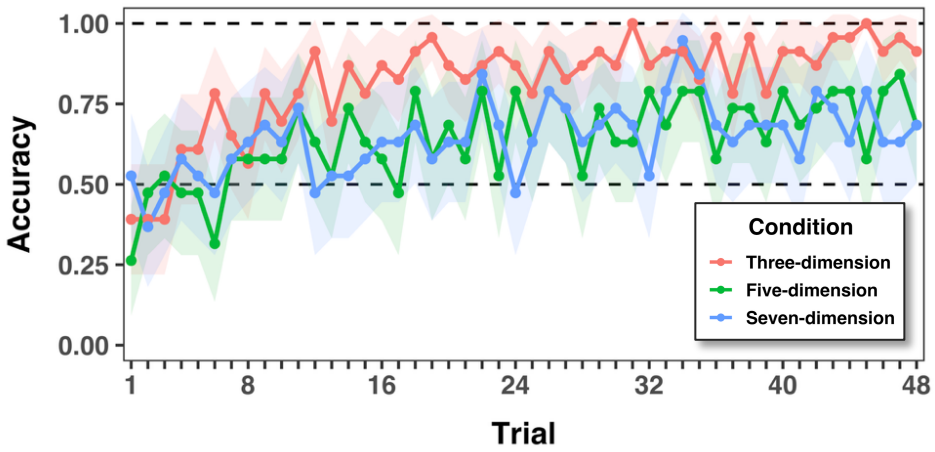
2.2.2 Cursor Traces and Sampling

We recorded the mouse traces on each training and testing trial along with the trial length and coded them into the metrics of sampling. The temporal interval between the recorded data points is unified into a second scale by proportionally mapping the total number of data points to the related trial length in seconds. Due to the experimental setting, only at most one feature can be revealed at a time point, which forces sampling to be dimension exclusive. Taking the axiomatic assumption that people move the cursor to where they are to visually sample from and the association between where one directs their eyes and what one attends to, we take the processed cursor traces as credible measures of sampling and thus distributed attention. We henceforth use the term "sampling" to denote the cursor-trace data for ease of reference.

2.3 Results

We first examined the group-level performance during the training session. Figure 3 shows that there exhibits a steady accuracy increase as the training goes on for all conditions. The accuracy of the 3-dimension group grows faster and asymptotes at a higher value than the other two groups. The general accuracy increase suggests that people gradually learned the feature-to-category map. As the number of dimensions goes

up, it becomes harder for people to hit a high accuracy with the same number of training trials. We further investigate how people adjust their sampling strategies during learning under each condition.



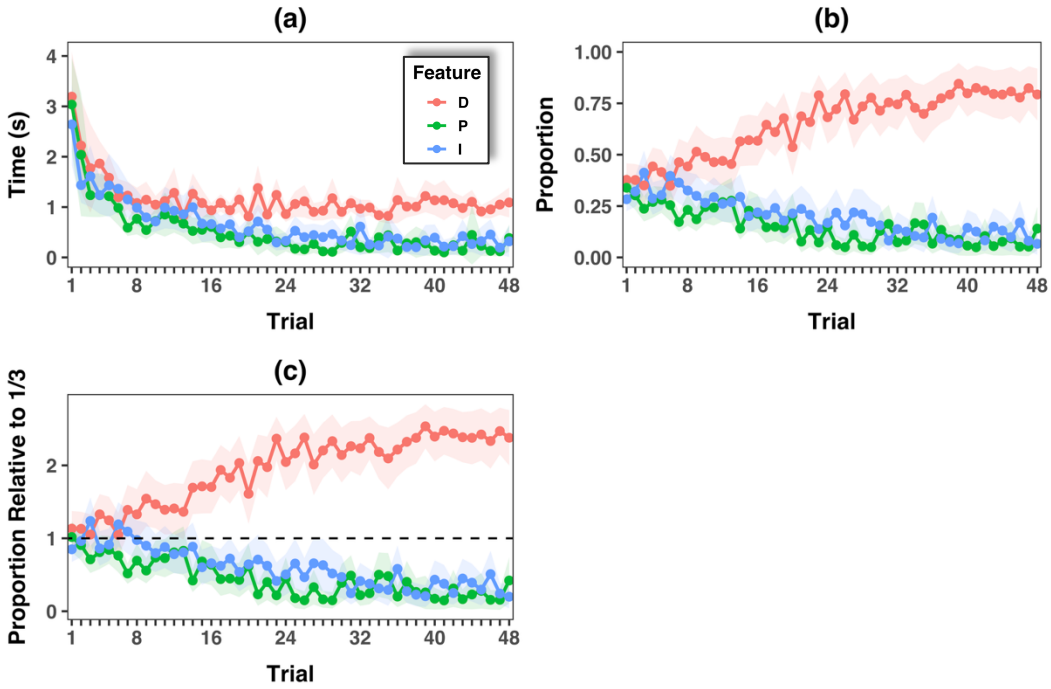
Different colors represent accuracies from different conditions. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 3 Accuracy during training under three conditions

2.3.1 Three-dimension

First, we start with the attention dynamics on a group level. Figure 4a shows that the time for sampling each feature decreases as a function of the training trial. Participants start to sample all three features for several seconds and show a particularly fast drop in sampling duration of all features during the first 10 to 15 trials. Until then, the discrepancies among features are of small magnitude. After about the 15th trial, the time of sampling D feature keeps steady while those of P and I features keep decreasing but in a slower rate. To observe the relative portion of different features, we normalize the dimension-wise time

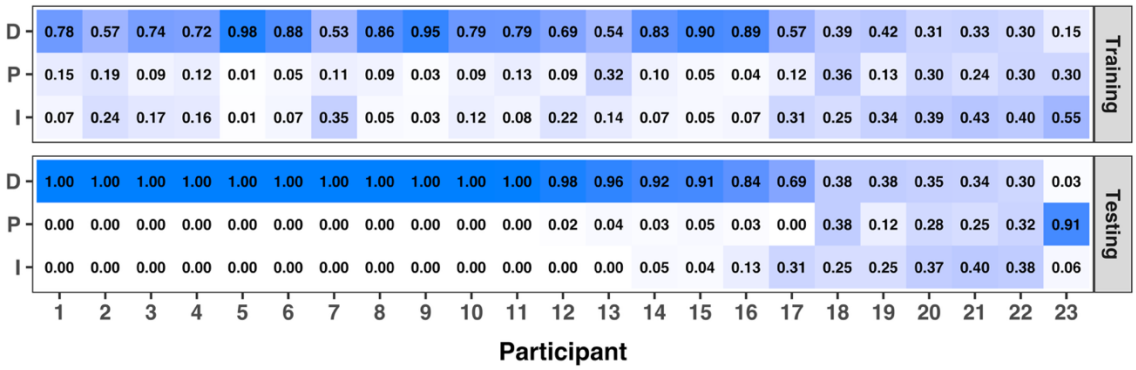
of sampling to generate the proportion of time sampling each feature on each trial. Figure 4b shows a balanced proportion of three features at the beginning. Then, D feature takes the largest proportion (over 50%) after the first 15 trials and steadily increases till the end when it takes almost 80% time of a trial. The proportions of sampling P and I features decrease over training, both fluctuating around 10% for the last 10 trials. Figure 4c shows that D feature grows to take more than the baseline of even proportions with no bias in sampling, while P and I features end up with lower-than-baseline proportions as learning.



(a) Average time of sampling each feature on each training trial. (b) Average proportion of time that people sample each feature. (c) Average proportion of time that people sample each feature relative to the base proportion if people sample all features evenly (i.e., $\frac{1}{3}$ in the three-dimension condition). Different colors represent different features. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 4 Sampling patterns over training under the 3-dimension condition

The heatmap of proportions of time sampling each feature provides more information on individual sampling profiles. In Figure 5, each column shows the average proportion of time one samples each feature on the 48 training trials (the top panel) and 16 testing trials (the bottom panel). We can observe that the division of proportions is less discrepant in training, but much more skewed for most participants during testing. Participants 18 – 22 adhere to a consistent sampling profile from training to testing, each feature taking a relatively even portion. For some people, although they balance the proportions across features during training, they heavily concentrate on one feature during the testing phase, either the D (e.g., participants 2, 7, 13) or P feature (e.g., participant 23).



The values and color shades in each square represent the average proportion of time that one participant samples a feature across all training (top panel) and testing trials (bottom panel). Each column represents one participant's data. Participants are ordered by the proportion of time for sampling D feature during testing. To decrease the impact of particularly long sampling time on some trials, we take the average on all trial-wise proportions for each participant.

Figure 5 Heatmap of sampling under the 3-dimension condition

Given distinct attention profiles in training versus testing, we take the sampling behavior in the testing session as a robust reflection of attention related to categorization decisions. Because participants know these trials are to test their knowledge without feedback, they should no longer update what features they attend to and how they weigh them. Instead, they will sample whatever they think is important according to the rule they have learned. In contrast, how people sample different features during the training phase more depicts the learning process. For example, one may spend much time on the P feature before figuring out it is not a perfect cue.

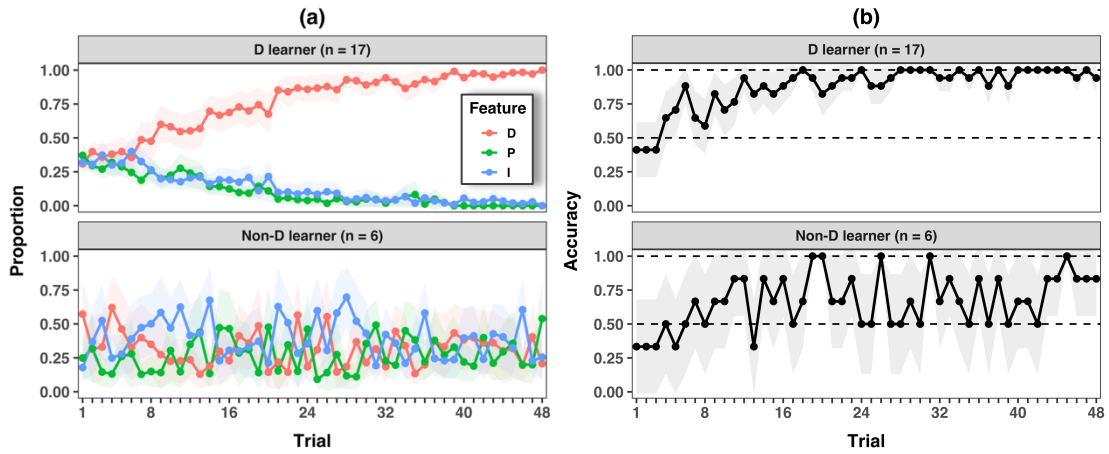
Therefore, we label the learner type for each participant according to their sampling profile in the testing session. The classification criteria are: if the average proportion of sampling one feature during testing trials is higher than 50%, they are called this-feature learner; otherwise, they are labeled as all(-feature) learner. A trivial inference is that, ideally, there should not be any I learner – people who dominantly rely on the I feature can not achieve an accuracy ≥ 55% for testing trials and should be excluded with the pre-screening criterion. This is supported by our data. The number of different types of learners under the three-dimension condition is shown in Table 2.

Table 2 Different types of learners under the 3-dimension condition

Label	**Criterion**	**Participants**
D learner	Average proportion of time sampling D feature ≥ 50%	18
P learner	Average proportion of time sampling P feature ≥ 50%	1
I learner	Average proportion of time sampling I feature ≥ 50%	0
All learner	People who fail all three criteria above	5

We then take a closer look into how different types of learners change sampling behavior over training. Note that there are only a few P or All learners, and their sampling curves are not too much distinguishable from one another. We thereby group them as non-D learners as contrast to D learners. A more proper interpretation of the labels is that D learners learn to dominantly use D to categorize amoebas while non-D learners do not dominantly use D for categorization – they may be using another feature or combinedly use several features which can involve D.

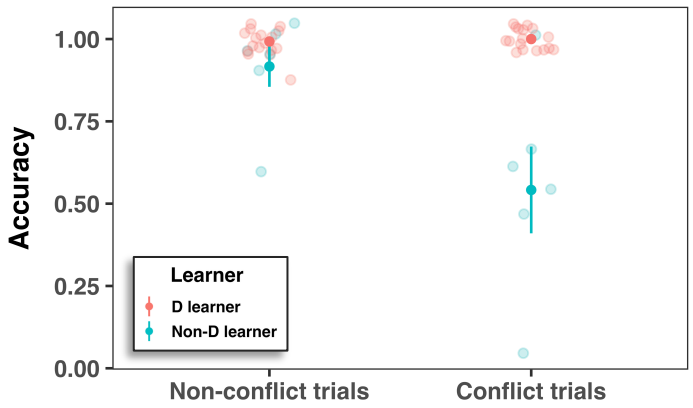
Figures 6a and 6b show the proportion of time sampling each feature and accuracy during training, with D learners in the top panels and non-D learners in the bottom panels. For D learners, the trial-wise proportion of time sampling the D feature increases until it takes almost all portions, while the proportions to P and I features start at around a decent 30% and decrease toward zero in the end. This reflects a strong learning process which results in a polarized distribution of proportions – D learners quickly and heavily direct attention resources to the most useful feature. However, non-D learners sample three features in an even manner and somehow fail to settle down on one single dimension. Those people might fail to figure out the optimal rule based on D feature but can still achieve a better-than-guess accuracy with some memory-engaged endeavors, given only three dimensions. Accordingly, the accuracy of D learners grows fast and steady over training, while the non-D learners get stuck in achieving a higher accuracy, ending up at around 75%.



(a) Average proportion of time that each learner group samples a feature on each training trial, with different color indicating different features. (b) Average accuracy of different learner groups on each training trial. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 6 Sampling and accuracy of learner groups under the 3-dimension condition

In addition to comparing sampling patterns of different learners, we also compare their accuracies on different types of testing trials. Theoretically, one can achieve comparable accuracies in both conflict and non-conflict trials if they rely on the D feature. People who fail to identify the diagnosticity of D feature but overly rely on the P feature will show a accuracy reduction in conflict trials compared to non-conflict trials. Although the number of non-D learners is small, we still observe this clear pattern in Figure 7. For D learners, the average accuracy on non-conflict and conflict testing trials are .993 and 1; yet for non-D learners, the average accuracy is.917 on non-conflict trials but only .542 on conflict trials.



Individual accuracies are represented by transparent dots while the group averages are represented by solid dots with error bars. Different colors represent different types of learners.

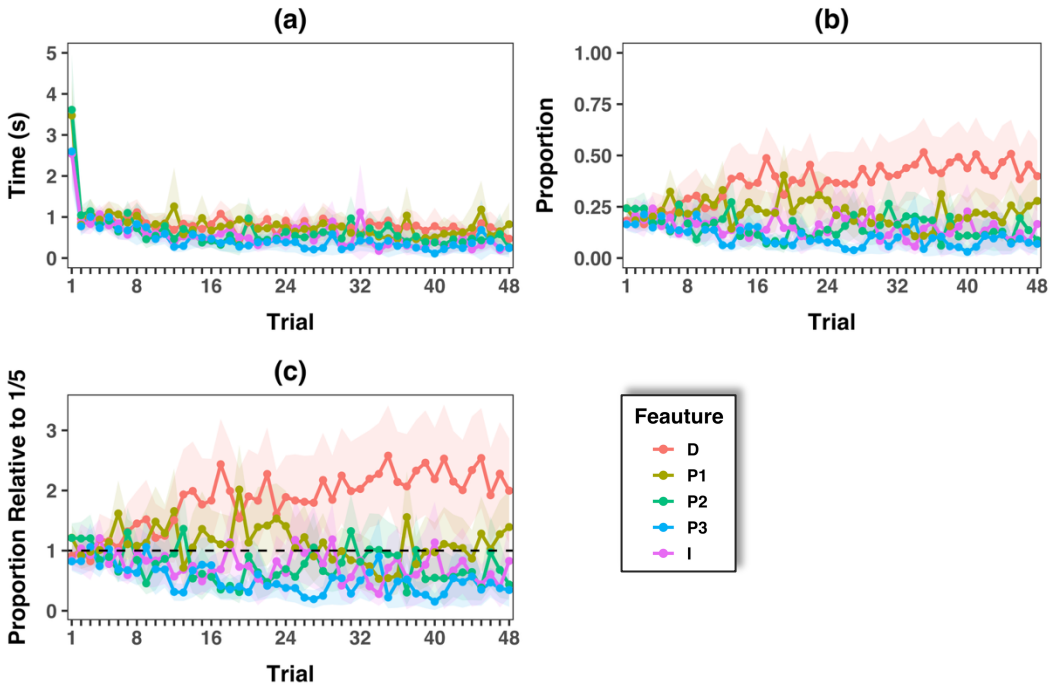
Figure 7 Accuracy of learner groups on testing trials under the 3-dimension condition

2.3.2 Five-dimension

For amoebas with five features, there are three P features with the same diagnosticity. All features are randomly sampled from a set of candidate organelles. For each participant, we label the P feature that takes the largest proportion of time being sampled in the training phase as P1 feature, and so forth as P2 and P3 features. The order of P features indicates the most, second, and third preferred feature for each person when they sample and learn during the training phase. Because these preferences are subjective and idiosyncratic, this rank-order procedure retains meaningful sampling patterns when aggregating over the participants.

Figure 8 shows the group-level change in time durations and proportions of sampling each feature across training trials. According to Figure 8a, the trial-wise duration of each feature fast decreases during the first few trials and then diverges, some staying relatively

flat yet some dropping mildly. Figure 8b better illustrates the diverging pattern of feature-specific proportion. All features begin taking an even amount of sampling around 20%. As training continues, D feature takes increasing proportions up to 40% ~ 50% while others decay. A special pattern is that in the first half of training, P1 is quite competitive and earns a larger proportion of sampling than others even though it does not exceed the share of D feature. With more training, the D feature wins over P1. Compared to the three-dimension condition where the D feature takes an overwhelming amount, the five-dimension group on average struggles with exploring a larger number of features and concentrating on a small subset of information. Figure 8c shows that, as participants go through the training phase, D feature takes a larger-than-even proportion of time being sampled, the proportion of P1 fluctuates around the baseline proportion, while other features are generally suppressed below the baseline proportion in the later trials of the training phase.

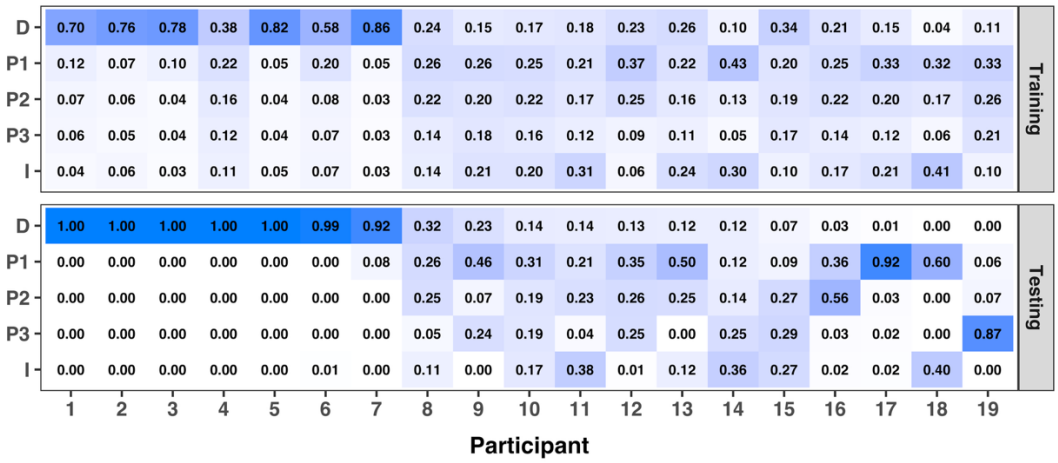


(a) Average time of sampling each feature on each training trial. (b) Average proportion of time that people sample each feature. (c) Average proportion of time that people sample each feature relative to the base proportion if people sample all features evenly (i.e., $\frac{1}{5}$ in the five-dimension condition). Different colors represent different features. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 8 Sampling patterns over training under the 5-dimension condition

The individual sampling profiles are displayed in Figure 9. Similar as the three-dimension data, the individual proportion distributions during the training phase show less dispersion than those during testing. This implies that people use a distributive strategy of attention when learning and exploring. However, compared to the three-dimension data, there are fewer participants who almost only sample the D feature during the testing phase (n = 17 out of 23 in 3-dimension data, n = 7 out of 19 in 5-dimension data). Instead, more of them sample multiple features on the testing trials. Also, some

people display different sampling patterns between two sessions. For example, participant 19 strongly prioritizes the P3 feature in testing trials that is least sampled from in the training phase.



The values and color shades in each square represent the average proportion of time that one participant samples a feature across all training (top panel) and testing trials (bottom panel). Each column represents one participant's data. Participants are ordered by the average proportion of time for sampling D feature during testing.

Figure 9 Heatmap of sampling under the 5-dimension condition

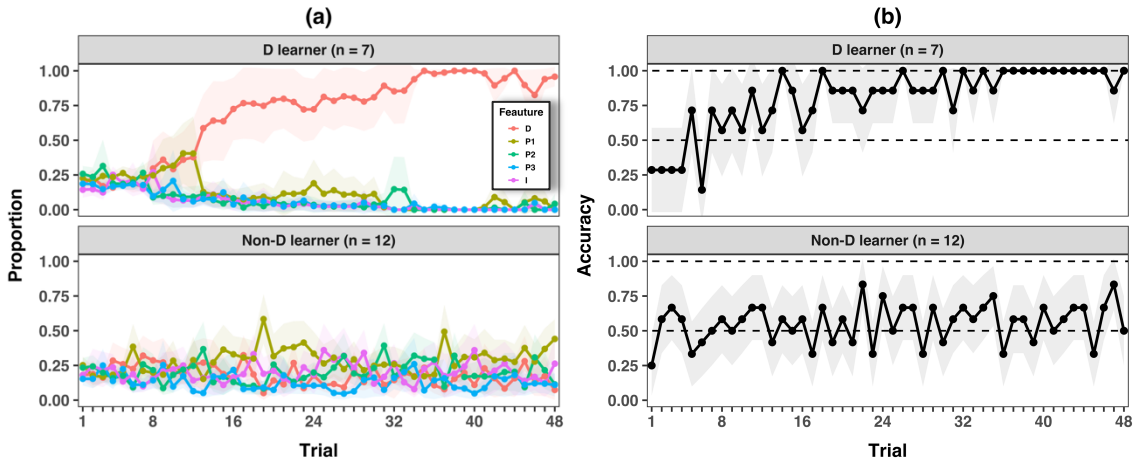
Based on these results, we group participants into different learner types using a similar criterion from the three-dimension condition. To accommodate the multiple P features, we define the P learner as people with the largest proportion over testing trials occurring on a P feature and the total proportion of all P features larger than 50%. This allows us to identify those who combinedly use multiple P features as P learners. Table 3 shows that, of all 19 participants, there are seven D learners and ten P learners with the rest as All learners. The number of people who predominantly sample D feature during learning,

i.e., ideally those who figure out the diagnosticity of D feature (n = 7), is substantially smaller than that in the three-dimension condition (n = 18).

Table 3 Different types of learners under the 5-dimension condition

Label	Criterion	Participants
D learner	Average proportion of time sampling D feature ≥ 50%	7
P learner	Sum of proportions of time sampling all P features ≥ 50% and the largest proportion occurs on a P feature	10
I learner	Average proportion of time sampling I feature ≥ 50%	0
All learner	People who fail all three criteria above	2

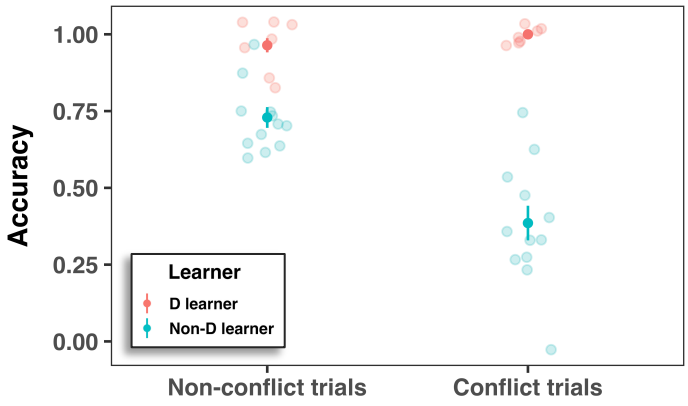
Figure 10a shows the proportion of time that D learner versus non-D learner groups sample each feature. For non-D learners, the proportions are relatively evenly distributed among features throughout training. Sometimes, P1 or P2 feature stands out and takes a bit larger proportion. For D learners, the proportion to D feature increases up to 100% while those to P and I features decreases to 0%. The results of different sampling profiles are reflected in Figure 10b, where $n = 7$ D learners almost achieve 100% accuracy at the end of training, while non-D learners struggle to maintain an accuracy over 60%. A notable difference in D-learners of the five- and three-dimension conditions is that during first ten trials, the proportion of P1 raises parallel to D and then gets suppressed. This is a unique characteristic with more dimensions and more P features in the information space – P features serve as a more tempting lure to grab attention during category learning.



(a) Average proportion of time that each learner group samples a feature on each training trial, with different color indicating different features. (b) Average accuracy of different learner groups on each training trial. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 10 Sampling and accuracy of learner groups under the 5-dimension condition

On the other hand, the accuracy discrepancy between the non-conflict and conflict testing trials supports that people are being trapped by P features more in this condition. Figure 12 shows an interaction between the trial type and learner type on the accuracies. D learners' accuracies on both types of trials are equally high (non-conflict = .964, conflict = 1), while non-D learners on average only do good in non-conflict trials with the accuracy at .729 but not conflict testing trials with a low accuracy at .385.



Individual accuracies are represented by transparent dots while the group averages are represented by solid dots with error bars. Different colors represent different types of learners.

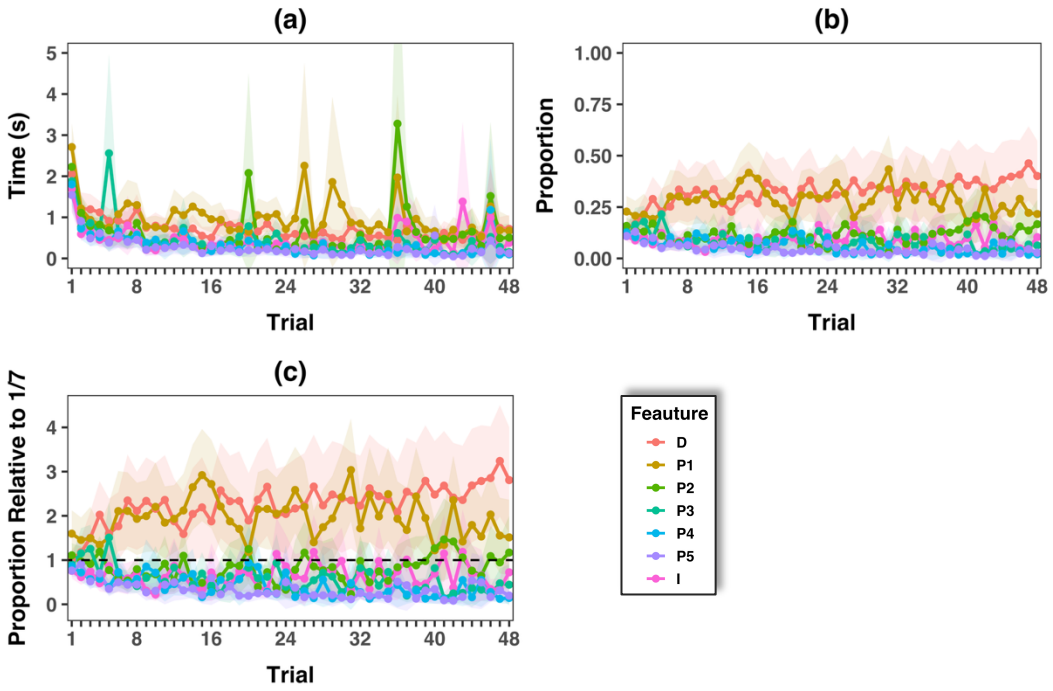
Figure 11 Accuracy of learner groups on testing trials under the 5-dimension condition

2.3.3 Seven-dimension

The seven-dimension condition imposes especially heavy cognitive demands as the sheer quantity of feature information intensely challenges the limit of processing capacity. We also order P features from P1 (the most favored during training) to P5 (the least favored during training) for each participant as what we do for the five-dimension data.

Figure 12 reveals some unique patterns in the time and proportion of sampling on training trials. The time durations of sampling each feature still decrease at the beginning like the previous conditions, while P1 instead of D feature generally takes a longer duration later. Moreover, the sampling time for features P1 and P2 shows intermittent spikes even in the later training phase (Figure 12a). Figure 12b shows that in addition to an increase in the proportion of D feature as in it appears in the three- and five-dimension conditions, the proportion of P1 emerges to pop up alongside D feature, almost taking a comparable

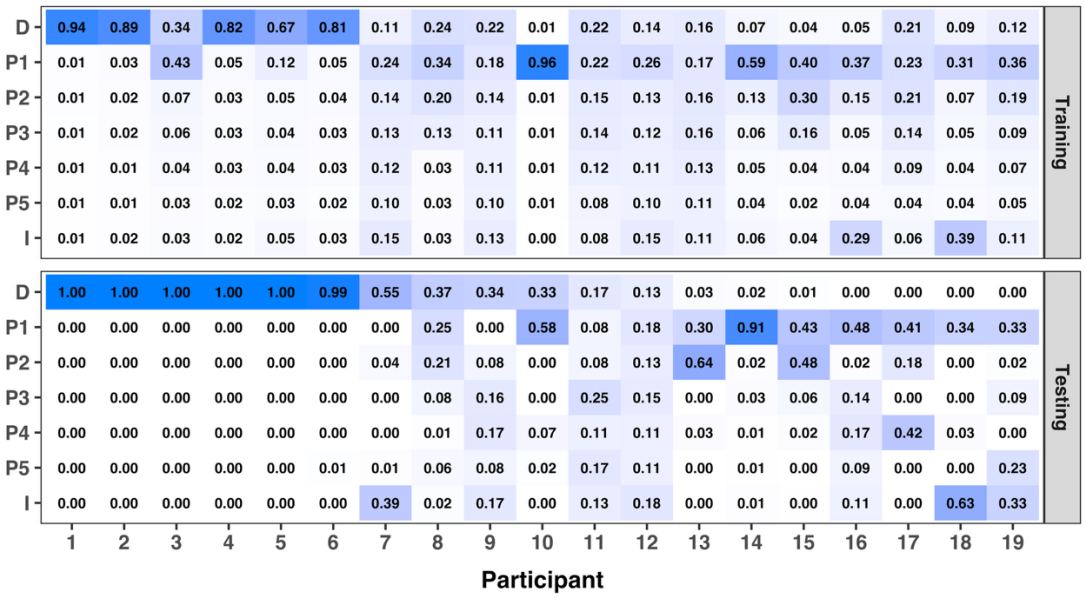
proportion as the D feature. This supports the argument that people are more likely to sample P features more and get trapped in cases of large numbers of dimensions. Figure 12c displays the ratio of the proportion of time sampling each feature relative to the baseline proportion if we assume participants show an even sampling strategy. It is more clearly shown that the ratios of D and P1 features grow over one, P2 (and I) feature(s) fluctuate around 1, while other features are suppressed lower than one.



(a) Average time of sampling each feature on each training trial. (b) Average proportion of time that people sample each feature. (c) Average proportion of time that people sample each feature relative to the base proportion if people sample all features evenly (i.e., $\frac{1}{7}$ in the seven-dimension condition). Different colors represent different features. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 12 Sampling patterns over training under the seven-dimension condition

Figure 13 displays participant's overall sampling partition over testing and training trials. In the training panel, more people sample a subset of features instead of evenly sampling all features as they do in three- and five-dimension data. This is consistent with the quest for simplicity to reduce the number of dimensions during learning (Galdo et al., 2022). In the testing panel, a small number of participants sample the D feature the most while more of them are overwhelmed by one or several P features. We also observe participants with inconsistent sampling profiles from training to testing (e.g., participants 3, 7, 13). These profiles will motivate our modeling decisions in Chapter 3 about the differences between the sampling probability and decision weight – as in the testing phase when one is told to use their knowledge, the proportion distribution should mostly contain information about the decision weight.



The values and color shades in each square represent the average proportion of time that one participant samples a feature across all training (top panel) and testing trials (bottom panel). Each column represents one participant's data. Participants are ordered by the average proportion of time for sampling D feature during testing.

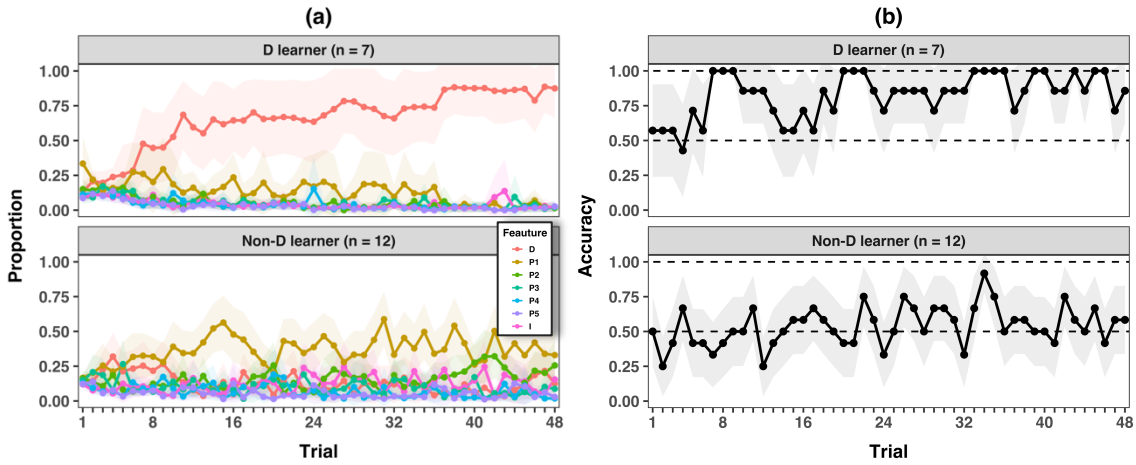
Figure 13 Heatmap of sampling under the seven-dimension condition

In Table 4, we still classify participants into four groups following the same criterion as the five-dimension condition. Similarly, we put P, I, and All learners together as a non-D learner group to compare the proportions of sampling each feature versus D learners. As Figure 14a shows, even D learners are not as efficient to sample the most useful D feature as those in the three- and five-dimension conditions. The proportion to the D feature does not reach 80% until the last ten trials while the P1 feature recurrently pops out in the training process. Notably, the proportions of P2~P5 and I features are strongly suppressed to nearly zero. This is evidence of attention constraints that reduce visual processing all information down to a small subset, which is D and P1 feature in this context. For the

non-D learners, it turns out that P1 is the dominant feature with the largest proportion, although its role is not as overwhelming as the D feature for D learners. Growth of P1's proportion is sluggish and levels off with fluctuations. These sampling patterns are mirrored in accuracies as well. In Figure 14b, we see that D learners' average accuracy increases with fluctuations and finally reaches 80%. However, non-D learners show weaker learning outcomes and lack of significant accuracy improvement, probably due to individual asynchronization of correct responses.

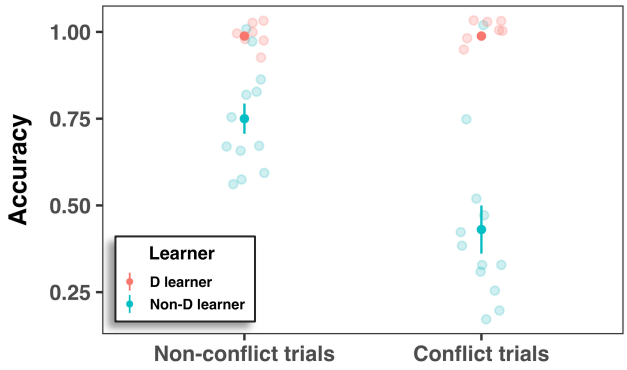
Table 4 Different types of learners under the 7-dimension condition

Label	Criterion	Participants
D learner	Average proportion of time sampling D feature ≥ 50%	7
P learner	Sum of proportions of time sampling all P features ≥ 50% and the largest proportion occurs on a P feature	9
I learner	Average proportion of time sampling I feature ≥ 50%	1
All learner	People who fail all three criteria above	2



(a) Average proportion of time that each learner group samples a feature on each training trial, with different color indicating different features. (b) Average accuracy of different learner groups on each training trial. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 14 Sampling and accuracy of learner groups under the 7-dimension condition

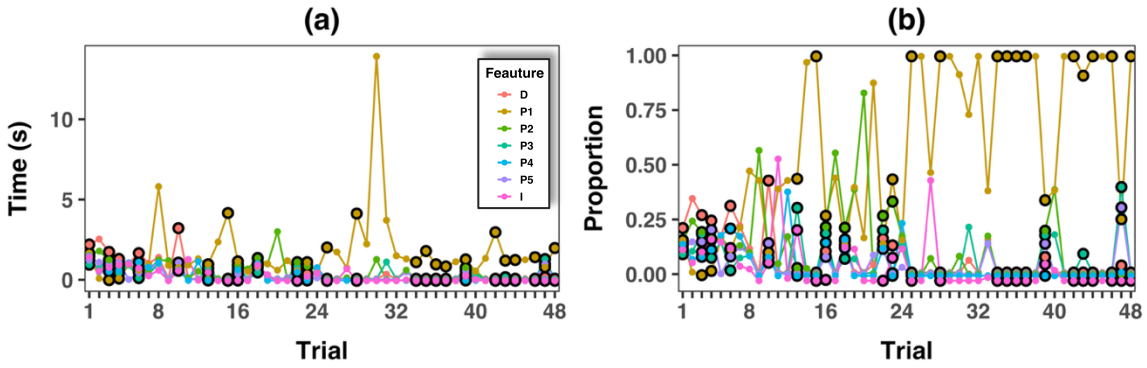


Individual accuracies are represented by transparent dots while the group averages are represented by solid dots with error bars. Different colors represent different types of learners.

Figure 15 Accuracy of learner groups on testing trials under the 7-dimension condition

The testing accuracy plot in Figure 15 also demonstrates that non-D learners are trapped. D learners' accuracies in both non-conflict and conflict trials are .988, while non-D learners are doing much worse in conflict trials (accuracy = .431) than non-conflict trials (accuracy = .75). Sampling P more than D feature hurts the accuracy when D and P features provide conflict information.

In addition to group-level analysis, it is valuable to look at individual sampling curves to have a concrete idea of how learners are trapped by P features. Figure 16 displays a non-D learner's data from the seven-dimension condition. This person's behavior of sampling P1 feature is reinforced by some correct responses after sampling it such as trial 25, 34 – 37, 42 – 44. They then decide to continue prioritize that P dimension, only moving away from it right after an error while returning to sampling it in most cases.

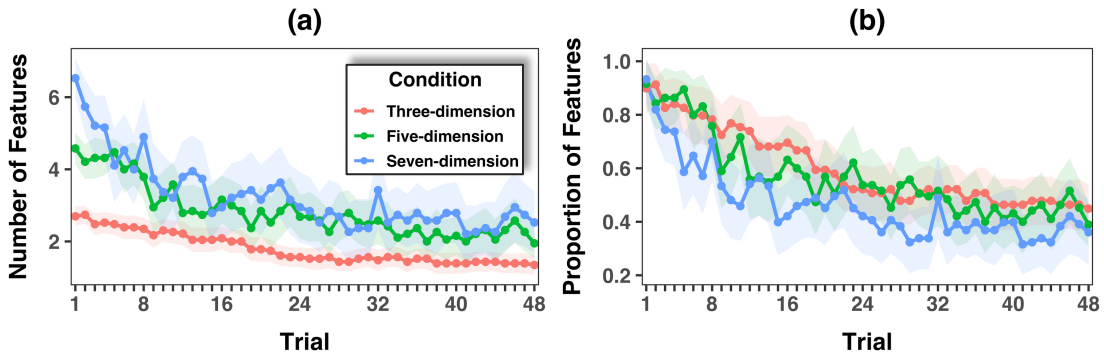


(a) Time (s) of sampling each feature. (b) Proportion of time sampling each feature within each trial. Different colors represent different features. Trials with a correct response are marked with black frames.

Figure 16 One participant's sampling pattern over training

2.3.4 Summary

These behavioral results are valuable to our theoretical arguments that selective attention and the dilemma between accuracy and simplicity goals are heavily induced in conditions with high information dimensionalities. In addition to behaviors under each of the three conditions, Figure 17 shows an overall comparison of the number of features being sampled and the relative proportion of features on each trial among all conditions. We can observe that participants in all three groups show a tendency of decreasing the number of dimensions being sampled as they learn. Particularly in the seven-dimension group, participant constrain the number of dimensions strongly such that the proportion of features being sampled falls below the other two conditions (Figure 17b).



(a) Number of features being sampled on each trial, averaging over all participants under each condition. A feature is defined as being sampled on certain when the sampling time is larger than a threshold of 200ms. (b) Proportion of features being sampled on each trial, averaging over all participants under each condition. The proportion of features are obtained through dividing the number of sampled features by the total number of available features.

Figure 17 Features sampled over training

We then unravel these two goals of selective attention with our modeling framework where attention is optimized to both maximize accuracy and minimize cognitive efforts, which realizes a balance of the goal of performance and simplicity.

Chapter 3. Model Structure

Under the framework of Adaptive Attention Representation Models (AARMs, Galdo et al., 2022), this variant elaborates the partial encoding mechanism (King, 2024) and builds in attention constraints with a cost function that includes generalized regularization. In this chapter, we specify the mathematical functions of cognitive processes in our model and provide pivotal simulations to justify the model's capabilities of producing diverse patterns of attention in complex environments.

3.1 Mathematical Specifications

3.1.1 Similarity-Based Categorization

Our model inherits the essential formulation of the GCM where categorization decisions are based on the degree to which a stimulus resembles (i.e., is similar to) previous stored exemplars. In a task to classify the probe i – item on the i-th trial – with D dimensions into one of C categories, $\boldsymbol{e^{(i)}} = \left[e_1^{(i)},\ e_2^{(i)},\ \ldots,\ e_D^{(i)}\right]$ is the vector of the probe with $e_j^{(i)}$ representing the numerical level of dimension j of probe i. Similarly, past exemplars – item on the n-th trial – coded as $\boldsymbol{x^{(n)}} = \left[x_1^{(n)},\ x_2^{(n)},\ \ldots,\ x_D^{(n)}\right]$ are stored in the memory along with their true category $F^{(n)}$. The attention vector related to probe i is denoted as $\boldsymbol{\alpha^{(i)}} = \left[\alpha_1^{(i)}, \alpha_2^{(i)}, \ldots, \alpha_D^{(i)}\right]$ and the attention to exemplar n is $\boldsymbol{\alpha^{(n)}} = \left[\alpha_1^{(n)}, \alpha_2^{(n)}, \ldots, \alpha_D^{(n)}\right]$.

To classify the i-th probe, one compares it with all previously stored exemplars along each dimension, which result in a dimension-wise exemplar-probe distance.

$$d_j\left(e^{(i)}, x^{(n)}\right) = \left|e_j^{(i)}, x_j^{(n)}\right|$$

Attention gates which dimensions are processed by working memory and can thereby be stored in long-term memory for future retrieval. Consequently, when probe i is compared to exemplar n, attention to $\boldsymbol{e^{(i)}}$ informs which dimensions are prioritized in categorization decisions – $f\left(\alpha_j^{(i)}\right)$, while memory of an exemplar $x^{(n)}$ determines which dimensions can be retrieved from memory to further engage in decisions – $g\left(\alpha_j^{(n)}\right)$. This is the basic mechanism of partial encoding (King, 2024). In this sense, the activation of dimension j of $\boldsymbol{x^{(n)}}$ induced by probe $\boldsymbol{e^{(i)}}$ in the psychological space is an exponential function of the objective distance modulated by attention (encoding strength) to the probe and memory of the exemplar.

$$a_j\left(\boldsymbol{e^{(i)}}, \boldsymbol{x^{(n)}}\right) = e^{-f\left(\alpha_j^{(i)}\right) d_j\left(e^{(i)}, x^{(n)}\right) g\left(\alpha_j^{(n)}\right)}$$

Given the dimension-wise activation, the exemplar-wise activation related to the probe i is calculated through the multiplicative pooling rule (Nosofsky, 1986).

$$a\left(\boldsymbol{e^{(i)}}, \boldsymbol{x^{(n)}}\right) = \prod_j a_j\left(\boldsymbol{e^{(i)}}, \boldsymbol{x^{(n)}}\right)$$

According to the Luce choice rule, the predicted probability of probe i as category c is the total activation of exemplars under category c relative to the total activation of all past exemplars. In this calculation, $\mathbb{I}_{\{F^{(n)}=c\}}$ is an indicator function that returns one when the event within the braces $F^{(n)} = c$ is true, i.e., $\boldsymbol{x^{(n)}}$ belongs to the Category c.

$$P^{(i)}(c) = \frac{\sum_n a\left(\boldsymbol{e}^{(i)}, \boldsymbol{x}^{(n)}\right) \mathbb{I}_{\{F^{(n)}=c\}}}{\sum_k \sum_n a(\boldsymbol{e}^{(i)}, \boldsymbol{x}^{(n)}) \mathbb{I}_{\{F^{(n)}=k\}}}$$

Therefore, the response $\boldsymbol{R}^{(i)}$ is predicted through a multinomial distribution defined by the vector of probability of response, $\boldsymbol{P}^{(i)} = \left[P^{(i)}(1), P^{(i)}(2), \ldots, P^{(i)}(C)\right]$.

$$\boldsymbol{R}^{(i)} \sim multinomial(\boldsymbol{P}^{(i)})$$

In addition to categorization responses, our model can further predict on sampling data. Let $t_j^{(i)}$ denote the sampling duration on the dimension j of probe i and $s_j^{(i)}$ denote the binary outcome of sampling or not. If the duration exceeds certain threshold of information processing t_0, we believe that people sample information from that dimension of that item ($s_j^{(i)} = 1$). Otherwise, they do not sample it either due to never looking at it or a lack of time to encode it.

$$s_j^{(i)} = \begin{cases} 1, & t_j^{(i)} > t_0 \\ 0, & t_j^{(i)} \leq t_0 \end{cases}$$

Provided the association between attention and gaze, our model maintains that sampling is primarily guided by attention through a probabilistic connection: the higher $\alpha_j^{(i)}$ is, the more likely one visually samples dimension j and encodes it. Sampling is also exposed to physical properties of features or individual preferences, e.g., people look at visually salient features more, people look at features on preferred locations more, etc. To characterize the gaze-attention association with additional variations, we use a two-parameter logistic function to compute the probability of sampling – $psamp_j^{(i)}$ as a function of attention, where the slope term b_1 reflects the sensitivity of sampling behavior

regarding attention while the intercept term $b_{0,j}$ captures the dimension specific factor, such as individual preferences or salience.

$$psamp_j^{(i)} = \frac{1}{1 + e^{-\left(b_{0,j} + b_1 \alpha_j^{(i)}\right)}}$$

Therefore, we can predict whether a dimension is sampled or not on trial i with a random draw from the binomial distribution parameterized by $psamp_j^{(i)}$.

$$s_j^{(i)} \sim binomial\left(n = 1, psamp_j^{(i)}\right)$$

3.1.2 Choice of Linking Functions

In addition to these foundations above, the linking function f that maps attention from the entire real line to a latent space of a positive scale and the related g are yet to decide. This thesis introduces two linking functions $f: \mathbb{R} \to \mathbb{R}_{>0}$, exponential and logistic functions that have different functional properties that effect the gradient space of α. Correspondingly, the function g that maps attention on the probe i to memory strength is decided based on the specific f so that the numeric magnitudes of $f\left(\alpha_j^{(i)}\right)$ and $g\left(\alpha_j^{(n)}\right)$ are comparable.

1. Logistic linking function

Logistic linking function maps $\mathbb{R}$ to decimal values on $(0, 1)$. In this sense, the larger value of $\alpha_j^{(i)}$ takes on the real line, the higher $f^1\left(\alpha_j^{(i)}\right)$ will be, meaning that people will assign higher decision weights to the j-th dimension when classifying the i-th probe (Nosofsky, 1986). Because we take learners as rational cognitive entities, this logistic

function is invariant to external factors but only relies on the utility of a dimension for categorization.

$$f^1\left(\alpha_j^{(i)}\right) = \frac{1}{1 + e^{-\alpha_j^{(i)}}}$$

To highlight in our model, the decision weight $f^1\left(\alpha_j^{(i)}\right)$ is different from the sampling weight $psamp_j^{(i)}$ which involves two additional sets of parameters – the intercepts and slope – to capture non-decision factors only for sampling. Instead, we relate the memory strength of previous exemplars to how much people sample that information, i.e., sampling weights. The probability of sampling one dimension of an item reflects the extent to which that piece of information enters the memory, no matter driven by the "goal-related" attention or external factors, characterized by $\boldsymbol{b_0} = \left[b_{0,1}, \ldots, b_{0,D}\right]$ and b_1. To maintain a comparable scale, we also apply a logistic g function. In this case, $g^1\left(\alpha_j^{(n)}\right)$ is the same as $psamp_j^{(n)}$.

$$\mathrm{g}^1\left(\alpha_j^{(n)}\right) = \frac{1}{1 + e^{-\left(b_{0,j} + b_1 \alpha_j^{(n)}\right)}}$$

Correspondingly, a decay parameter δ is introduced to allow flexibility when projecting the stimulus space onto the psychological space. Large values of δ enlarge the magnitude of difference in dimension activation, resulting in a steeper gradient field and thus a classifier relying more on neighborhood exemplars (Nosofsky, 2011).

$$a_j\left(e^{(i)}, x^{(n)}\right) = e^{-\delta f^1\left(\alpha_j^{(i)}\right) d_j\left(e^{(i)},\, x^{(n)}\right) g^1\left(\alpha_j^{(n)}\right)}$$

2. Exponential linking function

The exponential linking function maps $\mathbb{R}$ onto the full positive line $\mathbb{R}_{>0}$. This allows $f^2\left(\alpha_j^{(i)}\right)$ to go beyond 1 and may enhance the discriminability of the model. A higher $\alpha_j^{(i)}$ will result in a more intense shrinkage of the psychological space of similarity along that attended dimension *j*, which makes the dimension *j* more discriminable.

$$f^2\left(\alpha_j^{(i)}\right) = e^{\alpha_j^{(i)}}$$

To maintain a comparable scale resulted from f and g and to match the memory strength with the sampling weight, we define $g^2\left(\alpha_j^{(n)}\right)$ as an exponent as well.

$$g^2\left(\alpha_j^{(n)}\right) = e^{b_{0,j}+b_1\alpha_j^{(n)}}$$

Also, the natural wide magnitude of exponents does not necessitate the decay parameter when calculation similarity or activation.

$$a_j\left(e^{(i)}, x^{(n)}\right) = e^{-f^2\left(\alpha_j^{(i)}\right)d_j\left(e^{(i)},\, x^{(n)}\right)g^2\left(\alpha_j^{(n)}\right)}$$

The exponential function allows explosively increase when $\alpha_j^{(i)} > 0$, which results in steep gradient for in certain support of $\alpha_j^{(i)}$.

3.1.3 Attention Constraints in Loss

According to AARM framework, instances are accumulatively stored into memory as the strength-based representations $\left[\boldsymbol{e}^{(i)}, \boldsymbol{\alpha}^{(i)}\right]$, where strengths are constrained by attention. These attention vectors $\boldsymbol{\alpha}^{(i)}$ are updated through feedback-based learning to minimize certain loss.

Originally, the loss function is defined as cross entropy, $\mathbb{CE}$, the negative log of probabilities of correct response (i.e., accuracy), and minimizing loss is to maximize the probability of being correct.

$$\mathbb{CE} = -log\big(P(correct)\big)$$

However, as we argue thus far, people pursue error minimization but also computational simplicity. Mathematically, $\boldsymbol{\alpha}^{(i)} = \left[\alpha_1^{(i)}, \alpha_2^{(i)}, \dots, \alpha_D^{(i)}\right]$ should be updated such that a subset (i.e., lower number of attended dimensions) of $\alpha_j^{(i)}$ of small values (i.e., lower amount of attention) will realize a sufficiently low cross entropy.

This pattern in category learning is like the sparse representation problems in statistics – researchers search for a small subset of regressors with sparse coefficients that largely minimize residual squared errors with ℓ_p regularization (Tibshirani, 1996). In theoretical and applied science, ℓ_p regularization is an effective algorithm to penalize the complexity in regression (Donoho & Elad, 2003; Fu, 1998; Hu et al., 2017; Tibshirani, 2011).

Generally, the ℓ_p norm term of coefficients $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_p = \left(\sum_j \beta_j^p\right)^{\frac{1}{p}}$, is added to the original loss function. Thereby, the sparse optimization problem to minimize error turns into an unconstrainted optimization process to minimize the modified loss $\boldsymbol{\mathcal{L}}$.

$$\boldsymbol{\mathcal{L}} \coloneqq error + w\|\boldsymbol{\beta}\|_p$$

In this loss function, w is the parameter that reflects the tradeoff between minimizing error and penalizing complexity. Minimizing this specific loss prevents excessively increasing coefficients after certain point, when adding more coefficients or enlarging

current coefficients will no longer reduce the error very much but, instead, increase the ℓ_p norm of coefficients.

In the context of categorization, we adopt this approach and add the ℓ_p norm of attention into the loss function. Specifically, the ℓ_p regularization in our model penalizes the probability of sampling to minimize the physical efforts caused by eye movements. In other words, $\left\|psamp^{(i)}\right\|_p$ reflects the sampling cost on the i-th trial.

$$\left\|psamp^{(i)}\right\|_p = \left(\sum_j \left(psamp_j^{(i)}\right)^p\right)^{\frac{1}{p}}$$

$$= \left(\sum_j \left(\left[1 + e^{-\left(b_{0,j} + b_1 \alpha_j^{(i)}\right)}\right]^{-1}\right)^p\right)^{\frac{1}{p}}$$

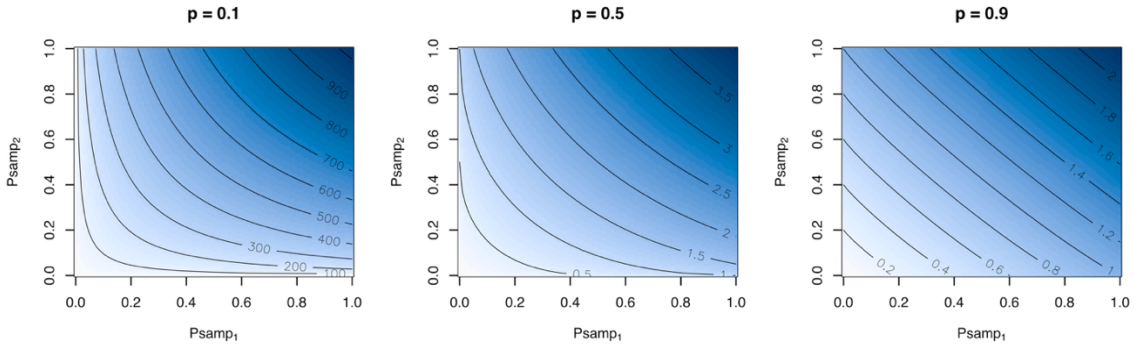
Then, the loss function is defined as a weighted sum of cross entropy and sampling costs – ℓ_p norm of sampling probabilities. w is the weight of the simplicity goal of minimizing the sampling costs relative to the accuracy goal of maximizing performance.

$$loss\left(\alpha_j^{(i)}\right) = \mathbb{CE} + w\left\|\boldsymbol{psamp}^{(\boldsymbol{i})}\right\|_p$$

$$= -log\left(P(correct)\right) + w\left(\sum_j \left(\left[1 + e^{-\left(b_{0,j} + b_1 \alpha_j^{(i)}\right)}\right]^{-1}\right)^p\right)^{\frac{1}{p}}$$

We maintain that by varying p values within $p < 1$, the ℓ_p norm on sampling probabilities can produce various attention constraints. Figure 18 shows a series of simulated values of $\|\boldsymbol{psamp}\|_p$ represented by the blue shades in the two-dimensional attention space. When $p < 1$, the equal-value contours of $\|\boldsymbol{psamp}\|_p$ are concave and naturally cause a low amount of total attention. As p approximates zero such as $p = .1$, $\|\boldsymbol{psamp}\|_p$ favors either the equally-small psamp_1 and psamp_2 or a large $psamp$

constrained to only one dimension – this realizes the dimension reduction tendency. Also, small p values can cause a similar effect as competition, without the dimension-exclusive assumption that increasing attention in one dimension urges the same amount of decrease in others as the sum-to-constant assumption. These results support that the variation of $p < 1$ allows different aspects of the simplicity in attention. Also, according to Galdo et al. (2022)'s work, among all values of $p > 0$, models with structural constraints of $p < 1$ perform the best. Both of these inform our decision to take p as a free parameter < 1 during the model fitting.



In this two-dimensional case, $\boldsymbol{psamp} = [psamp_1, psamp_2]$ and $\|\boldsymbol{psamp}\|_p = \left(psamp_1^p + psamp_2^p\right)^{1/p}$. These three plots contain results from $p = .1, .5, .9$. As p shrinks toward zero, the contours of $\|\boldsymbol{psamp}\|_p$ is more concave.

Figure 18 Contour plots of ℓ_p-norm of sampling probabilities

3.1.4 Optimizing Constrained Attention Through Gradient Descent

Framing learning as an optimization problem, searching for an attention state that minimizes $\boldsymbol{\mathcal{L}}$ will realize both goals of accuracy and simplicity. Our model assumes that loss function is minimized through the gradient descent on $\alpha_j^{(i)}$ across trials.

$$\alpha_j^{(i+1)} = \alpha_j^{(i)} - \gamma_0 \frac{\partial}{\partial \alpha_j^{(i)}} loss\left(\alpha_j^{(i)}\right)$$

$$= \alpha_j^{(i)} + \gamma_0 \frac{\partial}{\partial \alpha_j^{(i)}} log\big(P(correct)\big) - \gamma_0 w \frac{\partial}{\partial \alpha_j^{(i)}} \left\| psamp^{(i)} \right\|_p$$

During the training phase with feedback, our model can capture the update of $\boldsymbol{\alpha^{(i)}}$ and thus predict the response probability $P^{(i)}(c)$ and dimension-wise sampling probability $psamp_j^{(i)}$ for each trial. The analytical solution to derivatives of loss is provided in the Appendix A.

3.2 Model Simulations

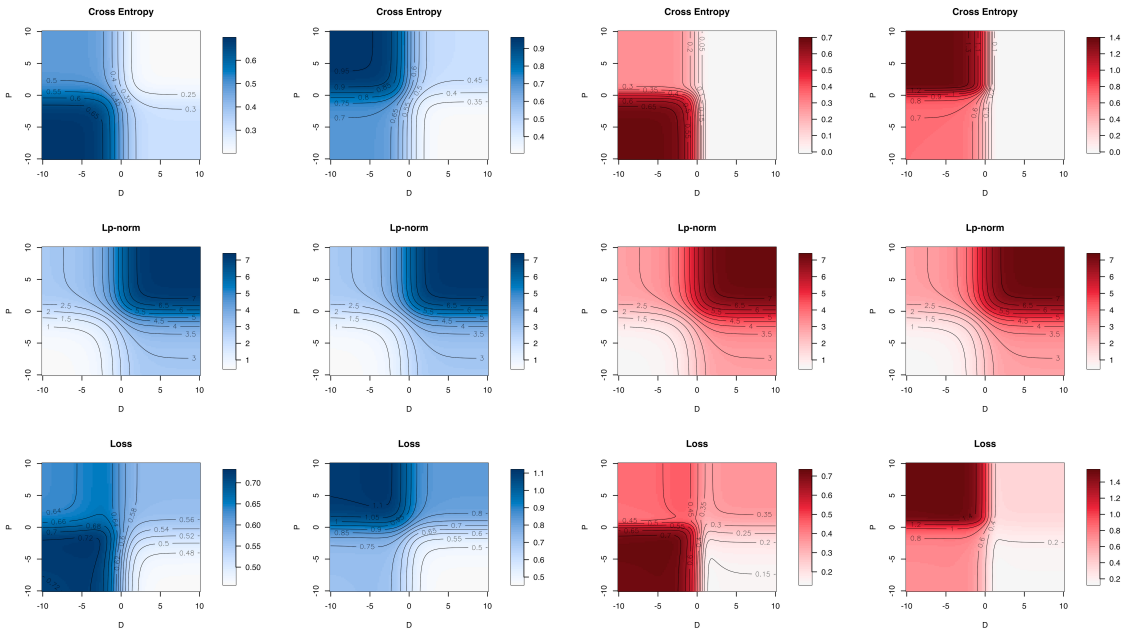
Taking our categorization task set-up (e.g., D, P, and I features, 48 learning trials, etc.), we run a series of simulations to compare the variation of loss and its components regarding different attention vectors and the gradient descent process with different parameterization.

3.2.1 Loss and Its Compositions

We simulate how values of cross entropy, ℓ_p norm of sampling costs, and their weighted combination, loss, vary with attention to D and P dimensions (α_D, α_P).

Figure 19 shows that when $w = .05$, loss is primarily characterized by the cross entropy and modulated by ℓ_p norm, as plots in the bottom row look like the top row. In the loss plots of the non-conflict probe (the bottom left of blue and red panels), we observe two optimum spots – one in the bottom right with large α_D and small α_P (global optimum) while the other in the top left with large α_P yet small α_D (local optimum). This is a sign when people can get stuck in the local optimum of large α_P. Experimentally, we observe this behavioral pattern when people prioritize the P dimension, especially when facing a higher number of dimensions.

In terms of the cross-entropy plots that solely reflects the attention-accuracy association, the exponential linking function results in a steeper accuracy gradient around $\boldsymbol{\alpha = 0}$ than the logistic. This means that the accuracy change is more sensitive to variations in α_D and α_P around 0, and vice versa, the update of α_D and α_P will be more intensely responsive to accuracy (or errors) with the exponential linking function. Also, for the conflict probe shown in the right columns of red and blue panels, the logistic linking function favors high α_D and low α_P the most, while the exponential linking function suffices to high α_D and is relatively careless about the values of α_P.



Each plot represents certain values associated with an attention space $\boldsymbol{\alpha} = [\alpha_D, \alpha_P]$. The x axis depicts α_D while y axis depicts α_P. Results from models with the logistic linking function are in blue and the exponential function in red. On each colored panel, the left column is based on the non-conflict probe with D and P providing consistent information about categories, while the right column is based on the conflict probe. In these simulations, $p = .5$, $w = .05, \boldsymbol{b_0} = 0, b_1 = 1$, $g\left(\alpha_j^{(n)}\right) = 1$ (perfect memory); for the model with the logistic linking function, $\delta = 1$. There are three dimensions (one D, one P, and one I) each of two levels for this stimulus set. Simulations are based on the exemplar matrix of all eight unique exemplars.

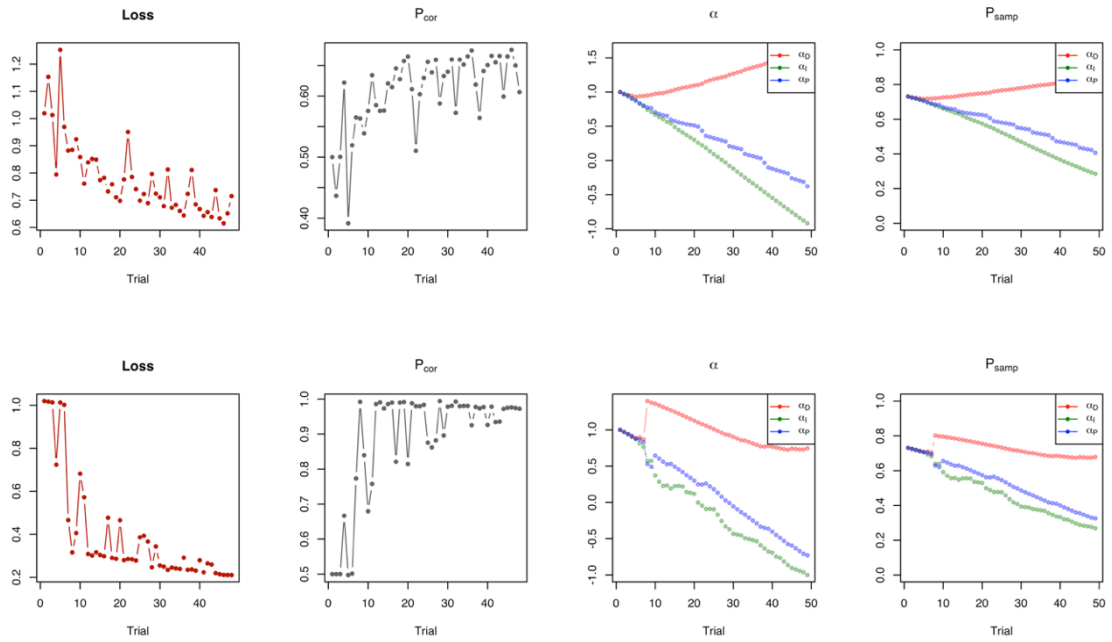
Figure 19 Cross entropy, ℓ_p norm, and loss regarding different attention vectors

3.2.2 Gradient Descent

We also simulate gradient descent processes with certain parameter values and different linking functions, with the partial encoding component $g\left(\alpha_j^{(n)}\right)$ included.

First, we simulate the attention update through gradient descent when learning a series of three-dimensional stimuli. In Figure 20, we see that given the same learning sequence,

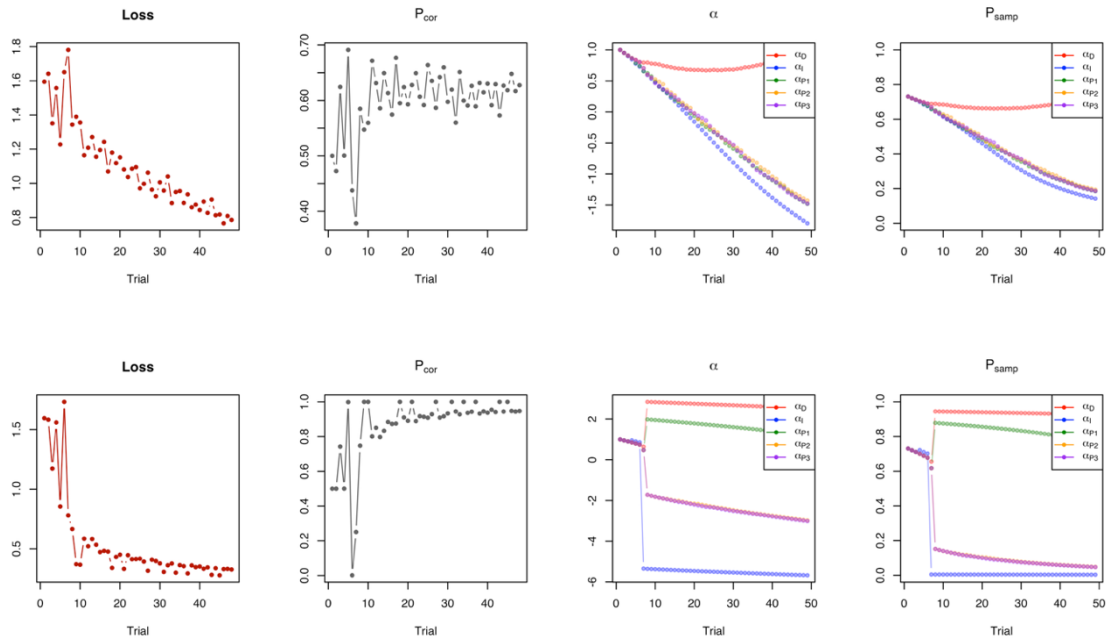
the same starting values of $\boldsymbol{\alpha}$, the model with the logistic linking function in the first row learns slowly, ended up with an accuracy at around 65% after 48 training trials when $\gamma = 1, \delta = 1$. Given the same learning rate, the model with the exponential function on the second row is more efficient in improving the accuracy. With the same weight $w = .05$, the model with the exponential function constrains the probabilities of sampling all dimensions early on, while the model with the model with the logistic function increases the probability of sampling D dimension to improve accuracy.



Each column respectively contains simulated results of loss, probability of correct response, attention, and probability of sampling each feature over the gradient descent process. The first row displays results from the logistic function while the second row from the exponential function. In these simulations, $p = .5, w = .05, \boldsymbol{b_0} = \mathbf{0}, b_1 = 1, \gamma = 1$; for the model with logistic linking functions, $\delta = 1$. The learning set involves 48 three-dimensional exemplars of the same random order.

Figure 20 Simulated gradient descent with three-dimensional exemplars

Furthermore, we run similar simulations on the learning set of five dimensions – one D, one I, and three P features. Figure 21 shows that the exponential linking function allows a faster accuracy increase and reallocation of attention than the logistic function with the same learning rate $\gamma = 1$ and $\delta = 1$.



Each column respectively contains simulated results of loss, probability of correct response, attention, and probability of sampling each feature over the gradient descent process. The first row displays results from the logistic function while the second row from the exponential function. In these simulations, $p = .5, w = .05, \boldsymbol{b_0} = \mathbf{0}, b_1 = 1, \gamma = 1$; for the model with logistic linking functions, $\delta = 1$. The learning set involves 48 five-dimensional exemplars of the same random order.

Figure 21 Simulated gradient descent with five-dimensional exemplars

Specifically, with the five-dimension learning set, we observe a large divergence between α_D and α_P triggered by conflict exemplars in simulations from the exponential function. With exponential-scaled loss, the steep gradient significantly pushes attention away from dimensions that provide incorrect category information to concentrate on the diagnostic dimensions. The sharp divergence of attention in conflict trials theoretically saves people from being stuck with suboptimal information – P or I dimension. Note that we structurally randomize the learning set here so that there are two conflict exemplars every eight trials just like the real experimental setup. Therefore, we observe salient patterns of attention divergence with early conflict exemplars. This may indicate that with the exponential linking function, the early exposure to conflict trials can effectively drive attention to diagnostic dimensions, while it may become trivial when applying the logistic linking function.

Chapter 4. Model Fitting

In this chapter, we fit our models to the data from our experiment under three conditions. We elaborate the methods of model fitting and model performance in this chapter.

4.1 Methods

For models with both linking functions, we estimate following free parameters: the starting value of attention α_0, attention constraint parameter p, and the slope b_1 and intercepts for each dimension $\boldsymbol{b_0} = [b_{0,1}, \dots, b_{0,D}]$. We assume that the initial latent attention to different dimensions be the same, $\alpha_{0,1} = \dots = \alpha_{0,D} = \alpha_0$, because people should have no prior info about this categorization task. In terms of attention constraints, although both w and p are related to penalizing attention vectors, this thesis fixes w in the model fitting procedure for parameter identifiability. Also, we set the learning rate parameter γ_0 at 1 for simplicity and identifiability. Specifically, in the model with the logistic linking function, we add one more parameter δ in activation calculation (see details in Chapter 3). Therefore, we have $(D + 4)$ parameters in models with the logistic linking function and $(D + 3)$ parameters in models with the exponential linking function, where D is the number of dimensions that vary between conditions.

Remember that our model can predict the trial-wise probabilities of correct response and probabilities of sampling each dimension. To calculate the likelihood of data given model

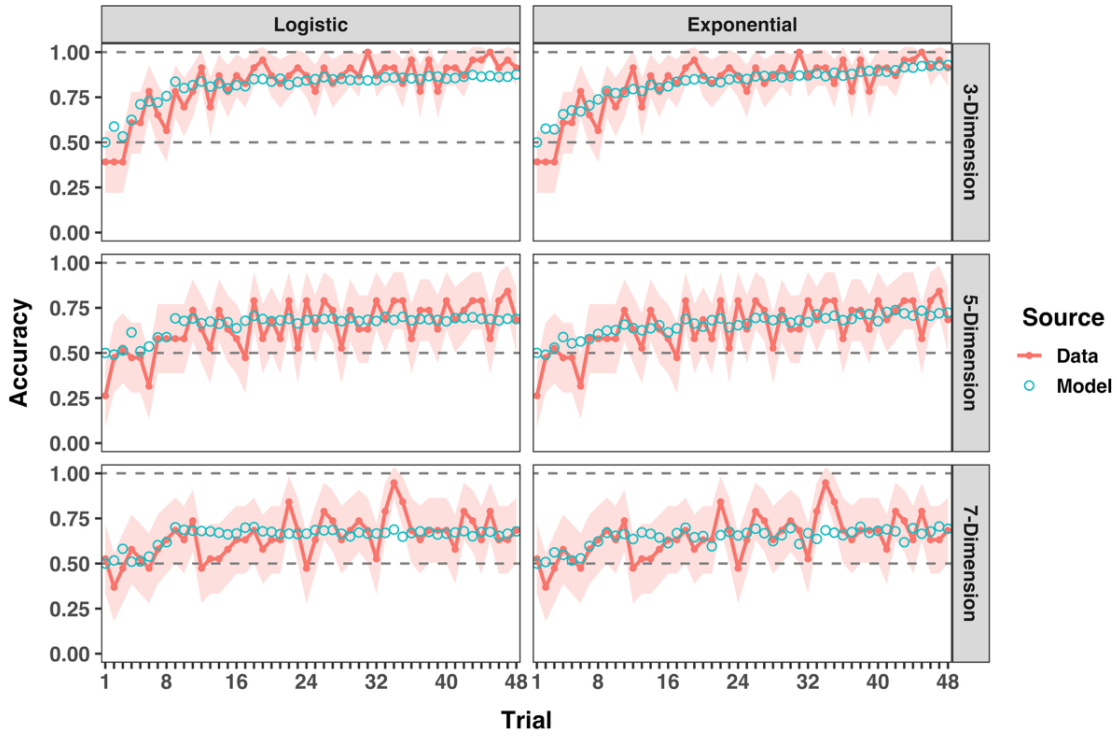
predictions, we code data into the binary format. Given two possible categories in this task, the response $\boldsymbol{R}^{(i)}$ is coded as the 0/1 vector of length $c = 2$ with 1 on the reported category. The sampling data on each trial $\boldsymbol{S}^{(i)} = \left[S_1^{(i)}, S_2^{(i)}, \dots, S_D^{(i)}\right]$ is coded as the vector of length D with 1 on the dimensions that are revealed for at least $200ms$ yet 0 otherwise. We take the durations $\geq 200ms$ sufficient for information to be sampled and encoded into memory.

Then, these two models with different linking functions were fitted to the training data of each participant. We obtain maximum likelihood estimation of parameter values through the BFGS method that sets no limitations to the raw inputs. We manually transform them from the real line onto the assumed scale within the likelihood function. After that, we use estimated parameters to simulate individual probability of response and dimension-wise probability of sampling and aggregate over participants on each trial. Finally, we compare predicted response and sampling metrics with the real data on the group level for each condition.

4.2 Results

4.2.1 Accuracy

Figure 22 suggests that models with both linking functions capture the group-level accuracy patterns throughout training for all three conditions. The model with the exponential linking function performs slightly better than the model with the logistic linking function in terms of reproducing the gradual increase of accuracy, especially in the later training phase.

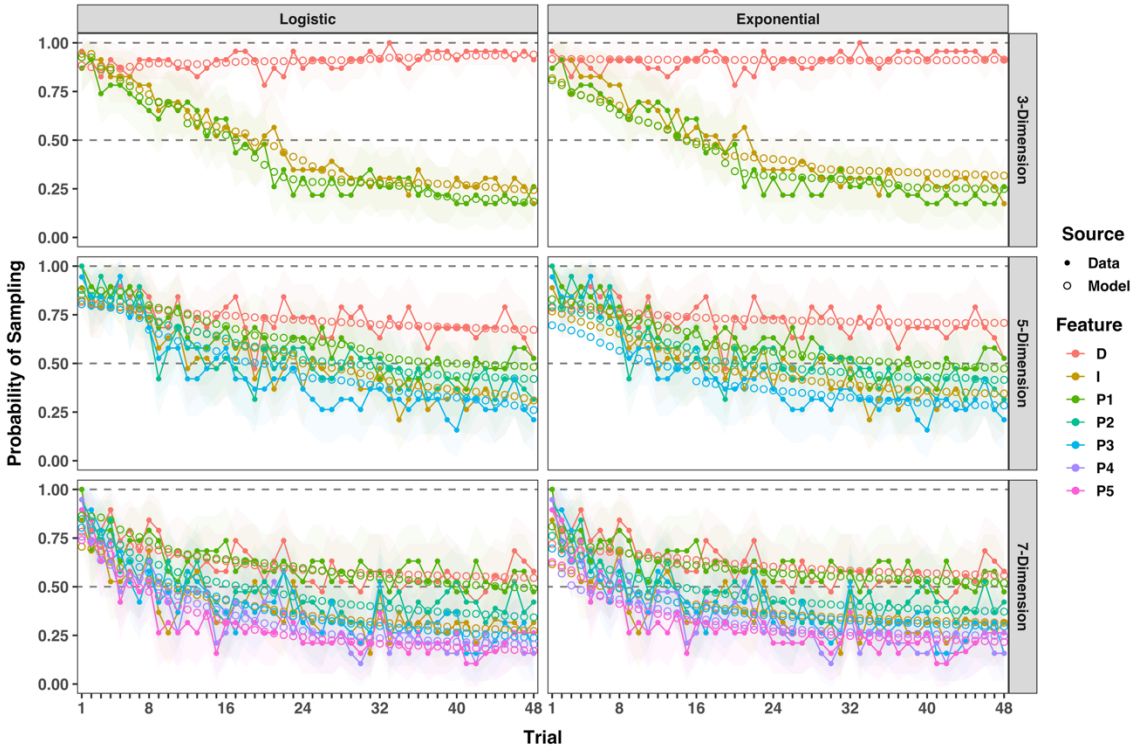


Model fits of accuracy data. Real data are shown in red filled dots connected with solid lines and model simulated data are shown in teal hollow circles. The left column contains results from models with logistic linking functions, while the right column contains results from models with exponential linking functions. Each row contains respectively the three-, five-, and seven-dimension data. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 22 Real vs. simulated accuracy over training

Taking the response as the sole metric for likelihood calculation, the average AIC from models with exponential functions are a bit better than those from models with logistic functions, $\Delta AIC_{exponential-logistic} = -1.3, -2.7, -1.8$ for the three-, five-, and seven-dimension conditions.

4.2.2 Sampling



Model fits of sampling data. Real data are displayed in filled dots connected with solid lines and model simulations are in hollow circles, with different colors representing different features. The left column contains results from the logistic linking function, while the right column from the exponential function. Each row contains respectively the 3-, 5-, and 7-dimension data. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 23 Real vs. simulated probability of sampling over training

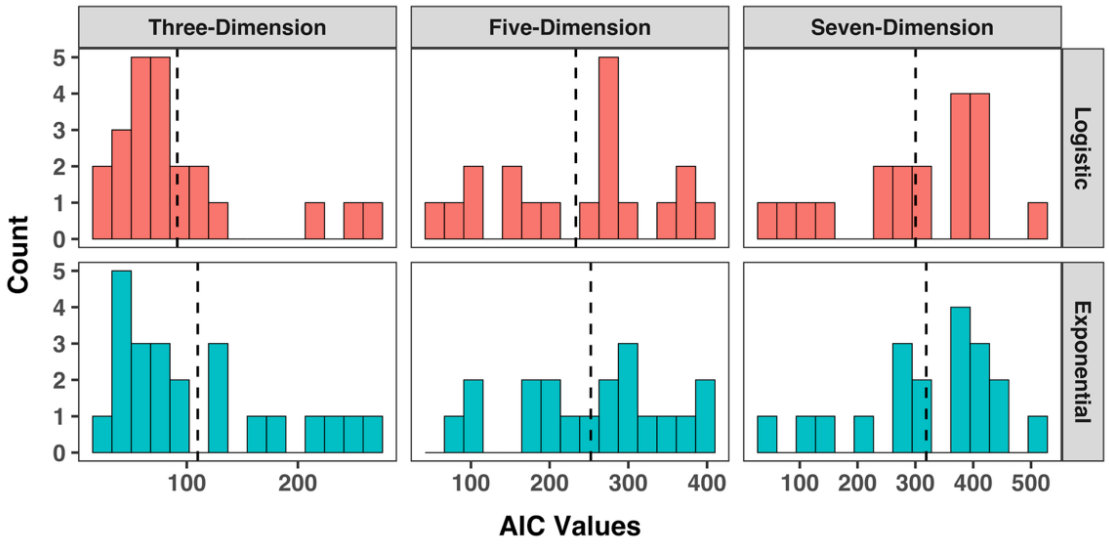
We then compare the model predicted probability of sampling for each dimension on each trial against the real sampling behavior averaging over participants under three

conditions. Figure 23 displays that both models learn to assign higher probabilities of sampling to the D feature under the three-dimension condition. Under five- and seven-dimension conditions, models learn to prioritize not only D dimension but also the most preferred P1 dimension while sampling. This suggests that our model is capable of reproducing selective attention patterns such as prioritizing D features and being trapped with certain P features.

Comparing simulated profiles of two types of models, the models with the logistic linking function are better at capturing the probability of sampling at the beginning of training than models with the exponential function. Taking the sampling behavior as the metric for likelihood calculation, the average AIC values calculated from the logistic function model are lower than those from the exponential function, $\Delta AIC_{exponential-logistic} = 19.7, 13, 21$ for the three-dimensional, five-dimensional, and seven-dimensional data.

To conclude, in two previous chapters, we present two models that both can approximate human learner's sampling profiles and learning outcomes in high-dimensional information space. Model simulation and model fitting results conjunctively validates our proposal of applying ℓ_p regularization of sampling costs alongside the cross entropy to compose a loss function that captures both performance and simplicity goals. Models with the exponential linking function is slightly better at recovering accuracies, while the model with the logistic linking function outperforms in probabilities of sampling. The overall AIC values for models with the logistic function are lower and better than those

from models with the exponential function (Figure 24). Nonetheless, these distinctions are not significant to strongly favor one linking function over the other in our case.



Each column contains results from the dataset in one condition while each row contains results from one linking function. The average AIC values are marked with dashed lines. The average AIC for the 3-dimension condition is 91.5 with the logistic and 110 with the exponential function; the average AIC for the 5-dimension condition is 233 with the logistic and 252 with the exponential function; the average AIC for the 7-dimension condition is 300 with the logistic and 319 with the exponential function.

Figure 24 Histograms of AIC values

Chapter 5. Conclusions and Discussion

5.1 General Conclusions

This thesis provides behavioral evidence and computational models to unravel features of selective attention during category learning. Our experiment systematically increases the number of dimensions towards the capacity of information processing, which reveals how people explore the high-dimensional information space and exploit a subset of dimensions for categorization decisions. We observe that as the number of dimensions approximates the attention limit, instead of distributing attention to all dimensions, people constrain the scope of attended dimensions and concentrate on a small subset. As a result, people are more likely to exploit a subset of sub-optimal information (i.e., P dimensions) and fall into learning traps. These patterns suggest that the high-dimensional learning set induces heavier burdens on attention, urging the capacity-limited attention system to emphasize the goal of simplicity and thus more possible to form an inaccurate representation of the environment.

Furthermore, inspired by AARM, we refine a modeling approach that realizes both goals of selective attention – maximizing accuracies and economizing cognitive efforts. To penalize sampling costs associated with eye movements, we apply the ℓ_p regularization on probabilities of sampling with the free parameter p continuous on $[0, 1]$. By adding ℓ_p norm of sampling probabilities to the loss function thus involving it into the optimization

process, the models implement the gradient descent over a functional representation of both goals of error minimization and effort reduction. At each iteration, our models update attention toward a state that in general brings higher accuracies and/or lower sampling probabilities. The simulation results and model fits show that our approach encapsules various aspects of attention constraints and recovers both accuracy and sampling patterns in the real data well.

Compared to the structural constraints of ℓ_p norm (Galdo et al., 2022), we allow more flexible functional constraints on attention by applying the ℓ_p regularization technique from sparse representation problems in statistics to cognitive modeling for "sparse attention problems". The ℓ_p regularization with $p < 1$ searches for a subset of attended dimensions with small coefficients (i.e., probability of sampling) that still maximizes the accuracy as possible. Our approach effectively frames the search for attention states that ensure decent performance without exhausting attention resources as an unconstrainted optimization process by taking ℓ_p norm as part of the loss. It can well capture the subject-to-subject differences in balancing the goal of accuracy and simplicity of selective attention.

5.2 Discussion and Future Directions

To model constraints of selective attention, the hallmark solution from Galdo et al. (2022) assumes one-to-one mapping from the algorithmic level to the computational level and applies separate parameters for different aspects of the computational efficiency goal. They use κ in the Ridge regularization to minimize the amount of attention, λ in the

Lasso regularization to minimize the number of attended dimensions, and β in gradient inhibition for competition. Instead, our model retains the interwoven association between different aspects of the quest for simplicity by promoting an elegant solution with ℓ_p regularization. When fitting partial encoding AARMs to our datasets with Galdo et al. (2022)'s solution to attention constraints, these models do not recover the behavioral patterns very well in some datasets. See the model fits from the PE-AARMs with the Lasso regularization (λ) and competition (β) in the Appendix B. Our models outperform in reproducing accuracy and sampling patterns for all datasets.

One additional assumption during the model fitting in this thesis is that we set $w = .05$ to only estimate p. However, these two parameters may play different roles when realizing attention constraints. p captures the specific form of simplicity, with a near-zero value reflecting heavier dimension reduction tendency in addition to limiting the total amount of attention. In contrast, w may reflect the weight of the overall simplicity goal relative to the accuracy goal. We fit our model with w as a free parameter to our data (Appendix C). Although the model fits are promising, some parameter estimates are extreme and hard to interpret. We will look deeper into the identifiability issue in the future.

Within these specific datasets, our model performance is invariant to the choice of linking functions. In Chapters 3 & 4, we present two linking functions – logistic and exponential – that map the real line to the positive scale, both of which can recover behavioral data well. There is certain concern on the inconsistent scale between $f\left(\alpha_j^{(i)}\right)$, $g\left(\alpha_j^{(n)}\right)$ and $psamp_j^{(i)}$ when applying the exponential function. Yet, this should have little impact

because $f\left(\alpha_j^{(i)}\right)$ and $g\left(\alpha_j^{(n)}\right)$ are used to compute the probability of response on $[0, 1]$, which attenuates the impact from different magnitudes of $f\left(\alpha_j^{(i)}\right)$ and $g\left(\alpha_j^{(n)}\right)$.

In addition to incorporate the ℓ_p regularization mechanism, another specification of our models compared to the original AARMs is that we clarify how decision weights $f\left(\alpha_j^{(i)}\right)$, memory strength $g\left(\alpha_j^{(n)}\right)$, and sampling weights or probabilities $psamp_j^{(i)}$ relate to latent attention $\alpha_j^{(i)}$ or $\alpha_j^{(n)}$. We maintain that memory strength and sampling weights are similar in how they relate to attention – they are driven by attention while variant to factors such as individual bias or feature salience. When applying the logistic linking function, the memory strength $g\left(\alpha_j^{(n)}\right)$ is the same as the sampling weight $psamp_j^{(n)}$. However, decision weights relate to attention driven by the optimization process, independent from external factors. This is particularly important when the physical property of some dimensions is systematically changed to influence the sampling probabilities and memory strength, but not decision weights. In these scenarios, we can properly test our mathematics to equalize memory strength and sampling probabilities, which are different from decision weights. We will validify this idea in a following experiment, where we manipulate the salience of some dimension to intentionally interfere with sampling and memory rather than decisions.

Bibliography

Baddeley A. (2003). Working memory: looking back and looking forward. *Nature reviews. Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, *1*(4), 371–394. https://doi.org/10.1068/p010371

Blair, M. R., Watson, M. R., & Meier, K. M. (2009a). Errors, efficiency, and the interplay between attention and category learning. *Cognition*, *112*(2), 330–336. https://doi.org/10.1016/j.cognition.2009.04.008

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009b). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1196–1206. https://doi.org/10.1037/a0016272

Blanco, N. J., & Sloutsky, V. M. (2019). Adaptive flexibility in category learning? Young children exhibit smaller costs of selective attention than adults. *Developmental Psychology*, *55*(10), 2060–2076. https://doi.org/10.1037/dev0000777

Blanco, N. J., Turner, B. M., & Sloutsky, V. M. (2023). The benefits of immature cognitive control: How distributed attention guards against learning traps. *Journal of Experimental Child Psychology*, *226*, 105548. https://doi.org/10.1016/j.jecp.2022.105548

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 4–4. https://doi.org/10.1167/11.5.4

Castro, L., Savic, O., Navarro, V., Sloutsky, V. M., & Wasserman, E. A. (2020). Selective and distributed attention in human and pigeon category learning. *Cognition*, *204*, 104350. https://doi.org/10.1016/j.cognition.2020.104350

Chen, L., Meier, K. M., Blair, M. R., Watson, M. R., & Wood, M. J. (2013). Temporal characteristics of overt attentional behavior during category learning. *Attention, Perception, & Psychophysics*, *75*(2), 244–256. https://doi.org/10.3758/s13414-012-0395-8

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*(1), 73–101. https://doi.org/10.1146/annurev.psych.093008.100427

Dolguikh, K., Tracey, T., & Blair, M. R. (2021). The ubiquity of selective attention in the processing of feedback during category learning. *PLOS ONE*, *16*(12), e0259517. https://doi.org/10.1371/journal.pone.0259517

Donoho, D. L., & Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via L^1 minimization. *Proceedings of the National Academy of Sciences*, *100*(5), 2197–2202. https://doi.org/10.1073/pnas.0437847100

Dukas R. (2004). Causes and consequences of limited attention. *Brain, behavior and evolution*, *63*(4), 197–210. https://doi.org/10.1159/000076781

Fu, W. J. (1998). Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, *7*(3), 397–416. https://doi.org/10.2307/1390712

Galdo, M., Weichart, E. R., Sloutsky, V. M., & Turner, B. M. (2022). The quest for simplicity in human learning: Identifying the constraints on attention. *Cognitive Psychology*, *138*, 101508. https://doi.org/10.1016/j.cogpsych.2022.101508

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–340. https://doi.org/10.1037/a0019042

Hu, Y., Li, C., Meng, K., Qin, J., & Yang, X. (2017). Group sparse optimization via $L_{p,q}$ regularization. *The Journal of Machine Learning Research, 18*(1), 960–1011. https://doi.org/10.48550/arXiv.1601.07779

Jolicœur, P., & Dell'Acqua, R. (1999). Attentional and structural constraints on visual encoding. *Psychological Research*, *62*(2–3), 154–164. https://doi.org/10.1007/s004260050048

King, N. (2024). *Learning in the context of partial information* [Unpublished Master's Thesis]. The Ohio State University.

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. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*(6), 812–863. https://doi.org/10.1006/jmps.2000.1354

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636–645. https://doi.org/10.3758/BF03213001

Lavie N. (1995). Perceptual load as a necessary condition for selective attention. Journal of experimental psychology. *Human perception and performance, 21*(3), 451–468. https://doi.org/10.1037//0096-1523.21.3.451

Lavie, N., Beck, D. M., & Konstantinou, N. (2014). Blinded by the load: attention, awareness and the role of perceptual load. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*(1641), 20130205. https://doi.org/10.1098/rstb.2013.0205

Lee, W. J., Li, A. X., Lee, J. E., & Hayes, B. K. (2024). Learning traps and change blindness in dynamic environments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*(9), 1345–1360. https://doi.org/10.1037/xlm0001390

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*(6), 493–497. https://doi.org/10.1016/S0960-9822(03)00135-0

Matsuka, T., & Corter, J. E. (2008). Observed attention allocation processes in category learning. *Quarterly Journal of Experimental Psychology*, *61*(7), 1067–1097. https://doi.org/10.1080/17470210701438194

McColeman, C. M., Barnes, J. I., Chen, L., Meier, K. M., Walshe, R. C., & Blair, M. R. (2014). Learning-induced changes in attentional allocation during categorization: A sizable catalog of attention change as measured by eye movements. *PLoS ONE*, *9*(1), e83302. https://doi.org/10.1371/journal.pone.0083302

Meier, K. M., & Blair, M. R. (2013). Waiting and weighting: Information sampling is a balance between efficiency and error-reduction. *Cognition*, *126*(2), 319–325. https://doi.org/10.1016/j.cognition.2012.09.014

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77. https://doi.org/10.1016/j.cogpsych.2016.01.001

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

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (1st ed., pp. 18–39). Cambridge University Press. https://doi.org/10.1017/CBO9780511921322.002

Nosofsky, R. M., & Hu, M. (2023). Category structure and region-specific selective attention. *Memory & Cognition*, *51*(4), 915–929. https://doi.org/10.3758/s13421-022-01365-4

Palmer J. (1990). Attentional limits on the perception and memory of visual information. Journal of experimental psychology. *Human perception and performance, 16*(2), 332–350. https://doi.org/10.1037//0096-1523.16.2.332

Pashler, H., Johnston, J. C., & Ruthruff, E. (2001). Attention and performance. *Annual review of psychology*, 52, 629–651. https://doi.org/10.1146/annurev.psych.52.1.629

Paskewitz, S., & Jones, M. (2020). Dissecting EXIT. *Journal of Mathematical Psychology*, *97*, 102371. https://doi.org/10.1016/j.jmp.2020.102371

Fu, W. J. (1998). Penalized Regressions: The Bridge versus the Lasso. Journal of *Computational and Graphical Statistics, 7*(3), 397–416. https://doi.org/10.2307/1390712

Plebanek, D. J., & Sloutsky, V. M. (2017). Costs of selective attention: when children notice what adults miss. *Psychological science, 28*(6), 723–732. https://doi.org/10.1177/0956797617693005

Poletti, M., Rucci, M., & Carrasco, M. (2017). Selective attention within the foveola. *Nature Neuroscience*, *20*(10), 1413–1417. https://doi.org/10.1038/nn.4622

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 811–829. https://doi.org/10.1037/0278-7393.31.5.811

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of experimental psychology. General, 147*(11), 1553–1570. https://doi.org/10.1037/xge0000466

Schall, J. D., & Thompson, K. G. (1999). Neural selection and control of visually guided eye movements. *Annual Review of Neuroscience*, *22*(1), 241–259. https://doi.org/10.1146/annurev.neuro.22.1.241

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

Symons, A. E., Dick, F., & Tierney, A. T. (2021). Dimension-selective attention and dimensional salience modulate cortical tracking of acoustic dimensions. *NeuroImage*, *244*, 118544. https://doi.org/10.1016/j.neuroimage.2021.118544

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (2011). Regression Shrinkage and Selection via The Lasso: A Retrospective. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(3), 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

Turner, B. M. (2019). Toward a common representational framework for adaptation. *Psychological Review*, *126*(5), 660–692. https://doi.org/10.1037/rev0000148

Turner, B. M., & Sloutsky, V. M. (2024). Cognitive Inertia: Cyclical Interactions Between Attention and Memory Shape Learning. *Current Directions in Psychological Science, 33*(2), 79–86. https://doi.org/10.1177/09637214231217989

Weichart, E. R., Evans, D. G., Galdo, M., Bahg, G., & Turner, B. M. (2022). Distributed neural systems support flexible attention updating during category learning. *Journal of Cognitive Neuroscience*, *34*(10), 1761–1779. https://doi.org/10.1162/jocn_a_01882

Weichart, E. R., Unger, L., King, N., Sloutsky, V. M., & Turner, B. M. (2024). "The eyes are the window to the representation": Linking gaze to memory precision and decision weights in object discrimination tasks. *Psychological Review*, *131*(4), 1045–1067. https://doi.org/10.1037/rev0000475

West, G. L. (2010). Capacity limits during perceptual encoding. *Journal of Vision*, *10*(2), 1–12. https://doi.org/10.1167/10.2.14

Appendix A. Loss Derivation

$$\frac{\partial}{\partial \alpha_j^{(i)}} loss = \frac{\partial}{\partial \alpha_j^{(i)}} \mathbb{CE} + \frac{\partial}{\partial \alpha_j^{(i)}} w \left\| psamp^{(i)} \right\|_p$$

$$= -\frac{\partial}{\partial \alpha_j^{(i)}} log\big(P(correct)\big) + w \frac{\partial}{\partial \alpha_j^{(i)}} \left\| psamp^{(i)} \right\|_p$$

Deriviation of cross entropy

$$\frac{\partial}{\partial \alpha_j^{(i)}} \log\big(P(correct)\big) = \frac{\partial}{\partial \alpha_j^{(i)}} \sum_c \left[log\left(P^{(i)}(c)\right) \mathbb{I}_{\{F^{(i)}=c\}} \right]$$

$$= \sum_c \left[\frac{\partial}{\partial \alpha_j^{(i)}} log\left(P^{(i)}(c)\right) \mathbb{I}_{\{F^{(i)}=c\}} \right]$$

$$\frac{\partial}{\partial \alpha_j^{(i)}} \log\left(P^{(i)}(c)\right) = \frac{1}{P^{(i)}(c)} \frac{\partial}{\partial \alpha_j^{(i)}} P^{(i)}(c)$$

$$= \frac{1}{P^{(i)}(c)} \frac{\partial}{\partial \alpha_j^{(i)}} \frac{\sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=c\}}}{\sum_k \sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=k\}}}$$

$$= \frac{1}{P^{(i)}(c)} \frac{1}{\left[\sum_k \sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(n)}=k\}}\right]^2}$$

$$\left(\sum_k \sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=k\}} \frac{\partial}{\partial \alpha_j^{(i)}} \sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=c\}} \dots \right.$$

$$\left. \dots - \sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=c\}} \frac{\partial}{\partial \alpha_j^{(i)}} \sum_k \sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=k\}} \right)$$

$$= \frac{1}{P^{(i)}(c)} \frac{1}{\left[\sum_k \sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(n)}=k\}}\right]^2}$$

$$\left(\sum_k \sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=k\}} \sum_n \frac{\partial}{\partial \alpha_j^{(i)}} a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=c\}} \cdots\right)$$

$$\left(\ldots - \sum_n a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=c\}} \sum_k \sum_n \frac{\partial}{\partial \alpha_j^{(i)}} a\left(e^{(i)}, x^{(n)}\right) \mathbb{I}_{\{F^{(i)}=k\}}\right)$$

$$\frac{\partial}{\partial \alpha_j^{(i)}} a\left(e^{(i)}, x^{(n)}\right) = \frac{\partial}{\partial \alpha_j^{(i)}} \prod_m a_m\left(e^{(i)}, x^{(n)}\right)$$

$$= \frac{\partial}{\partial \alpha_j^{(i)}} \prod_{m \neq j} a_m\left(e^{(i)}, x^{(n)}\right) a_j\left(e^{(i)}, x^{(n)}\right)$$

$$= \prod_{m \neq j} a_m\left(e^{(i)}, x^{(n)}\right) \frac{\partial}{\partial \alpha_j^{(i)}} a_j\left(e^{(i)}, x^{(n)}\right)$$

$$= \prod_{m \neq j} e^{-f\left(\alpha_m^{(i)}\right) d_m\left(e^{(i)}, x^{(n)}\right) g\left(\alpha_m^{(n)}\right)} \frac{\partial}{\partial \alpha_j^{(i)}} e^{-f\left(\alpha_j^{(i)}\right) d_j\left(e^{(i)}, x^{(n)}\right) g\left(\alpha_j^{(n)}\right)}$$

$$= \prod_{m \neq j} e^{-f\left(\alpha_m^{(i)}\right) d_m\left(e^{(i)}, x^{(n)}\right) g\left(\alpha_m^{(n)}\right)} e^{-f\left(\alpha_j^{(i)}\right) d_j\left(e^{(i)}, x^{(n)}\right) g\left(\alpha_j^{(n)}\right)} \ldots$$

$$\ldots \frac{\partial}{\partial \alpha_j^{(i)}} \left(-f\left(\alpha_j^{(i)}\right) d_j\left(e^{(i)}, x^{(n)}\right) g\left(\alpha_j^{(n)}\right)\right)$$

$$= \prod_m e^{-f\left(\alpha_m^{(i)}\right) d_m\left(e^{(i)}, x^{(n)}\right) g\left(\alpha_m^{(n)}\right)} \left(-d_j\left(e^{(i)}, x^{(n)}\right) g\left(\alpha_j^{(n)}\right)\right) \frac{\partial}{\partial \alpha_j^{(i)}} f\left(\alpha_j^{(i)}\right)$$

For the logistic linking function, $f\left(\alpha_j^{(i)}\right) = \frac{\delta}{1+e^{-\alpha_j^{(i)}}}$, $g\left(\alpha_j^{(n)}\right) = \frac{1}{1+e^{-\left(b_{0,j}+b_1 \alpha_j^{(n)}\right)}}$

$$\frac{\partial}{\partial \alpha_j^{(i)}} f\left(\alpha_j^{(i)}\right) = \delta f\left(\alpha_j^{(i)}\right)\left(1 - f\left(\alpha_j^{(i)}\right)\right)$$

For the exponential linking function, $f\left(\alpha_j^{(i)}\right) = \mathrm{e}^{\alpha_j^{(i)}}$, $g\left(\alpha_j^{(n)}\right) = e^{b_{0,j}+b_1 \alpha_j^{(n)}}$

$$\frac{\partial}{\partial \alpha_j^{(i)}} f\left(\alpha_j^{(i)}\right) = f\left(\alpha_j^{(i)}\right)$$

Deriviation of ℓ_p norm of sampling probabilities

$$\frac{\partial}{\partial \alpha_j^{(i)}} \left\| psamp^{(i)} \right\|_p = \frac{\partial}{\partial \alpha_j^{(i)}} \left(\sum_m \left[psamp_m^{(i)} \right]^p \right)^{1/p}$$

$$= \frac{1}{p} \left(\sum_m \left[psamp_m^{(i)} \right]^p \right)^{1/p-1} \frac{\partial}{\partial \alpha_j^{(i)}} \sum_m \left[psamp_m^{(i)} \right]^p$$

$$= \frac{1}{p} \left(\sum_m \left[psamp_m^{(i)} \right]^p \right)^{1/p-1} \sum_m \frac{\partial}{\partial \alpha_j^{(i)}} \left[psamp_m^{(i)} \right]^p$$

$$= \frac{1}{p} \left(\sum_m \left[psamp_m^{(i)} \right]^p \right)^{1/p-1} \frac{\partial}{\partial \alpha_j^{(i)}} \left[psamp_j^{(i)} \right]^p$$

$$= \frac{1}{p} \left(\sum_m \left[psamp_m^{(i)} \right]^p \right)^{1/p-1} p \left[psamp_j^{(i)} \right]^{p-1} \frac{\partial}{\partial \alpha_j^{(i)}} psamp_j^{(i)}$$

$$= \left(\sum_m \left[psamp_m^{(i)} \right]^p \right)^{1/p-1} \left[psamp_j^{(i)} \right]^{p-1} \frac{\partial}{\partial \alpha_j^{(i)}} psamp_j^{(i)}$$

$$\frac{\partial}{\partial \alpha_j^{(i)}} psamp_j^{(i)} = \frac{\partial}{\partial \alpha_j^{(i)}} \frac{1}{1+e^{-\left(b_{0,j}+b_1 \alpha_j^{(i)}\right)}}$$

$$= \left[1 + e^{-\left(b_{0,j}+b_1 \alpha_j^{(i)}\right)} \right]^{-2} \left(-\frac{\partial}{\partial \alpha_j^{(i)}} \left(1 + e^{-\left(b_{0,j}+b_1 \alpha_j^{(i)}\right)} \right) \right)$$

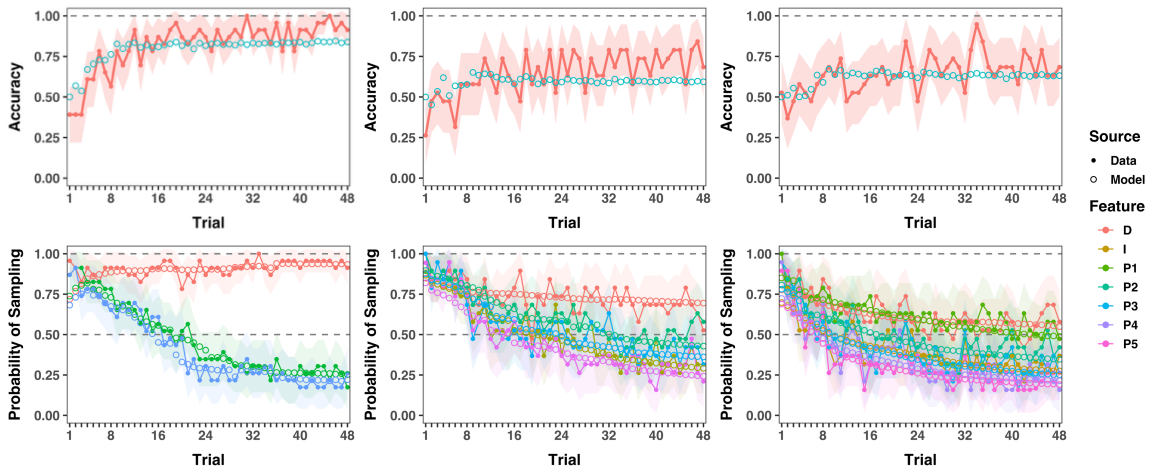
$$= \left[1 + e^{-\left(b_{0,j}+b_1 \alpha_j^{(i)}\right)} \right]^{-2} e^{-\left(b_{0,j}+b_1 \alpha_j^{(i)}\right)} b_1$$

$$= b_1 \frac{1}{1+e^{-\left(b_{0,j}+b_1 \alpha_j^{(i)}\right)}} \left(1 - \frac{1}{1+e^{-\left(b_{0,j}+b_1 \alpha_j^{(i)}\right)}} \right)$$

$$= b_1 psamp_j^{(i)} \left(1 - psamp_j^{(i)} \right)$$

Appendix B. Model Fits of PE-AARMs with Competition and Lasso Regularization

We fit the AARMs with the partial encoding component and Galdo et al. (2022)'s solution to attention constraints – the Lasso regularization (λ) and competition (β). We set the slope b_1 to 1. For simplicity, we only test models with the logistic linking function and compare those to our models with the logistic function. The average AIC values for the three-, five-, and seven-dimension conditions are $94.4, 241$, and 307, which are generally higher than our solution with the logistic linking function ($\Delta AIC = 2.9, 8, 7$).

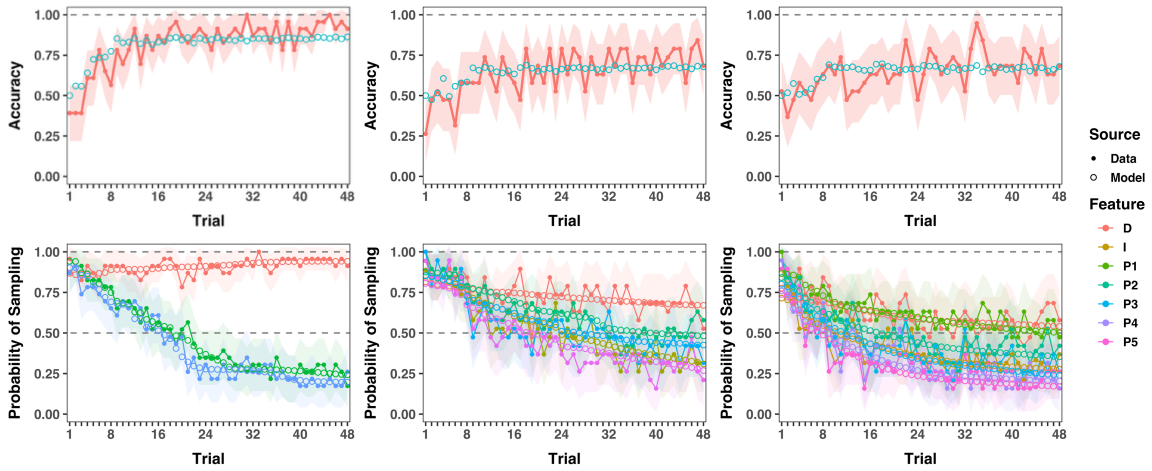


Model fits of accuracy and sampling data. Real data are shown in filled dots connected with solid lines and model simulated data are shown in hollow circles. Accuray plots are on the first row and sampling plots are on the second row. Each column contains three-, five-, and seven-dimension conditions respectively from left to right. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 25 Real vs. simulated accuracy and sampling from model with competition and Lasso regularization

Appendix C. Model Fits of PE-AARMs with ℓ_p Regularization with Free w

We fit our model with the logistic linking function to three datasets while allowing w as a free parameter between 0 and 1. The model can recover sampling and accuracy patterns, while there are identifiability issues such as extremely large estimates of slope b_1 in the three-dimension condition, some estimated weight parameters as zero.



Model fits of accuracy and sampling data. Real data are shown in filled dots connected with solid lines and model simulated data are shown in hollow circles. Accuray plots are on the first row and sampling plots are on the second row. Each column contains three-, five-, and seven-dimension conditions respectively from left to right. Colored bands show the 90% confidence interval based on the sampling distribution.

Figure 26 Real vs. simulated accuracy and sampling from model with ℓ_p regularization, free w